

SESSION

COMPUTATIONAL METHODS FOR MICROARRAY, GENE EXPRESSION ANALYSIS, AND GENE REGULATORY NETWORKS

Chair(s)

TBA

Optimization of a Microarray Probe Design Focusing on the Minimization of Cross-hybridization

F. Horn¹, H.-W. Nützmann², V. Schroeckh², R. Guthke¹, and C. Hummert¹

¹Research Group Systems Biology / Bioinformatics

²Department of Molecular and Applied Microbiology

Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany

Abstract—*Microarrays are extensively used for high-throughput gene expression analyses in molecular biology. Microarray analysis is reliable if the probe binds specifically to the intended target transcript. Cross-hybridizations of microarray probes is one of the main systematic errors which is influenced by microarray probe design. Newly released genome annotations make it possible and necessary to improve given probe designs in order to reduce this source of error.*

*We present a new method which evaluates and optimizes existing probe designs in a modular way. The workflow can include existing software and it can be adapted to additionally required probe design criteria. A microarray probe design optimization which focuses on the avoidance of cross-hybridization was exemplarily done for *Aspergillus nidulans*. We show the high impact of the underlying structural genome annotation on the probe design process. The new design was experimentally evaluated with the help of the mean variance of internal technical replicates.*

Keywords: microarray, probe design, cross-hybridization, *Aspergillus nidulans*, optimization

1. Introduction

Microarray technique represents one of the most common methods to carry out genome-wide research based on sequenced genomes. A microarray experiment consists of many different steps which are all vulnerable to errors. Signal intensities strongly depend on the probe sequence, because different sequences generate varying physical properties, which are important for hybridization [1]. The properties of the probe sequences may be predicted and they are used for the microarray probe design [2].

The main objective of the design process is to increase the reliability of signal intensities by reducing systematic errors caused by the probe sequences. Among other criteria, the hybridization process itself is modeled with the help of criteria, like melting temperature uniformity, GC-content, prediction of secondary structures and Free Gibbs energy [3].

In order to guarantee a high discrimination between targets and non-targets, the probe design is checked for cross-hybridization. Cross-hybridization is a non-target binding between a probe and a transcript fragment which is not

intended to match the probe. In fact, cross-hybridizations are one of the main sources of systematic error that affect tiling arrays [4] and even the well-established microarrays from Affymetrix [5]. Several studies have shown that nucleotide sequences are capable of hybridization, even when the complementary region between probe and transcript has only a 70% identity [1], [6]. Besides this identity threshold, non-specific bindings additionally need a longest continuous complementary substring of a certain minimum length [6], [3]. Signal intensities in the data may result from unspecific bindings and may lead to false-positively detected target genes.

There are approaches to cope with cross-hybridizations by creating new alternative Chip Definition Files (CDFs) of existing custom microarray probe designs [7], [8]. These methods correct and avoid the impact of cross-hybridizations by disregarding a certain fraction of the probes during data analysis. It is evident that the same level of information can be obtained with less probes spotted onto the microarray. The reannotation of oligonucleotide libraries is therefore the first step in order to obtain up-to-date microarray probe designs [9]. It is preferable to exclude existing cross-hybridizing oligonucleotides during the process of optimizing microarray probe designs [10]. This removal leads to a reduced production cost for each utilized data point. New alternative probes can be spotted onto the microarray which leads to a higher genome coverage rate or a higher number of replicates per gene.

Many different algorithms have been proposed for designing microarray probes [2]. Each algorithm has a different scope of application and consequently utilizes different probe design criteria and, as a consequence, perform differently. The different foci make it difficult to directly evaluate and compare the quality of the proposed algorithms with a theoretical optimization criterion. In fact, the limitations of the applied experimental protocol determine suitable probe design criteria and narrow down the set of available methods. It is favorable to use an extendable and adjustable general framework where different probe design criteria can be integrated [11], [12]. This allows to adjust for application-specific design criteria and enables the reuse of existing modular software.

In this work, we present a workflow which evaluates and

optimizes an already given reference probe design concerning the avoidance of cross-hybridization. The optimization of the probe design is exemplarily done for a microarray for *Aspergillus nidulans* which is a model organism of filamentous fungi [13]. The obtained probe design minimizes unspecific bindings. We show that this design yields more reliable results. In addition to the avoidance of cross-hybridizations, it is possible to include different design criteria which are applied due to experimental constraints.

2. Results

2.1 Evaluation of reference probe design

The mapping of a given full-genome probe design for *Aspergillus nidulans* was examined by aligning the probe sequences against three structural genome annotations: two different versions available from the Broad institute and one version from the Central Aspergillus Data REpository (CADRE). (The annotations are referred to as BROAD (2008), BROAD (2010) and CADRE (2009), respectively.) For further information see methods and figure 1.

The given reference probe design contains 342 and 377 probes that cross-hybridize with BROAD (2008) and CADRE (2009) annotation, respectively (see table 1). Regarding the newer BROAD (2010) annotation, only 148 probes are considered as cross-hybridizing.

Using the BROAD (2008) annotation and the CADRE (2009) annotation respectively, 317 and 313 probes in the reference probe design do not match any transcript with a perfect sequence identity.

The reference probe design contains probes that do not match any transcript in the given annotation: 74 probes using BROAD (2008), 204 probes using CADRE (2009), and 993 probes using the newer BROAD (2010).

The reference probe design does not cover a number of predicted transcripts in each annotation: 442 transcripts in BROAD (2008), 478 transcripts in CADRE (2009), and as much as 968 transcripts in BROAD (2010).

The evaluation also calculated the thermodynamic properties of the probe sequences. The result reveals that the melting temperatures of the probes are in a narrow range between 80°C and 90°C. This desirable property is achieved with the help of a uniform GC content of 48%.

In summary, the reference probe design is not optimized for any of the used annotations. Depending on the used annotation version, 7...11% of all probes do not match a transcript unambiguously. The current annotation causes a poorer performance which can be seen explicitly at the decreased number of perfect probes (see table 1).

2.2 Probe design optimization

A large fraction of the reference probe design is not optimized for any genome annotation and needs improvement. The objective of the optimization was to get 50 nucleotides

long optimized oligonucleotides which use the BROAD (2008) annotation. The probes should be placed at the 5'-end because cDNA is used in the hybridization protocol.

The workflow of the proposed probe design method can be separated into three consecutive steps (see figure 2). In the first step new probe candidates are generated with the help of ArrayOligoSelector [14]. In the second step, probe candidates are evaluated with the help of evaluation tool to exclude cross-hybridizations (see above). The evaluation also calculates thermodynamic properties that are used in a following third step - a further selection. The selection step is necessary because only one probe sequence per gene is spotted.

The optimization showed that it was not possible to find a valid unique probe sequence for every transcript. In order to achieve a higher gene coverage, design criteria have to be mitigated. New probe candidates are iteratively generated from intervals of elongated transcript sequences. 1,303 probes were found in the smallest interval of 600 basepairs (see table 2). In the next two steps the interval is extended to 1,200 and 2,000 basepairs which only led to 30 and 24 additional probes, respectively. In a last step, probes that are capable of cross-hybridization are exceptionally allowed. The relaxation of this last criterion increased gene coverage with 53 additional probes. In total, the softening of the design criteria leads to 107 additionally covered genes in the presented study.

Finally, there are 188 genes without a valid probe sequence which leads to a transcript coverage rate of 98,2%.

The comparison of the resulting new probe design with the given reference probe design shows that the new probe design is optimized for the BROAD (2008) annotation (see table 1). The new design consists of 10,512 probes (99,5%) which match perfectly and do not show any cross-hybridization. Notably, the comparison with the reference probe design demonstrates that 254 genes are additionally covered in the optimized design while avoiding systematic errors.

Remarkably, there are also 214 extra covered genes if the CADRE (2009) annotation is used as basis. This result is achieved by a lower number of genes with systematic errors. The number of potentially cross-hybridizing probes is only 133 in comparison to 377 probes in the reference probe design. Only three specific probes match a transcript without a total sequence identity whereas this number is much higher in the reference probe design with 313 probes. Changes in the annotation lead to 143 probes that do not match any given transcript in contrast to 204 probes in the reference probe design. The number of uncovered genes is 264 which corresponds to a gene coverage rate of 97,5%.

For the current BROAD (2010) annotation the gene coverage of the probe design is reduced to 90,5% and the number of covered genes (9,561 vs. 9,592) is comparable between both versions of the probe design. Nevertheless, the new

Table 1: Results of probe classification and gene coverage

Annotation Probe design	BROAD (2008)		CADRE (2009)		BROAD (2010)	
	old	new	old	new	old	new
Number of probes	10,676	10,566	10,676	10,566	10,676	10,566
Perfect probes	9,943 (93.1%)	10,513 (99.5%)	9,782 (91.6%)	10,287 (97.4%)	9,535 (89.3%)	9,535 (90.2%)
Cross-hybridizing	342 (3.2%)	53 (0.5%)	377 (3.5%)	133 (1.3%)	148 (1.4%)	63 (0.6%)
Not identical match	317 (3.0%)	0 (0.0%)	313 (2.9%)	3 (0.0%)	0 (0.0%)	0 (0.0%)
Not matching	74 (0.7%)	0 (0.0%)	204 (1.9%)	143 (1.4%)	993 (9.3%)	968 (9.2%)
Total number of genes	10,701	10,701	10,546	10,546	10,560	10,560
Covered genes	10,259 (95.9%)	10,513 (98.2%)	10,068 (95.5%)	10,282 (97.5%)	9,592 (90.8%)	9,561 (90.5%)
Uncovered genes	442 (4.1%)	188 (1.8%)	478 (4.5%)	264 (2.5%)	968 (9.2%)	999 (9.5%)

Probes from the reference probe design (old) and the optimized (new) probe design have been mapped to different genome annotations. Probes either show no systematic error (perfect probes), hybridize with multiple genes (cross-hybridizing), match one gene without total sequence identity (not identical match), or do not match any transcript at all (not matching). The lower part of the table shows how many genes of the annotation are perfectly covered by the corresponding probe design.

Table 2: Composition of the gene coverage

	Number of genes
Reference probe design (validated probes)	9,103
Probe design optimization:	
Sequence range: 0...600 bp	1,303
Sequence range: 0...1,200 bp	30
Sequence range: 0...2,000 bp	24
Ignoring cross-hybridizations	53
Uncovered genes	188
Total	10,701

The gene coverage of the probe design results from different steps. A high number of genes are covered by validated probes from the reference probe design. The probe design optimization leads to an additional number of covered genes which are obtained by iteratively mitigating the probe design criteria. First, the transcript sequences are extended and at last the cross-hybridization criterion is relaxed. In the end, some genes remain that are not covered by any valid probe.

probe design still minimizes systematic errors. 63 probes are prone to cross-hybridizations in contrast to 148 probes in the reference probe design. A high number of 968 probes do not match any transcript at all which is again comparable to the performance of the reference probe design.

In summary, the new probe design reduces systematic errors regardless of the structural annotation used. Concerning the cross-hybridizations, the improvements become apparent. For BROAD (2008) and CADRE (2009) the gene coverage of the optimized probe design is higher as compared to the reference probe design.

2.3 Impact of genome annotation

The evaluation of different probe designs clearly highlights the big impact of the underlying structural genome annotation on the results (see table 1).

The new probe design was optimized for the BROAD (2008) annotation and the gene coverage could be increased to 98.2%. The optimization also takes effect for the CADRE (2009) annotation with a gene coverage rate of 97.5%. In comparison to the current BROAD (2010) annotation, the gene coverage rate is dramatically decreased to 90.5% which

is comparable with the coverage rate of the reference probe design. The same trend for gene coverage can be seen for the reference probe design where the gene coverage rate also decreases to 90.8% if the BROAD (2010) annotation is used.

The differences in gene coverage result from probes which are vulnerable to systematic errors. The new probe design shows only a small fraction of probes that are prone to cross-hybridization in the BROAD (2008) annotation. This number doubles if the CADRE (2009) annotation is used. In the BROAD (2010) annotation only a few cross-hybridizing probes occur. This results from the increased number of error prone probes that do not match any transcript at all. The number of unmatched probes constitutes the largest error source which is affected by the change in genome annotation.

In the probe design optimized for BROAD (2008), the number of probes that are not classified as perfect increases from 54 (0.6%) over 279 (2.7%) to 1031 (9.8%) for the BROAD (2008), CADRE (2009), and BROAD (2010) annotation, respectively. The same trend holds for the non-perfect probes from the reference probe design which increases from 733 (6.9%) over 894 (8.4%) to 1141 (10.7%). It is noteworthy that a change in the annotation basis can cause almost 10% of all probes to be classified as invalid.

2.4 Experimental Validation

The new probe design is optimized for the minimization of systematic errors in respect to the BROAD (2008) annotation. Especially, the avoidance of cross-hybridization should significantly increase the reliability of experimental data. An indicator for improved reliability is a lower mean variance of internal technical replicates over each array. For this purpose, a highly reproducible experiment with the reference and the new probe design was performed (see methods). Microarray raw data was obtained from *Aspergillus nidulans* - *Streptomyces rapamycinicus* interaction experiments. The co-cultivation was performed because most of the secondary metabolite gene clusters are silent under laboratory condi-

tions and the fungal-bacterial interaction leads to specific activations [15], [16]. (Microarray data is available at Gene Expression Omnibus - GSE25266.)

First, a microarray experiment using the reference probe design was performed. The following second experiment used the same experimental setup except that the new optimized probe design was used. It is not possible to compare the variance of probes for each single gene individually because an altered probe sequence has an essential impact on the signal intensities. Probes with the same nucleotide sequences have a high Pearson correlation coefficient of 0.928 whereas altered probe sequences result in a low correlation coefficient of 0.554.

Overall, the internal technical replicates should however show the desirable property of a lower mean variance over each array. The first experiment with the reference probe design used 4,148 internal technical replicates for 164 genes whereas the second experiment with the new probe design had 1,368 internal technical replicates for 157 genes. The mean variance of the internal technical replicates for the reference probe design range from 4.27...4.7 for the biological sample of the *A. nidulans*-*S. rapamycinicus* interaction and *A. nidulans* wildtype, respectively (see table 3). The new probe design shows a lower mean variance of internal replicates, namely 3.55 for the wildtype and 3.69 for the interaction sample. This change corresponds to a reduction of the mean variance with a ratio of 0.76...0.86. The application of a Shapiro-Wilk test indicated a normal distribution of signal intensities with a p-value < 0.05. An F-test with a subsequent Holm-correction confirmed the significance of the change in variance. All adjusted p-values are below 0.05. The lower mean variance over each array of the new probe design is significant. In summary, the statistical analysis of experimental results obtained from technical replicates supports the applied method and shows that the new probe design yields more reliable results.

Table 3: Mean variance of technical replicates over each array

Sample/Replicates	Old design	New design	ratio
<i>A. nidulans</i> rep1	4.79	3.73	0.78
<i>A. nidulans</i> rep2	4.69	3.85	0.82
<i>A. nidulans</i> mean	4.70	3.55	0.76
<i>A. nidulans</i> + <i>S. rapamycinicus</i> rep1	4.00	3.76	0.94
<i>A. nidulans</i> + <i>S. rapamycinicus</i> rep2	4.51	4.12	0.91
<i>A. nidulans</i> + <i>S. rapamycinicus</i> mean	4.27	3.69	0.86

Mean variance of internal technical replicates which were included in the first microarray experiment using the reference probe design and in the second experiment using the optimized probe design. Two technical replicates were used for each of the biological samples (*A. nidulans* and *A. nidulans* + *S. rapamycinicus*). Mean variances and the ratio between both experiments are given for each replicate and for the mean of each biological sample.

3. Discussion

3.1 Probe Design Optimization

The reliability of used probe designs need to be checked whenever new genome annotations are available [10], [9]. For *A. nidulans* the evaluation of the given reference probe design showed this necessity as it contains many systematic errors and the possibility to cover a higher number of transcripts is not fully exploited. The approach combines both steps - the evaluation of reference probe designs and the design of new probes. Frequently, a probe design already exists and probe sequences that satisfy the design criteria do not need to be recalculated.

It is challenging to find the right software which applies all probe design criteria described above. The usage of a modular workflow which allows for the flexible integration of different design criteria helps to adjust the oligonucleotide design to the specific experimental requirements. This approach allows the integration of own probe design criteria and existing software. A similar workflow with different steps has been proposed and implemented in the tool Teolenn [11]. This framework was not considered due to the missing integration of re-evaluation of existing probe designs.

For the generation of probe candidates many different software tools have been proposed. In the proposed workflow we decided to use ArrayOligoSelector [14] which applies a large fraction of required design criteria and was recommended in an evaluation of custom microarray applications [2]. The tool chosen is interchangeable and should be orientated at the specific probe design requirements.

In this working example, hybridization are only considered if the alignment has a minimum sequence identity of 90% (see methods). This way, cross-hybridization can not be fully excluded because it was shown that it already occurs at a identity of 70% [6]. If the evaluation tool uses a more stringent cut-off, more probes are classified as invalid and more genes are not covered by any probe. The setting of this threshold is always a trade-off because the aim is to cover as many genes as possible while excluding cross-hybridizations. Hybridization with *S. rapamycinicus* transcripts was not checked because poly-dT-priming ensures that only eukaryotic RNA is amplified.

Due to the experimental objectives, the position of the probe and the GC content range were used as design criteria. The filtering for a narrow GC content range is a fast calculable filter criterion and effectively obtains a close melting temperature uniformity. The computational costly application of the Nearest-Neighbor Model [17] gives a more precise estimation of the melting temperature. A direct application of this methods for probe design is limited because it assumes that both nucleotide strands interact freely in a solution which is not the case for microarrays.

Generally, if more probe design criteria are applied more

probe candidates are excluded leading to a lower number of valid probe sequences. Overall, the used approach utilizes only a small set of all possible probe design criteria. Despite that, it was not possible to find a valid probe for 188 genes. Several factors contribute to this number of uncovered genes: If the gene annotation allows for transcripts which are shorter than the desired probe length or consist of highly repetitive sequence stretches, it is apparently not possible to find a valid probe sequence for them. In addition, a few transcripts share the same 3'-end, represent different splice variants, or are positioned within the same locus but on different strands. Finally, some sequences are at different loci, but have a high sequence similarity which may result from gene homology.

3.2 Impact of annotation databases

It is crucial to decide what structural genome annotation should be used as reference for the probe design. The reason are new genome assemblies and differences in the formal definition of the characteristics of a gene. Large fractions of the annotation of *Aspergillus nidulans* are done automatically with the help of bioinformatic tools. It is evident that with ongoing research the annotation of transcripts is subject to change. A large fraction of the oligonucleotide libraries can not be unambiguously matched to existing structural genome annotations [9]. The progress in laboratory research and, consequently, the related manual curation of genome annotations lead to more robust genome annotations.

3.3 Experimental Validation

The quality of the designed probes, and therefore the quality of the proposed approach, is eventually assessed by experimental validation. Probe sequences may be evaluated with spike-in experiments [18], self-hybridization experiments with the analysis of gene coverage [11], correlation of experimental data with probe design criteria [11], [12], experimental selection of probes [12], and the usage of internal technical replicates [19]. Without a transcriptome golden standard the impact of modifications can not be directly linked to the overall improvement of the array design. Spike-in experiments, Northern Blots, and qRT-PCR can only focus on a selection of chosen transcripts and are therefore not suited to assess a whole microarray probe design. Furthermore, it is not distinguishable which specific probe design criterion has an effect on the results because the criteria are mutually dependent. An altered probe sequence, for instance, does not only change the sequence similarity but also the physical properties of the probe and the hybridization. Nevertheless, it is necessary for an improvement of the design process.

In this study we used internal replicates to assess the quality of the new probes. Internal technical replicates allow to check for the performance of probes regardless of the experimental influences. A significant decrease of mean variances of internal replicates over each array was observed.

This shows that the probes have a higher signal reproducibility. The optimized microarray probe design is more reliable as it has been shown with the help of statistically significant lower mean variance of the internal technical replicates.

4. Materials and Methods

4.1 Probe design evaluation

The probe design from febit biomed GmbH (Heidelberg, Germany) was as used as 'reference probe design' (see GSE25266 and [15]). It was analyzed regarding the structural genome annotations from BROAD institute [20] (two different versions downloaded October, 10th 2008 and February, 18th 2010) and from CADRE [21] (downloaded February, 16th 2009). The annotation versions are referred to as 'BROAD2008', 'BROAD2010', and 'CADRE2009', respectively.

Probe sequences were aligned locally to the known corresponding transcripts with the help of FASTA (Parameters: expectation value 1.0, alignment type 0) [22]. The thermodynamic properties of each probe and the hybridization were calculated with the nearest-neighbor model [17], which is implemented in the freely available software MELTING (Parameters: '-Hdnadna -N0.2 -P0.0001 -Ksan98a') [23]. A probe is considered to match a transcript if there is at least one 16 basepairs long common subsequence and if both sequences share a sequence identity not less than 90%. Although literature suggests that hybridization already occurs at 70% sequence identity [6], a less stringent cut-off was applied. A stricter constraint dramatically decreases the number of valid probe sequences and prevents a full-genome probe design. All probes are finally classified into four classes. Probes that i) match perfectly, ii) cross-hybridize, iii) do not match any transcript, and iv) hybridize, but are not fully identical with the target sequence.

4.2 Generation of new probe candidates

New probe candidates were generated for genes where no perfect matching probe is given in the reference probe design. Different available algorithms could be applied for this step. In this study, we integrated the public available tool ArrayOligoSelector (Parameters: target GC percentage 48.0, length of oligonucleotides 50), number of oligos per gene 5) [14], which utilizes sequence similarity, a given GC content range, tests for low-complexity regions, and recognition of self-complementary sequences. The transcript sequence was trimmed to the first 600 basepairs to reduce computational time and to meet the probe design objective of placing the probe near the 3'-end. The generated probe candidates were checked with the help of the evaluation tool described above. This guarantees that new probe candidates meet the given cross-hybridization criterion and that systematic errors are avoided.

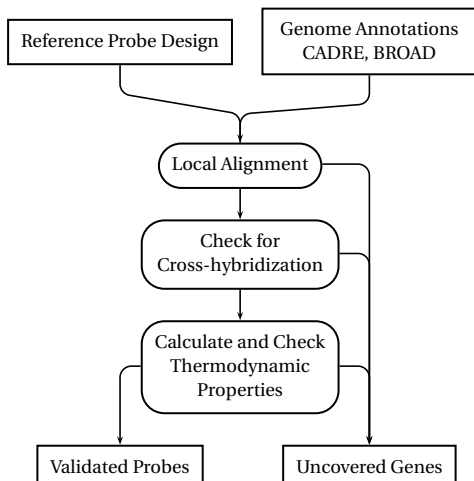


Fig. 1: **Schematic overview of evaluation process.** A reference probe design is locally aligned to selected genome annotation databases. Probes that cross-hybridize are filtered and thermodynamic properties of the hybridization are calculated for further assessment.

4.3 Selection of validated probes

The aim of covering the full genome of *Aspergillus nidulans* allows only to spot one oligonucleotide for each gene considering the given spotting density constraint. Validated newly generated probe candidates are preferred if they are positioned at the 3'-end of the transcript. If several probes exist within an overlapping close interval of 50bp, the following second design criterion is applied: Probes with a GC content closest to the mean GC content of the reference probe design are chosen if the difference to the mean is below 8%. This ensures similar thermodynamic properties of all probes. After the application of these criteria, at most one single probe candidate per gene remains.

4.4 Iterative softening of design criteria

We start with a transcript sequence ranging from the 3'-end to 600 basepairs. In order to get a better gene coverage, the used transcript sequence range was iteratively extended to 1,200 and 2,000 basepairs for the remaining uncovered genes. Finally, the stringent cross-hybridization criterion was relaxed for the remaining uncovered genes. Hence, probe candidates are even considered if they are vulnerable to cross-hybridization. Probe sequences were chosen manually for genes of high biological interest and without a valid probe candidate. The manually chosen sequences minimize the number of cross-hybridizations and fall within the narrow range of the desired mean GC content ($\pm 8\%$).

Merging the valid probes from the reference probe design with the selected new probe candidates resulted in the new and optimized probe design (see GSE25266 and figure 2).

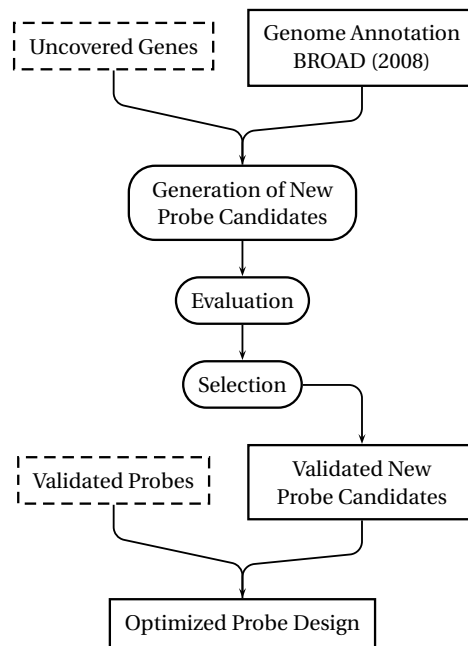


Fig. 2: **Workflow of probe design optimization.** New probe candidates are generated for the genes where there are no current valid probe sequences. Probe candidates are evaluated with the evaluation tool. If more than one probe candidate is valid, different selection criteria are applied to select the best optimized probe. The final new optimized probe design is obtained by the combination of these probe candidates with the validated probes from the reference probe design. (Dashed lines represent results from the evaluation of the reference probe design.)

In summary, in this study the following probe design criteria have been applied: cross-hybridization, sequence complexity, lack of self-binding, GC content, and position on reverse strand.

4.5 Experimental validation

Microarray raw data was obtained from *Aspergillus nidulans* - *Streptomyces rapamycinicus* interaction experiments [15]. The fungus was incubated over night in liquid *Aspergillus* minimal medium (AMM) and shifted into fresh medium. *Actinomycetes* were cultivated in M79 medium and 5 ml of the culture was added to 100ml AMM and both organisms were further incubated at 37°C. The reference culture is incubated without bacteria. After 3 h, each sample was split into two identical technical replicates and total-RNA was isolated using RiboPure-Yeast Kit (Applied Biosystems) according to the manufacturers instructions. cDNA synthesis, labeling and microarray measurements were done by febit biomed GmbH. In the first experiment, the reference probe design was used. The same samples were used for the second experiment where the new probe

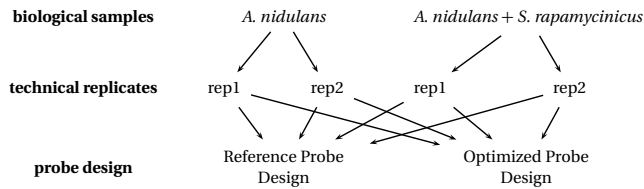


Fig. 3: **Schematic overview of Experimental Design.** In the first sample *A. nidulans* is cultivated without *S. rapamycinicus* and in the second sample it is co-cultivated with *S. rapamycinicus*. Each sample was split in two identical technical replicates. For each replicate a microarray experiment is performed with the reference and the new optimized probe design. The microarrays contain internal technical replicates that are used for the experimental validation.

design was utilized (see figure 3). All microarray data is compliant to the MIAME standard and can be accessed at GEO (<http://www.ncbi.nlm.nih.gov/geo/>) with the accession number GSE25266.

Both microarrays contain several internal technical replicates which can be used to assess the quality of microarray design. The comparability of both experiments is shown with the help of Pearson correlation coefficients of the signal intensities. The mean variance of the internal technical replicates were calculated over each array. The application of a Shapiro-Wilk tests for a normal distribution of signal intensities. The significance of the change in variances are evaluated by an F-test and a subsequent Holm-correction.

5. Conclusion

We proposed a workflow for the evaluation and optimization of existing microarray probe designs. This workflow is capable of integrating existing software and adjusting the probe design according to the experimental requirements. Exemplarily, this approach has been applied for a full-genome microarray for *Aspergillus nidulans* with the focus on avoiding systematic errors, especially cross-hybridizations. The reduction of cross-hybridization improves the reliability of the probe design which can be seen in a reduced mean variance of internal technical replicates over each array. We showed the high influence of different structural genome annotations on the design process. It is recommended to check for cross-hybridizations based on a current version of genome annotation prior to microarray data analysis.

Acknowledgment

This work was supported by the International Leibniz Research School for Microbial and Molecular Interactions (ILRS) and the Jena School for Microbial Communication (JSMC).

References

- [1] T. R. Hughes *et al.*, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer." *Nat Biotechnol*, vol. 19, no. 4, pp. 342–347, Apr 2001.
- [2] S. Lemoine *et al.*, "An evaluation of custom microarray applications: the oligonucleotide design challenge." *Nucleic Acids Res*, vol. 37, no. 6, pp. 1726–1739, Apr 2009.
- [3] Z. He *et al.*, "Empirical establishment of oligonucleotide probe design criteria." *Appl Environ Microbiol*, vol. 71, no. 7, pp. 3753–3760, Jul 2005.
- [4] S. Graf *et al.*, "Optimized design and assessment of whole genome tiling arrays." *Bioinformatics*, vol. 23, no. 13, pp. i195–i204, Jul 2007.
- [5] J. Mieczkowski *et al.*, "Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements." *BMC Bioinformatics*, vol. 11, p. 104, 2010.
- [6] M. Kane *et al.*, "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays," *Nucleic Acids Res.*, vol. 28, no. 22, pp. 4552–7, Nov 2000.
- [7] F. Ferrari *et al.*, "Novel definition files for human GeneChips based on GeneAnnot." *BMC Bioinformatics*, vol. 8, p. 446, 2007.
- [8] M. Dai *et al.*, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." *Nucleic Acids Res*, vol. 33, no. 20, p. e175, 2005.
- [9] P. B. T. Neerinx *et al.*, "Oligorap - an oligo re-annotation pipeline to improve annotation and estimate target specificity." *BMC Proc*, vol. 3 Suppl 4, p. S4, 2009.
- [10] H.-H. Chou, "Shared probe design and existing microarray reanalysis using PICKY." *BMC Bioinformatics*, vol. 11, p. 196, 2010.
- [11] L. Jourden *et al.*, "Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments." *Nucleic Acids Res*, vol. 38, no. 10, p. e117, Jun 2010.
- [12] F. Bidard *et al.*, "A general framework for optimization of probes for gene expression microarray and its application to the fungus *Podospora anserina*." *BMC Res Notes*, vol. 3, p. 171, 2010.
- [13] W. Vongsangnak and J. Nielsen, *Aspergillus: Molecular Biology and Genomics*. Caister Academic Press, Jan 2010, ch. Bioinformatics and Systems Biology of *Aspergillus*, pp. 61–84.
- [14] Z. Bozdech *et al.*, "Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray." *Genome Biol*, vol. 4, no. 2, p. R9, 2003.
- [15] V. Schroeckh *et al.*, "Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*." *Proc Natl Acad Sci U S A*, vol. 106, no. 34, pp. 14 558–14 563, Aug 2009.
- [16] A. A. Brakhage and V. Schroeckh, "Fungal secondary metabolites - strategies to activate silent gene clusters." *Fungal Genet Biol*, Apr 2010.
- [17] J. SantaLucia and D. Hicks, "The thermodynamics of DNA structural motifs." *Annu Rev Biophys Biomol Struct*, vol. 33, pp. 415–440, 2004.
- [18] I. V. Yang, "Use of external controls in microarray experiments." *Methods Enzymol*, vol. 411, pp. 50–63, 2006.
- [19] D. L. Leiske *et al.*, "A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray." *BMC Genomics*, vol. 7, p. 72, 2006.
- [20] J. E. Galagan *et al.*, "Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*." *Nature*, vol. 438, no. 7071, pp. 1105–1115, Dec 2005.
- [21] J. E. Mabey *et al.*, "Cadre: the Central *Aspergillus* Data REpository." *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D401–D405, Jan 2004.
- [22] W. R. Pearson, "Flexible sequence similarity searching with the FASTA3 program package." *Methods Mol Biol*, vol. 132, pp. 185–219, 2000.
- [23] N. L. Novère, "MELTING, computing the melting temperature of nucleic acid duplex." *Bioinformatics*, vol. 17, no. 12, pp. 1226–1227, Dec 2001.

EMMA: An EM-based Imputation Technique for Handling Missing Sample-Values in Microarray Expression Profiles.

Amitava Karmaker^{1*}, Edward A. Salinas², Stephen Kwek³

¹University of Wisconsin-Stout, Menomonie, Wisconsin 54751, USA

²Johns Hopkins University, Baltimore, Maryland 21218, USA

³Microsoft Corporation, Redmond, Washington 98052, USA

Abstract - Data with missing sample-values are quite common in many microarray expression profiles. The outcome of the analysis of these microarray data mostly depends on the quality of underlying data. In fact, without complete data, most computational approaches fail to deliver the expected performance. So, filling out missing values in the microarray, if any, is a prerequisite for successful data analysis. In this paper, we propose an Expectation-Maximization (EM) inspired approach that handles a substantial amount of missing values with the objective of improving imputation accuracy. Here, each missing sample-value is iteratively filled out using an updater (predictor) constructed from the known values and predicted values from the previous iteration. We demonstrate that our approach significantly outperforms some standard methods in terms of treating missing values, and shows robustness in increasing levels of missing rates.

Keywords: Microarray, Biological Data Mining, Missing Sample Value Estimation, EM Algorithm.

1 Introduction

Since the last decade, microarray technology has been applied as one of the widely used tools for gene expression profiling across various experimental conditions [1]. It generates enormous amounts of data that can be visualized and analyzed by computational tools. As a precondition for effective data analysis, microarray profiling data need to be preprocessed to ensure superior data quality. Due to intrinsic experimental settings and erroneous hybridization processes, very often microarray data contain missing values (probes), which deteriorate the subsequent analysis significantly. Studies found that on an average a microarray dataset contains ~5% missing values, and ~60% of the genes typically have at least one feature (sample) value missing [2]. Nevertheless, several data analysis algorithms, namely principal component analysis (PCA) [3], support vector machines (SVM) [4-6], singular value decomposition

(SVD) [7], artificial neural network (ANN) [8] etc., require fairly complete datasets to perform stably. In addition, unsupervised clustering (e.g. hierarchical clustering[9]) suffers from missing values while constructing clusters using distance measures. Moreover, because of higher expenses, sometimes replications of experiments are not often feasible. In order to ensure better analysis, incomplete microarray data are required to be preprocessed and reasonably complete.

In this paper, we propose an iterative technique to handle missing values in microarray data. Our method is inspired by the EM (Expectation Maximization) algorithm, which is widely used for missing value imputation in data preprocessing. In our algorithm, we try to implement an updater which will eventually estimate the most appropriate values replacing the imputed values in the preceding iterations. Unlike other methods, our technique can estimate the unknown values in the dataset and fill out the entries in the single dataset without incorporating "reference datasets". Empirically, we tested our method on six publicly available *Saccharomyces cerevisiae* (Yeast) microarray datasets and evaluated the performance measures. Our findings outperform other existing techniques considerably and tend to be quite robust in higher missing rates.

2 Related works

As we know, microarray is a large matrix of expression levels of genes (rows) under differential experimental conditions/derived from various samples (columns). The general hypothesis behind estimating missing values for microarray is to capture the inherent association among the underlying rows and columns, and infer new values for the missing ones taking this relationship into account as a whole. To preserve better correlations among the data values, sometimes gene expressions with missing values are discarded from further considerations. But it might not be an option if the most of the gene expressions have some of their values missing. Another simple way to deal with missing values is to impute average gene expression over the row [10]. Besides these, Troyanskaya et al. [11] proposed a

*Corresponding author

estimation method based on singular value decomposition (SVDimpute) [11]. Another Bayesian principal component analysis (BPCA) [12] based imputation algorithm was presented by Oba et al. [12], which assumes higher covariance among the gene expressions to estimate unknown values. These global imputation approaches are suitable for datasets with considerably large number of samples (~30) having strong global associations among them (e.g. temporal/time series datasets). On the other hand, there are quite a number of techniques for local missing value estimation. For example, k-nearest neighbor based KNNimpute [11], least square (LSImpute) [13], local least square (LLS) [14], etc. can handle relatively smaller datasets. To start with, these methods select neighborhood genes by Euclidean distance measures or Pearson's Correlations as required. The next step involves predicting the missing values based on selected genes' expression pattern. Still, these methods are error prone due to noise and insufficient samples. One of the shortcomings of all these methods is to integrate multiple datasets from diverse sources and consolidate those for analysis. Combination of datasets without proper relevance may critically degenerate the quality of neighborhood gene analysis, as pointed out for KNNimpute [12]. To this end, Tuikkala et al. [15] devised a gene ontology based technique GOImpute, which separates functionally related genes for further imputation. This method outperforms KNNimpute, but its performance is dependent on the availability of enough genes and accuracy of their annotations. Finally, we found another order statistics based approach called integrative Missing Value Estimation (iMISS) [16], which improves LSS algorithm and subsequently beats the GOImpute in terms of imputation accuracy.

**Table 1: Description of the test datasets
(Datasets denoted with (*) are time series)**

Dataset	No of genes/ instances	No. of samples/ attributes	Description
Diauxic*	5289	7	Metabolic transition from fermentation to respiration
Adaptive	3685	4	Evolutionary adaptability
Phosphate	5257	8	Polyphosphate metabolism
Alpha-factor*	4053	18	Yeast Cell cycle-regulation
Elutriation*	5192	14	
CDC15*	4833	13	

3 Methods and materials

3.1 Description of datasets used

To test our algorithms, we collected microarray profile data of *Saccharomyces cerevisiae* (Yeast) from the Princeton *Saccharomyces* Genome Database SGD Lite (<http://sgd-lite.princeton.edu/>), a publicly accessible yeast microarray data repository. Our datasets are composed of both time series and non-time series data. The first dataset (Diauxic) we selected is a time series, spotted cDNA microarray gene expression profiles dealing with metabolic shift from fermentation to respiration in yeast [17]. The second dataset (Adaptive) is on the study of adaptability of yeast and their differential gene expressions under diverse stress conditions [18]. Another dataset (Phosphate) reports the resulting gene expressions for phosphate accumulation and polyphosphate metabolism [19]. The rest of the datasets (Alpha-factor, Elutriation, CDC15) were created from Spellman time series cell-cycle datasets [20] based on the methods used for yeast cultures. These three datasets are all temporal and comprise higher sample dimensions. The characteristics of the datasets used are furnished in Table 1.

3.2 The Expectation-Maximization (EM) Algorithm

A popular way of dealing with missing values is to use the Expectation-Maximization (EM) algorithm introduced by Dempster, Laird and Rubin [21]. Here, the data source is assumed to be from a certain (mixture of) parametric model(s). EM algorithm tends to perform very well in parameter estimation. EM iteratively performs the following two steps.

Estimation (E) step: Estimate the parameters in the probabilistic model for the data source by using the known attribute-values and estimates of the missing attribute values obtained in the previous iteration of the M-step.

Maximization (M) step: Fill in the missing values to maximize the likelihood function that is refined in the E-step.

There are two drawbacks in using EM algorithm to fill up missing values. Firstly, it assumes that the data source comes from some parametric model (or a mixture of parametric models) with a finite mixture of Gaussian (k-Gaussians) being the most commonly used. Due to this assumption, most EM applications are applicable to numerical attributes only. Secondly, while EM can be proved to converge (with the appropriate parametric model assumption), the convergence process tends to be extremely slow. In particular, EM algorithm is useful when maximum likelihood estimation of a complete data model is relatively easy. Ouyang et al. [22] showed the use of microarray data for Gaussian mixture clustering and imputation. This research originated when we tried to investigate whether imputation accuracy can be improved by using EM algorithm in filling up

missing numerical attribute-values, which is literally appropriate for microarray data.

EMMA (η_{known} , η_{missing})

//Here $\eta_{\text{known}} = 0.0$, $\eta_{\text{missing}} = 1.0$, $H_i = \text{Linear Regressor}$

Initialize:
Fill the missing values using its mean (for continuous values).

Update:
Repeat the following two steps until convergence (k iterations).

E-step:
for each attribute x_i **do**
 Construct an updater H_i for x_i .

M-step:
for each attribute x_i **do**
 if x_i 's value was missing then
 $\eta \leftarrow \eta_{\text{missing}}$
 else
 $\eta \leftarrow \eta_{\text{known}}$

 $x_i \leftarrow \eta H_i(x) + (1 - \eta) x_i$

Output:
The final updaters for filling in the missing values.

Figure 1: Pseudo code of our algorithm, EMMA.

3.3 Our iterative technique for handling missing values

Instead of using common parametric models, we assume that the value of each attribute is somehow dependent on the values of the other attributes, which can be captured to a certain extent by simple linear regressor. In fact, this assumption is quite rational for analyzing microarray data derived from particular stand-alone (not distributed) experiment, be it temporal or not.

Inspired by EM approach, we propose an iterative algorithm, which is EMMA (EM on MicroArray) by the name. In the E-step, we build a linear regressor, which we call updater H_i , for each attribute x_i using the other attributes as input. In the M-step, we update the predicted value of those attributes based on these models constructed in the E-step as shown in Figure 1. The refined values are then used in the subsequent iterations to construct the updaters. Initially, if the sample value is missing, we use the mean values for first imputation. Because of the property of convergence at local maximum (saddle point) of EM algorithm, we need to start up with somewhat known (filled out) values. That is why we initiate with mean value imputation for the missing ones.

We continue this process iteratively until a certain number of iterations is reached or the attributes cease to

change much. The rate of refinement of certain sample value is moderated by the parameter η (eta). Our experiment sets η to 1.0 (specified by η_{missing}) for attributes (samples) with missing values, as they can be replaced with completely new values. On the other hand, η was valued at 0.0 (specified by η_{known}) for attributes (samples) without missing values to restrict drastic changes of values over iterations. These values of η are not fully optimized in order to prevent overfitting. Besides, we obtained outperforming results using these non-optimized parameters, and also values of η may be fine-tuned for better yields.

3.4 Experimental settings

To construct test datasets, we removed the gene with missing values from these datasets, so that we can calculate the accuracy of missing value imputations more precisely. The experiments use source code from the machine learning software WEKA[23]. Missing values are artificially added to the data sets to simulate randomized missing values. To introduce $m\%$ missing values per attribute x_i in a data set of size n , we randomly selected mn instances and replaced its x_i -value with an ‘‘unknown’’ (In WEKA, missing values are denoted as ‘?’) label. Missing values were added in the original data sets from which both training data sets and test data sets were generated. In each set of experiment, we used increasing levels of ‘missingness’ - missing rate: $m = 1\%$, 5% , 10% , 20% , 25% , and 50% . We find that often at $m \geq 10\%$, the majority of the instances (genes) have some missing values, while at $m \geq 25\%$, all instances (genes) have some missing values. Moreover, as ours is an iterative approach, we recorded the performance metrics at increasing number of iterations, $T = 1, 2, 5, 10, 15, 20, 25, 50$ respectively.

3.5 Performance Evaluation

In order to validate the efficacy of our imputation method, we used Root Mean Squared Error (RMSE)[24, 25] (See Equation 1) performance metric, which estimates the relative closeness of the predicted and actual values. To minimize the variances in the RMSE measures, we created as many as ten datasets at same missing level ($m\%$), and finally took average of all ten performance figures. There are other formulations of RMSE than that in Equation 1. The reason why we choose this expression because in the ideal case (null imputation algorithm), this RMSE measure will stand out as zero (null). This ensures that we can use this metric to compare same datasets using various algorithms. The closer the predicted (estimated) values to the actual values, the lesser the RMSE values are, resulting in the values of RMSE ~ 0.0 for almost correct prediction.

$$RMSE = \sqrt{\frac{\text{mean}\{\bar{Y}_{\text{predicted}} - \bar{Y}_{\text{actual}}\}^2}{\text{mean}\{\bar{Y}_{\text{actual}}\}^2}} \quad (1)$$

Table 2: RMSE measures of six datasets at different iterations for different missing rates ($m\%$)

Missing rate (m)	1%	5%	10%	20%	25%	50%
# of Iterations (T)	Diauxic					
1	0.05617	0.12377	0.18004	0.26307	0.30318	0.45581
2	0.05598	0.12312	0.17718	0.25392	0.29049	0.42738
5	0.05591	0.12305	0.17669	0.25264	0.28854	0.43352
10	0.05590	0.12307	0.17679	0.25288	0.28897	0.45119
20	0.05590	0.12307	0.17681	0.25296	0.28914	0.45561
50	0.05590	0.12307	0.17681	0.25296	0.28915	0.45695
	Adaptive					
1	0.06932	0.15188	0.21863	0.31998	0.35122	0.50188
2	0.06924	0.15117	0.21685	0.31247	0.34323	0.48526
5	0.06922	0.15112	0.21699	0.31202	0.34435	0.49516
10	0.06921	0.15112	0.21700	0.31203	0.34454	0.50235
20	0.06921	0.15112	0.21700	0.31203	0.34454	0.50522
50	0.06921	0.15112	0.21700	0.31203	0.34454	0.50554
	Phosphate					
1	0.07581	0.16862	0.23738	0.33168	0.37086	0.51028
2	0.07585	0.16858	0.23661	0.32842	0.36520	0.49984
5	0.07588	0.16864	0.23704	0.32938	0.36700	0.52675
10	0.07588	0.16864	0.23713	0.32951	0.36745	0.55547
20	0.07588	0.16864	0.23714	0.32952	0.36748	0.56368
50	0.07588	0.16864	0.23714	0.32952	0.36748	0.56678
	CDC15					
1	0.06420	0.15783	0.22104	0.31270	0.35536	0.51542
2	0.06374	0.15700	0.21827	0.30564	0.34639	0.49566
5	0.06370	0.15689	0.21771	0.30640	0.34725	0.52152
10	0.06370	0.15689	0.21767	0.30644	0.34780	0.55042
20	0.06370	0.15689	0.21767	0.30638	0.34780	0.55335
50	0.06370	0.15689	0.21767	0.30637	0.34779	0.56756
	Alpha-Factor					
1	0.04435	0.11903	0.19110	0.29814	0.34277	0.50708
2	0.04814	0.11370	0.18466	0.29244	0.33753	0.50076
5	0.04720	0.11530	0.18688	0.29501	0.34183	0.50746
10	0.04750	0.11553	0.18666	0.29532	0.33986	0.51684
20	0.04718	0.11592	0.18554	0.29375	0.33845	0.52423
50	0.04699	0.11596	0.18471	0.29380	0.33879	0.53709
	Elutriation					
1	0.05573	0.11865	0.17985	0.27823	0.31824	0.48531
2	0.05529	0.11899	0.17841	0.27091	0.30842	0.46726
5	0.05528	0.12042	0.17588	0.27105	0.31368	0.47391
10	0.05528	0.11941	0.17756	0.27294	0.31711	0.47882
20	0.05528	0.11949	0.17623	0.27251	0.31790	0.47941
50	0.05528	0.11923	0.17573	0.27300	0.31804	0.48761

4 Results and Discussion

As we mentioned before, we took the performance measures (in RMSE) at different iteration counts for increasing missing value levels. The findings are reported in the Table 2. Because EMMA is based on an iterative approach, we observed the change of accuracy with the number of iterations the underlying updater H_i is called upon

for imputation. For almost all cases, it seems that RMSE is higher at the very first turn, and lessens within a few iterations ($T = \sim 2-5$). These values remain fairly steady throughout higher iterations ($T > 10$). This happens due to the property of the regressor we used. Basically, linear regressor here tends to fit the data points within first few runs, and adjusted regressor does not rectify too much in the following turns. Moreover, the RMSE values across different missing rates remain

relatively robust, and do not swing erratically with the proportion of missing values induced in the datasets. For example, the RMSE counts for Diauxic dataset ($T = 10$) are 0.05590, 0.15112, 0.12307, 0.17679, 0.25288, 0.28897, 0.45119 at the missing rate of 1%, 5%, 10%, 20%, 25%, 50% respectively. As expected, as the missing rate increases, the performance of our technique deteriorates (see Table 2), but the degree of imputation accuracy does not fall as much as the magnitude of missing rates. This suggests that our algorithm can handle higher number of missing values efficiently.

To evaluate the performance of our method comparing to other extant methods, we obtained the RMSE numbers of those methods on the same datasets. As reported by Hu et al.[16], for Diauxic, Adaptive and Phosphate dataset, the RMSE measures (10% missing rate) of KNNimpute [11] and LLS based method [14] are $\sim 0.6-0.8$; while their integrative approach improved those numbers by decreasing almost ~ 0.05 . On the other hand, the error numbers for our algorithm at the same settings are in the range of $\sim 0.17-0.24$ ($T = 10$), which is an improvement over the aforementioned methods by a significant margin. Besides, we maintain a competitive edge against these methods at higher missing rates (See Figure 2~4).

For a dataset with m rows (genes), n columns (samples) and k iterations, the computational complexity of our algorithm is approximately $O(mn^2k)$. So the running time scales up with the dimensions of the data matrix for the microarray. Instead of using linear regressor, we can also use any other hypothesis that can handle numeric class values (e.g. decision stump, multi-layer perceptrons, etc.).

All the datasets we used here were originated from same experiments using common microarray platform (spotted cDNA microarray). One of the advantages of our method is we do not need to integrate expression profiles from different experimental settings (cDNA vs. Affymetrix). Still, the performance of our algorithm depends on the variance of the datasets. Also, to infer linear regression, the dataset has to contain enough attributes (columns) to fit on the hypothesis.

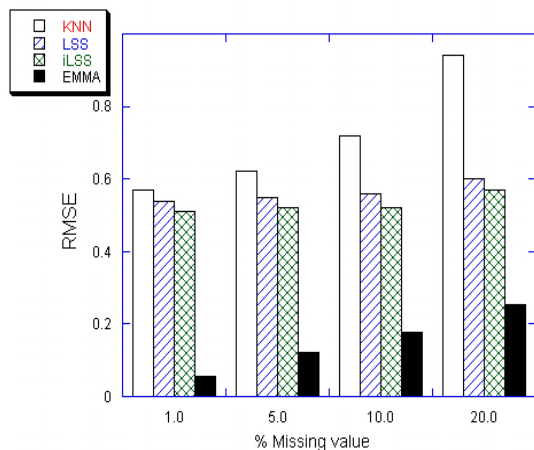


Figure 2: Comparison of performances for KNN, LSS, iLSS and EMMA for Diauxic dataset

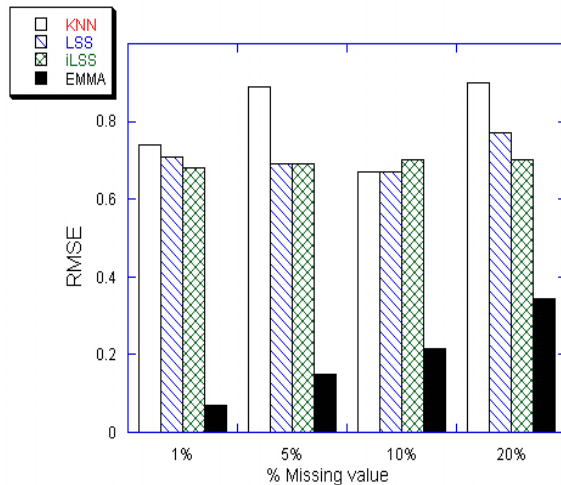


Figure 3: Comparison of performances for KNN, LSS, iLSS and EMMA for Adaptive dataset

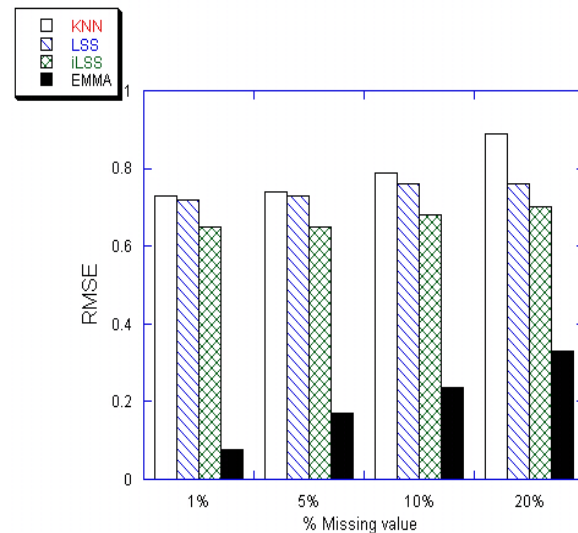


Figure 4: Comparison of performances for KNN, LSS, iLSS and EMMA for Phosphate dataset

5 Conclusions

To summarize, we present a novel method for treating missing sample values of genes in microarray data. Our technique is based on popular EM algorithm, and it far surpasses other existing state-of-the-art techniques in terms of imputation accuracy. In fact, being iterative in nature, our algorithm successfully grasps the innate relationships among the samples in subsequent runs, and stabilizes after the imputed and known values converge at some point. We

validated the strength of our algorithm by applying it to estimate missing values in both temporal and non-temporal benchmark datasets. In future, we expect to extend this technique to handle noise in microarray data in the preprocessing step.

6 References

- [1] D. B. Allison, *et al.*, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet*, vol. 7, pp. 55-65, Jan 2006.
- [2] A. G. de Brevern, *et al.*, "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering," *BMC Bioinformatics*, vol. 5, p. 114, Aug 23 2004.
- [3] S. Raychaudhuri, *et al.*, "Principal components analysis to summarize microarray experiments: application to sporulation time series," *Pac Symp Biocomput*, pp. 455-66, 2000.
- [4] V. Vapnik, "The Nature of Statistical Learning Theory," vol. Springer-Verlag, New York, 1995.
- [5] M. P. Brown, *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc Natl Acad Sci U S A*, vol. 97, pp. 262-7, Jan 4 2000.
- [6] G. Valentini, "Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles," *Artif Intell Med*, vol. 26, pp. 281-304, Nov 2002.
- [7] O. Alter, *et al.*, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc Natl Acad Sci U S A*, vol. 97, pp. 10101-6, Aug 29 2000.
- [8] J. Huang, *et al.*, "Clustering gene expression pattern and extracting relationship in gene network based on artificial neural networks," *J Biosci Bioeng*, vol. 96, pp. 421-8, 2003.
- [9] M. B. Eisen, *et al.*, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci U S A*, vol. 95, pp. 14863-8, Dec 8 1998.
- [10] A. A. Alizadeh, *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-11, Feb 3 2000.
- [11] O. Troyanskaya, *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp. 520-5, Jun 2001.
- [12] S. Oba, *et al.*, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, pp. 2088-96, Nov 1 2003.
- [13] T. H. Bo, *et al.*, "LSimpute: accurate estimation of missing values in microarray data with least squares methods," *Nucleic Acids Res*, vol. 32, p. e34, 2004.
- [14] H. Kim, *et al.*, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, pp. 187-98, Jan 15 2005.
- [15] J. Tuikkala, *et al.*, "Improving missing value estimation in microarray data with gene ontology," *Bioinformatics*, vol. 22, pp. 566-72, Mar 1 2006.
- [16] J. Hu, *et al.*, "Integrative missing value estimation for microarray data," *BMC Bioinformatics*, vol. 7, p. 449, 2006.
- [17] J. L. DeRisi, *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680-6, Oct 24 1997.
- [18] T. L. Ferea, *et al.*, "Systematic changes in gene expression patterns following adaptive evolution in yeast," *Proc Natl Acad Sci U S A*, vol. 96, pp. 9721-6, Aug 17 1999.
- [19] N. Ogawa, *et al.*, "New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis," *Mol Biol Cell*, vol. 11, pp. 4309-21, Dec 2000.
- [20] P. T. Spellman, *et al.*, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol Biol Cell*, vol. 9, pp. 3273-97, Dec 1998.
- [21] A. P. Dempster, *et al.*, "Maximum-likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. B39, pp. 1-38, 1977.
- [22] M. Ouyang, *et al.*, "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics*, vol. 20, pp. 917-23, Apr 12 2004.
- [23] E. Frank, *et al.*, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, pp. 2479-81, Oct 12 2004.
- [24] R. Jornsten, *et al.*, "DNA microarray data imputation and significance analysis of differential expression," *Bioinformatics*, vol. 21, pp. 4155-61, Nov 15 2005.
- [25] R. Jornsten, *et al.*, "A meta-data based method for DNA microarray imputation," *BMC Bioinformatics*, vol. 8, p. 109, 2007.

Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays

C. Hummert¹, F. Mech¹, F. Horn¹, M. Weber¹, S. Drynda², U. Gausmann³, R. Guthke¹

¹Research Group: Systems Biology / Bioinformatics, Hans Knöll Institute Jena

²Clinic of Rheumatology, Otto von Guericke University Magdeburg, Medical Faculty

³Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)

Abstract—Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. Nevertheless, the quality of the technology is still capable for further improvements. One of the main problems is cross-hybridization of the transcripts to non-corresponding probes on the array by unspecific binding.

Four different Affymetrix GeneChip arrays are analyzed, namely the Human Genome arrays HG-U133A, HG-U133B, HG-U133 Plus 2.0 and the Mouse Genome 430 2.0 array. It is shown that putative cross-hybridizations are common for the examined arrays (e.g., 45 % of all probes for the U133A). Furthermore, a considerable amount of probes does not match the annotated transcript correctly. A new set of CDFs is created avoiding putative cross-hybridization completely. It is compared with three other CDFs (Affymetrix, Dai *et al.*, Ferrari *et al.*) with the help of correlation between microarray and qRT-PCR results for two datasets. The newly created and the Ferrari CDFs perform significantly better than the original Affymetrix CDFs. The new CDFs are available as R-packages at <http://www.sysbio.hki-jena.de/software> and have been submitted to BioConductor.

Keywords: microarrays, unspecific binding, cross-hybridization, Chip Definition Files

1. Background

Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. They are applied to measure changes in expression levels for thousands of genes simultaneously. Until 2011, more than 20,000 measurement series based on microarray technology have been published in public repositories like NCBI's Gene Expression Omnibus.

Nevertheless, the quality of the technology is still capable for further improvements [1], [2]. Several studies tried to compare data derived from different types of arrays and showed a rather poor consistency [3], [4]. Although microarrays are commonly used, this is a daunting problem. In addition, although there has been extended work on this field [5], there is still a lack of standardized experimental protocols among different laboratories [6].

The main problem of microarray analysis is unspecific binding of transcripts by cross-hybridization. This means that RNA fragments hybridize to a probe which is not designed for this gene. It was shown that fragments longer than 8 nucleotides are able to hybridize and that cross-hybridization can emerge from alignments ranging from 10 to 16 nucleotides. Further, the 5'-ends were found to cross-hybridize more likely than the 3'-ends [7].

Unspecific binding may lead to false-positive and false-negative results following in incorrect hypotheses about gene expression [8], [9]. Affymetrix, a technology widely used [10], accounts for the influence of cross-hybridization by introducing internal controls: each probepair comprises a Perfect Match (PM) and a Mismatch (MM) probe which are statistically evaluated [11]. Unfortunately, this procedure cannot solve the problem of cross-hybridization completely [12] and further refinements are suggested [13]. For example, Wu *et al.* [7] stated that the MM probes can also cross-hybridize, even though by another mechanism as the PM probes. Therefore, they recommended ignoring the MM probes.

Generally, expressed transcripts are represented on the array by a series of probepairs called probesets. The signal intensities are summarized to a single value per probeset. A large number of single transcripts are represented by multiple probesets. Multiple probesets representing the same gene are expected to show similar fold changes calculated from the signal intensities of the hybridized samples. However, this is in fact not the case [14], [15], [16]. This problem arises from single probes in the probeset which are capable of cross-hybridization. Ways to deal with this problem is either a probe-based analysis, leaving out the probe-to-probeset summarization step [17], [18], or the composition of the probesets could be improved by setting up alternative Chip Definition Files (CDFs) based on information contained in different sequence databases. For example, the group of Ferrari *et al.* [19] created a set of custom CDFs based on the GeneAnnot database [20]. In these CDFs the probesets that match the same gene were merged into one probeset. Hence, the existence of more than one probeset per gene was eliminated, avoiding discordant expression signals for the same transcript.

Another set of custom CDFs relying on a broad repertoire

of databases like RefSeq or Unigene has been created by the group of Dai *et al.* [21]. Probesets matching the same gene were merged, but remained divided if they were able to discriminate different isoforms of a gene. Probes causing cross-hybridizations were removed from the new probesets, but the filter had been not very strict.

Several groups dealt with the question of the minimum probeset size [19], [21]. For example, the group of Lu *et al.* [22] sets the minimum probeset size to 4 probes because smaller probesets result in high error rates. In this study the minimum probeset size was set to 4 [19], [21]. From these new probesets custom CDFs and the corresponding Bioconductor libraries for Affymetrix GeneChips were created.

In the work presented here, a new set of CDFs is introduced avoiding putative cross-hybridization completely. These CDFs are compared with those from Affymetrix, Ferrari, and Dai by validation of the respective microarray results using qRT-PCR for two different datasets.

2. Results

Four different Affymetrix GeneChip arrays are analyzed, namely the HG-U133A, HG-U133B, HG-U133 Plus 2.0 designed for human samples, and the Mouse Genome 430 2.0 array. For the detection of putative cross-hybridizations, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using *blastn* [23] as described in the methods section.

The GeneChip HG-U133A consists of 22,283 probesets, each of 11–20 probepairs and 247,937 probepairs in total. Additional 1,155 probepairs are controls and are furthermore ignored. About 44 % of the PM probes (109,245) match exactly one single gene. 11 % of the probes (26,159) do not match any annotated gene. 45 % of the probes (112,533) match more than one gene and thus have cross-hybridization potential.

Furthermore, the direction of the probes was analyzed. Normally, sense strand RNA fragments are expected, although there are some loci in the human genome [24], as well as in the mouse genome [25], where both sense and antisense strands are transcribed. However, mixing up probes detecting sense or antisense strands in one single probeset could cause wrong expression results. Here, only probes matching the sense strand are considered as correct. For the U133A microarray all probes match the sense strand.

The GeneChip HG-U133B consists of 22,645 probesets, each of 11–20 probepairs and 249,491 probepairs in total. Again, there are additional probesets containing more than 11 probes as controls and are ignored (1,100). About 35 % of the probes (87,067) are found to match exactly one gene. 2 % of the probes (5,453) match more than one gene, so they possibly cross-hybridize, 5 % of the probes (12,805) match at least one gene but in the wrong direction (antisense

direction) and no gene in the sense direction, and 58 % of the probes (144,166) do not match any annotated gene.

The GeneChip HG-U133 Plus 2.0 consists of 54,675 probesets and 604,247 probepairs. Like in the other arrays, additional probesets containing more than 11 probes are controls and are discarded. Here, 37 % of the remaining probes (221,821) match exactly one gene, 23 % of the probes (141,146) match more than one gene, 11 % of the probes (65,327) match at least one gene but in the wrong direction (antisense direction) and no gene in the sense direction, and 29 % of the probes (175,953) do not match any annotated gene.

The Mouse Genome 430 2.0 array consists of 45,036 probesets and 496,457 probepairs. About 52 % of the counted probes (257,331) match exactly one gene and 5 % of the probes (27,112) match more than one gene. About 1 % of the probes (4,661) match genes only in the wrong direction and 42 % of the probes (207,353) do not match any annotated gene.

Nearly all Affymetrix probesets contain at least one probe which has cross-hybridization potential. In fact, for the HG-U133 Plus 2.0 Chip about 65 % of all probesets include more cross-hybridizing probes than non-ambiguous ones.

All probes matching exactly one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes, that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. Only the good probes are used to create the new CDFs as described in the methods chapter. Accordingly, for the HG-U133A microarray originally measuring 14,500 genes by 22,283 probesets the newly created CDF contains 12,400 probesets representing 12,400 genes. For the HG-U133 Plus 2.0 the number of probesets is reduced from 54,675 (representing 38,500 genes) to 18,800 (representing 18,800 genes). The HG-U133B comprises 22,645 probesets measuring the expression of 18,400 genes. Here, the number of probesets is reduced to 6,500 matching 6,500 transcripts. The Mouse 430 2.0 microarray consists of 45,036 probesets for 39,000 genes. With the new CDF there are 16,400 probesets matching 16,400 genes. Hence, the number of identifiable genes is reduced in order to achieve a higher specificity of the probesets. The result for the HG-U133 Plus 2.0 is in good agreement to the results of Barnes *et al.* [26], who used BLAT and the Golden Path database and achieved a number of 17,143 genes that can be measured.

Small probesets lead to higher error rates and result in lower statistical significance. In the Affymetrix CDFs the size is 11 for nearly all probesets, but in the newly created probesets the size is not fixed. Some probesets are smaller than those from Affymetrix due to the removal of the problematic probes. However, many probesets increase in size due to useful probes on the array that have not been used for the matching gene before and probesets measuring

the same gene being merged. For example, for the HG-U133 Plus 2.0 the mean probeset size increases from 11 to 17.

For the validation of all CDFs two test datasets are chosen: (i) the Etanercept (ETC) and (ii) the MAQC dataset. The first of the two datasets is derived from a study analyzing the effect of the TNF- α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points [27]. It is a typical dataset that arises in medical studies and is rather representative. One Affymetrix HG-U133A array experiment was performed for each time point. The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. The subset of the 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here.

qRT-PCR results are considered to reflect the real transcript concentrations with higher reliability than those determined by microarrays. Therefore, qRT-PCR experiments are regarded as a 'gold standard' for chip analyses [29], [30]. The Pearson correlation coefficient (PCC) of the microarray and the qRT-PCR data is computed for each gene using the different CDFs.

For the Etanercept dataset we performed qRT-PCR experiments for 16 genes. In total, this dataset now contains results from 51 microarrays and 816 qRT-PCR experiments. In addition, the genes with qRT-PCR data in both records are analyzed in more detail.

The performance of these CDFs were compared: the original Affymetrix CDFs (A), the two alternative CDFs of Ferrari *et al.* (F) [19] and Dai *et al.* (D) [21], and the new CDFs (H) presented here. The CDFs from Ferrari, using the GeneAnnot database, contain merged probesets (see background chapter), and cross-hybridization was not considered. The group of Dai offers a broad spectrum of different CDFs based on different databases. The one using RefSeq is chosen for comparison because it corresponds best to the new CDFs, using RefSeq as well. In the Dai CDFs different probesets matching a single gene are combined, although there are exceptions for genes comprising different isoforms. A check for cross-hybridization is also included. However, it applies a different algorithm than the new CDFs and the filter is much less strict.

For the probe to probeset summarization step two algorithms are used as described in the methods section: (i) the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and (ii) the Affymetrix Microarray Suite MAS5 [32]. These were compared repeatedly, but it is difficult or even impossible to decide which of the both algorithms performs better in any case [33], [34], [35].

For the Etanercept dataset, the mean correlation coefficient of all 16 genes for the Affymetrix CDF is 0.61 using the robust multi-array analysis algorithm (RMA) and 0.60 using the Affymetrix Microarray Suite MAS5. These

values include 31 probesets in total matching these 16 genes according to the Affymetrix annotation file. If only the best correlating probeset for each gene is considered, the average correlation coefficient increases to 0.73 for RMA and 0.71 for MAS5. However, this value is more of theoretical interest because the knowledge which probeset will perform best is gained not until the qRT-PCR experiments and correlation analysis is finished. On average, the incorporated probesets contain 5.58 putative cross-hybridizations calculated by BLAST (4.47 including only the best performing probesets).

The Dai CDF contains 23 probesets for the 16 genes of the Etanercept dataset. Their mean correlation coefficient increases to 0.67 for both RMA and MAS5 compared to the 0.60 using the Affymetrix CDF. Considering the best correlating Dai probesets only, the values further increase to 0.73 for RMA and 0.69 using MAS5. The mean size of the Dai probesets increases to 20.59 probes containing 8.82 putative cross-hybridizations. This number changes to 4.71 if normalized to a probeset size of 11. Here, normalization means the number of putative cross-hybridizations calculated for a hypothetical Dai probeset size of 11. Considering only the best Dai probesets, the number of putative cross-hybridizations decreases to 7.88 on average.

For the Ferrari CDF, the mean correlation coefficient equals 0.73 for RMA and 0.69 using MAS5 on average. The mean probeset size increases to 19.56, harboring 10.81 possible cross-hybridizations (6.07 if normalized).

Using the new CDF the mean correlation coefficient amounts to 0.72 for RMA and 0.68 for MAS5. The mean probeset size decreases to 10.25 with no cross-hybridizations at all. The detailed results are shown in the table below:

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset-size
TNF	A: 207113_s_at	0.88	0.85	N/A	8	11
	D: NM_000594_at	0.88	0.85	N/A	8	11
	F: GC06P031652_at	0.88	0.85	N/A	8	11
	H: gi_25952110	0.86	0.81	N/A	0	3
IL1B	A: 205067_at	0.95	0.90	0.37	6	11
	A: 39402_at	0.95	0.87	0.82	6	16
	D: NM_000576_at	0.96	0.89	0.74	12	27
	F: GC02M113303_at	0.96	0.89	0.74	12	27
H: gi_27894305	0.95	0.88	0.86	0	15	
IL6	A: 205207_at	0.69	0.71	0.81	3	11
	D: NM_000600_at	0.69	0.71	0.81	3	11
	F: GC07P022732_at	0.69	0.71	0.81	3	11
	H: gi_10834983	0.65	0.72	0.71	0	8
IL8	A: 202859_x_at	0.88	0.81	0.90	6	11
	A: 211506_s_at	0.86	0.73	0.98	6	11
	D: NM_000584_at	0.88	0.73	0.96	12	22
	F: GC04P074845_at	0.88	0.73	0.96	12	22
H: gi_28610153	0.89	0.73	0.95	0	10	
IL1RN	A: 212657_s_at	0.75	0.87	N/A	2	11
	A: 212659_s_at	0.77	0.84	N/A	4	11
	A: 216243_s_at	0.75	0.86	N/A	6	11
	A: 216244_s_at	0.13	0.07	N/A	4	11
	A: 216245_at	0.21	0.11	N/A	10	11
	D: NM_173841_at	0.80	0.88	N/A	12	33
	D: NM_000577_at	0.80	0.88	N/A	12	33
	D: NM_173842_at	0.80	0.88	N/A	12	33
	D: NM_173843_at	0.84	0.86	N/A	15	42
F: GC02P113591_at	0.83	0.86	N/A	16	44	
H: gi_27894315	0.78	0.88	N/A	0	23	
ICAM1	A: 202637_s_at	0.63	0.73	0.97	7	11
	A: 202638_s_at	0.62	0.72	0.98	4	11
	A: 215485_s_at	0.71	0.73	0.94	3	11
	D: NM_000201_at	0.70	0.76	0.99	14	33
	F: GC19P010247_at	0.70	0.77	0.99	14	33
H: gi_4557877	0.72	0.74	0.97	0	20	
SOD2	A: 215078_at	0.25	0.35	N/A	10	11

Continued on next page

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset-size
	A: 215223_s_at	0.15	0.28	N/A	7	11
	A: 216841_s_at	0.18	0.39	N/A	3	11
	A: 221477_s_at	0.32	0.44	N/A	10	11
	D: NM_001024466_at	0.16	0.33	N/A	6	12
	D: NM_000636_at	0.19	0.37	N/A	10	22
	D: NM_001024465_at	0.16	0.33	N/A	6	13
	F: GC06M160020_at	0.20	0.36	N/A	20	33
	H: gi_67782304	0.20	0.39	N/A	0	12
TRAF1	A: 205599_at	0.61	0.50	0.88	6	11
	D: NM_005658_at	0.61	0.50	0.88	6	11
	F: GC09M122704_at	0.61	0.50	0.88	6	11
	H: gi_53759116	0.59	0.47	0.89	0	5
ZFP36	A: 201531_at	0.84	0.86	N/A	5	11
	A: 213890_x_at	-0.01	-0.46	N/A	8	11
	D: NM_003407_at	0.84	0.86	N/A	5	11
	F: GC19P044589_at	0.84	0.86	N/A	5	11
	H: gi_141802261	0.85	0.82	N/A	0	6
PTGS2	A: 204748_at	0.91	0.71	0.97	4	11
	D: NM_000963_at	0.91	0.71	0.97	4	11
	F: GC01M184907_at	0.91	0.71	0.97	4	11
	H: gi_4506264	0.89	0.72	0.95	0	9
TNFAIP3	A: 202643_s_at	0.78	0.82	0.97	4	11
	A: 202644_s_at	0.87	0.85	0.93	6	11
	D: NM_006290_at	0.82	0.83	0.96	10	22
	F: GC06P138230_at	0.82	0.83	0.96	10	22
	H: gi_26051241	0.80	0.82	0.98	0	13
DUSP2	A: 204794_at	0.75	0.66	N/A	5	11
	D: NM_004418_at	0.75	0.66	N/A	5	11
	F: GC02M096230_at	0.75	0.66	N/A	5	11
	H: gi_12707563	0.74	0.60	N/A	0	6
ADM	A: 202912_at	0.80	0.67	0.92	5	11
	D: NM_001124_at	0.80	0.67	0.92	5	11
	F: GC11P010283_at	0.80	0.67	0.92	5	11
	H: gi_4501944	0.82	0.67	0.94	0	6
CROP	A: 203804_s_at	0.44	0.56	N/A	5	11
	A: 208835_s_at	0.43	0.36	N/A	5	11
	A: 220044_x_at	0.43	0.44	N/A	4	11
	D: NM_016424_at	0.49	0.50	N/A	13	32
	D: NM_006107_at	0.49	0.45	N/A	13	30
	F: GC17P046151_at	0.48	0.48	N/A	14	33
	H: gi_52426741	0.46	0.47	N/A	0	17
NFKBIA	A: 201502_s_at	0.81	0.73	N/A	4	11
	D: NM_020529_at	0.81	0.73	N/A	4	11
	F: GC14M034940_at	0.81	0.73	N/A	4	11
	H: gi_10092618	0.82	0.77	N/A	0	7
JUNB	A: 201473_at	0.44	0.44	0.94	7	11
	D: NM_002229_at	0.44	0.44	0.94	7	11
	F: GC19P012763_at	0.44	0.44	0.94	7	11
	H: gi_44921611	0.54	0.44	0.73	0	4
∅	all Affymetrix	0.61	0.59	0.88	5.58	11.16
	best Affymetrix	0.73	0.71	0.92	4.47	11.00
	Dai	0.67	0.67	0.91	8.82	20.59
	best Dai	0.73	0.69	0.91	7.88	18.69
	Ferrari	0.73	0.69	0.91	10.81	19.56
	Hummert	0.72	0.68	0.89	0.00	10.25

Evaluating the PM and MM probes statistically, the MAS5 software assigns 'present', 'absent' or 'marginal' to each expression value, and Affymetrix recommends to use only the 'present' detection call for further analysis. Following this recommendation and using only those results for the correlation analysis that are marked as 'present' the mean correlation coefficient increases from 0.59 to 0.66 (0.74 including only the best performing probesets). Hence, incorporating the Affymetrix detection call indeed improves the correlation, but using alternative CDFs is still better than using the Affymetrix probesets and the detection call.

Analyzing the MAQC reference dataset using the RMA suite, the results are almost in accordance with those of the Etanercept data described above. The mean correlation coefficient for all 1,000 genes is 0.47 for the Affymetrix CDF (0.71 incorporating only the best probeset for each gene). Using the Dai CDF, the mean correlation increases to 0.63 (0.64 for the best probesets). With the Ferrari and the new CDF the mean correlations are 0.63 and 0.58, respectively. The detailed results for all MAQC genes can be downloaded.

Discussion

Results from microarray experiments contain considerably high error rates [36]. Due to error propagation, it is of

particular importance to minimize errors in the beginning of the analysis chain [37]. Therefore, especially the pre-processing of the chip data has to be done as accurate as possible. Many efforts were spent on these problems before [38], such as the notable results of the 'Golden Spike Project' [6]. The question which statistical method should be adequately chosen is even more complicated if experimental data from different laboratories are incorporated in one single analysis [39].

For microarray analyses algorithms are essential which combine the 11-20 probepair intensities for a given gene and define a measure of expression that represents the amount of the corresponding mRNA species. In this study, two of these algorithms are compared, the robust multi-array analysis algorithm (RMA) and the Affymetrix Microarray Suite MAS5. Applying both algorithms to the Etanercept dataset RMA outperforms MAS5 on average. Other studies revealed similar results. However, their performance is assumed to be dependent on the actual dataset [40]. In fact, normalisation steps are applied after the probe to probeset summarization. Some of these steps depend on global parameters (e.g. mean of total gene expression) which depend on the total set of probesets. Therefore, identical probesets within different CDFs vary slightly in the final gene expression values.

Analyzing the probes of the Affymetrix microarrays discloses many inaccuracies. A large number of problematic probes are based on the fact that Affymetrix had to rely on genome annotation available at the time the chips were designed (U133A and U133B: 2001; U133 Plus 2.0 and Mouse 430 2.0: 2003). Because genome annotation improves permanently, the chip design does not properly match the present annotations anymore. Due to compatibility reasons, Affymetrix is not able to keep the design of their microarrays up to date.

The problem of cross-hybridization is well known. The first work on custom CDFs examining this error source was published by the group of Dai in 2005 [21]. They created a large amount of high quality custom CDFs related to different reference databases. Some probes, causing cross-hybridizations, are deleted from the probesets, but the filter is quite loose, so the number of problematic probes decreased but did not vanish. The use of the new CDFs can avoid full length, i.e., 25 mer long, cross-hybridizations completely. Cross-hybridization of shorter fragments are very difficult to handle due to the fact that the number of putative bindings grows exponentially the shorter the considered fragments are. Hence, if all putatively cross-hybridizing probes are excluded the amount of measurable genes will be reduced extremely.

The underlying gene annotation which is used for sequence alignment has a big impact on the number of cross-hybridizations. Manually curated mRNA sequences have a high chance of missing transcripts. Therefore, the inclusion of computational proposed gene annotations decreases the

number of false negative predicted cross-hybridizations. The drawback is that a number of false positive hybridizations increases. A more strict approach should be preferred, because it does not significantly decrease the number of covered transcripts as there is a high amount of available probes. In this study, the exclusion of XM-RefSeq-accessions results in smaller differences between the different CDFs in the number of putative cross-hybridizing transcripts. Interestingly, the correlation coefficients of the newly created probesets do not change significantly.

Evaluating the four different CDFs, we figured out that the usage of the original Affymetrix CDFs leads to poorer results than the usage of the custom CDFs, although the best Affymetrix probesets give equally good or even better results than the other CDFs. However, as already mentioned, this cannot be taken into account, because it is not known which probeset will perform best before the correlation analysis is completed. The Dai probesets perform better, but the problem of several probesets representing a single gene had not been solved. Although multiple probesets representing the same gene are expected to show similar signal intensities, this is in fact not the case [14], [15]. Thus, it is difficult to decide which of the probesets matching the same gene is the most reliable. The Ferrari and the new CDFs comprise only one probeset per gene, which is of great advantage. The Ferrari CDFs perform slightly better on the Etanercept dataset and both CDFs perform equally well on the MAQC data.

The analysis of the genes for which qRT-PCR results are available in the Etanercept dataset as well as in the MAQC dataset clearly shows higher correlation coefficients in the MAQC dataset. This is most likely due to the fact that the U133 Plus 2.0 arrays which were used in the MAQC dataset outperform the older U133A microarrays.

The results show that probesets consisting of more probes, i.e., larger probesets, lead to better correlation results in general, whereas smaller probesets perform poorer. This finding correlates to the results of the study of Cui *et al.* [14] that merges probesets matching the same transcript. Interestingly, probesets containing many putative cross-hybridizations do not considerably perform poorer than probesets containing only a few. This result is very surprising, because it is obvious that cross-hybridization is one of the main error sources in microarray experiments [8], [9]. The normalization step in the two summarizing algorithms RMA and MAS5 may explain for that because they possibly eliminate some cross-hybridization effects. Another explanation is that leaving out the problematic probes does not compensate the influence of cross-hybridization. Unspecific binding leads to two types of error: (i) false-positives because RNA fragments bind to problematic probes of the probeset, and (ii) gene expression events are missed or underestimated, leading to a false-negative error if the RNA fragments are already bound to problematic probes of other probesets (competitive binding).

Custom CDFs can only account for the first type of error by leaving out the problematic probes, the second effect could only be overcome by better array design.

The newly created CDFs perform slightly poorer than the Ferrari probesets (0.72 vs. 0.73) on the Etanercept dataset and equally well on the much larger MAQC dataset. On the one hand, the Ferrari CDFs can obviously counteract the negative effect by their much larger probesets in comparison to the new CDFs. On the other hand, using the new CDFs, putative cross-hybridizations are systematically excluded whereas using the Ferrari CDFs, the negative effect vanishes for statistical reasons due to the larger probesets. For exact studies, it is better to avoid a putative error source instead of averaging the cross-hybridization effects out as the Ferrari CDFs do. In addition, it has to be mentioned that the new CDFs provide as good or better results as the other CDFs using only about half the amount of probes (HG-U133A: 44 %, HG-U133B: 35 %, HG-U133 Plus 2.0: 37 %, Mouse Genome 430 2.0 Array: 52 %). Hence, designing new microarrays without the problematic probes, the dimension can be reduced by half without losing any information and minimize the costs of the technology tremendously. Future microarray design using only the good probes and incorporating probesets of large sizes like in the Ferrari CDFs will certainly provide optimal solutions.

Methods

Probe Analysis

For the detection of putative cross-hybridizations by sequence alignment, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using `blastn` [23]. For the U133A and the U133 Plus 2.0 the RefSeq release from 05/14/07 was used (download from `ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz`), for the U133B the release from 01/10/08, and for the Mouse 430 2.0 microarray the release from 05/09/08 (`~M_musculus/mRNA_Prot/mouse.rna.fna.gz`) was used. These parameters were applied: `ValW = 7`, `ValE = 1000`, `ValHspmax = 1`.

In this work all those RefSeq accession numbers beginning with XM or NM are used. The XM-identifiers indicate mRNA-RefSeq-accessions which are produced by computationally annotated genome submissions. The NM-identifier show that the RefSeq records are subsequently curated. Using both accessions in our model leads to more predicted cross-hybridizations which increases the reliability of the specificity of the probes.

The strand direction of the probes is analyzed. For each probe it is counted how many genes match and checked whether the match has the correct direction, i.e., the sense direction.

All BLAST hits for different transcript isoforms are merged, i.e., if the probe hybridizes to alternative splice variants of one gene but not to another gene, it is considered as unambiguous. Different gene isoforms of one gene are identified by screening the gene descriptions of the RefSeq database.

All probes matching only one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. For the creation of the new CDFs only the good probes are used. The probe sequences are annotated with GeneIDs derived from RefSeq. The GeneID is a database cross-reference qualifier, which supports access to the Entrez Gene database and provides a distinct tracking identifier for a gene or locus. Probes sharing the same GeneID are grouped together into a new probeset. The intersection between two different probesets is therefore always empty for all probesets. The size of the newly created probesets is variable and not fixed to 11 like in the Affymetrix CDFs.

Datasets

Two datasets were chosen for the validation of the different CDFs. The first of the two datasets chosen is derived from a study published by Koczan *et al.* [27] analyzing the effect of the TNF- α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points. One Affymetrix HG-U133A array was performed for each time point. The data are available at the Array Express archive [41] with the accession number E-MTAB-11.

Expression levels of 16 genes were measured by quantitative real-time RT-PCR (qRT-PCR) performed with TaqMan assay reagents according to the manufacturer's instructions on a 7900 High Throughput Sequence Detection System (Applied Biosystems, Foster City, CA, USA) using predesigned primers and probes (GAPDH Hs99999905_m1, ICAM1 Hs00164932_m1, TNFAIP3 Hs00234713_m1, IL1B Hs00174097_m1, NF κ BIA Hs00153283_m1, IL8 Hs00174103_m1, ADM Hs00181605_m1, TNF Hs00174128_m1, IL6 Hs00174131_m1, IL1RN Hs00277299_m1, SOD2 Hs00167309_m1, TRAF1 Hs00194638_m1, ZFP36 Hs00185658_m1, PTGS2 Hs00153133_m1, DUSP2 Hs00358879_m1, CROP Hs00538879_s1, JUNB HS00357891_s1).

The threshold cycle values (C_T) for specific mRNA expression in each sample were normalized to the C_T values of GAPDH mRNA in the same sample. This provides ΔC_T values that were used for the correlation analysis. In total, 816 qRT-PCR experiments were performed and complement the 51 microarray experiments (17 patients, 3 time points) described in [27]. The results of the qRT-PCR experiments can be downloaded.

The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. All available 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here. The MAQC data discussed in this publication are available in NCBI's Gene Expression Omnibus with accession number GSE5350. In addition, the nine genes for which qRT-PCR results are available in both datasets, are analyzed in more detail.

Comparison of the CDFs

For the comparison of different CDFs, the correlation between the microarray and the qRT-PCR experiments is used [29], [30]. As a performance index the Pearson correlation coefficient of the microarray results and the qRT-PCR experiments is calculated. Calculation of the Spearman correlation coefficient showed very similar results (data available at <http://sysbio.hki-jena.de/software>).

The raw chip data (CEL Files) are analyzed using the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and the Affymetrix Microarray Suite MAS5 [32] in combination with the different CDFs.

The MAS5 software assigns 'present', 'absent' or 'marginal' to each expression value, and Affymetrix recommends to use only the 'present' detection call for further analysis [32]. For an additional correlation analysis only the 'present' probesets are used to check if the calculated detection call from MAS5 gives a good prediction for the probeset quality.

Availability

The newly created CDFs as R-packages and additional files are available for download at <http://www.sysbio.hki-jena.de/software>. Using the CDFs does not interfere with all further steps of microarray analysis.

Acknowledgements

This work was supported by the ILRS - International Leibniz Research School for Microbial and Molecular Interactions (CH, FH) and by the ERASysBio+ project Linconet (MW).

References

- [1] S. Heber and B. Sick, "Quality assessment of Affymetrix GeneChip data," *OMICS A Journal of Integrative Biology*, vol. 10, no. 3, pp. 358–368, Fall 2006.
- [2] O. Modlich and M. Munnes, "Statistical framework for gene expression data analysis," *Methods in Molecular Biology*, vol. 377, pp. 111–130, May 2007.
- [3] P. K. Tan, T. J. Downey *et al.*, "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5676–5684, October 2003.
- [4] A.-K. Järvinen, S. Hautaniemi *et al.*, "Are data from different gene expression microarray platforms comparable?" *Genomics*, vol. 83, no. 6, pp. 1164–1168, June 2004.

- [5] A. Brazma, P. Hingamp *et al.*, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nature Genetics*, vol. 29, no. 4, pp. 365–371, December 2001.
- [6] S. E. Choe, M. Boutros *et al.*, "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset," *Genome Biology*, vol. 6, no. 2, p. R16, January 2005.
- [7] C. Wu, R. Carta, and L. Zhang, "Sequence dependence of cross-hybridization on short oligo microarrays," *Nucleic Acids Research*, vol. 33, no. 9, p. e84, May 2005.
- [8] Z. Chen, M. McGee *et al.*, "A distribution free summarization method for Affymetrix GeneChip® arrays," *Bioinformatics*, vol. 23, no. 3, pp. 321–327, February 2007.
- [9] A. C. Cambon, A. Khalyfa *et al.*, "Analysis of probe level patterns in Affymetrix microarray data," *BMC Bioinformatics*, vol. 8, no. 146, May 2007.
- [10] H. R. Ueda, S. Hayashi *et al.*, "Universality and flexibility in gene expression from bacteria to human," *The Proceedings of the National Academy of Sciences (US)*, vol. 101, no. 11, pp. 3765–3769, March 2004.
- [11] Affymetrix Inc, "GeneChip custom express array design guide. part no. 700506 rev. 4," Tech. Rep., 2003.
- [12] L. Zhang, M. F. Miles, and K. D. Aldape, "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnology*, vol. 21, no. 7, pp. 818–821, July 2003.
- [13] B. M. Bolstad, R. A. Irizarry *et al.*, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 3, pp. 185–193, January 2003.
- [14] X. Cui and A. E. Loraine, "Consistency analysis of redundant probe sets on Affymetrix three-prime expression arrays and applications to differential mRNA processing," *PLoS One*, vol. 4, no. 1, p. 4229, January 2009.
- [15] T. R. Hughes, M. Mao *et al.*, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer," *Nature Biotechnology*, vol. 19, no. 4, pp. 342–347, April 2001.
- [16] M. A. Stalteri and A. P. Harrison, "Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips," *BMC Bioinformatics*, vol. 8, no. 13, January 2007.
- [17] X. Liu, M. Milo *et al.*, "Probe-level measurement error improves accuracy in detecting differential gene expression," *Bioinformatics*, vol. 22, no. 17, pp. 2107–2113, September 2006.
- [18] G. Sanguinetti, M. Milo *et al.*, "Accounting for probe-level noise in principal component analysis of microarray data," *Bioinformatics*, vol. 21, no. 19, pp. 3748–3754, October 2005.
- [19] F. Ferrari, S. Bortoluzzi *et al.*, "Novel definition files for human GeneChips based on GeneAnnot," *BMC Bioinformatics*, vol. 8, no. 446, November 2007.
- [20] V. Chalifa-Caspi, I. Yanai *et al.*, "GeneAnnot: Comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes," *Bioinformatics*, vol. 20, no. 9, pp. 1457–1458, June 2004.
- [21] M. Dai, P. Wang *et al.*, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data," *Nucleic Acids Research*, vol. 33, no. 20, p. e175, November 2005.
- [22] J. Lu, J. C. Lee *et al.*, "Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays," *BMC Bioinformatics*, vol. 8, no. 108, March 2007.
- [23] S. McGinnis and T. L. Madden, "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Research*, vol. 32, pp. W20–W25, July 2004.
- [24] R. Yelin, D. Dahary *et al.*, "Widespread occurrence of antisense transcription in the human genome," *Nature Biotechnology*, vol. 21, no. 4, pp. 379–386, April 2003.
- [25] H. Kiyosawa, N. Mise *et al.*, "Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized," *Genome Research*, vol. 15, no. 4, pp. 463–474, April 2005.
- [26] M. Barnes, J. Freudenberg *et al.*, "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5914–5923, October 2005.
- [27] D. Koczan, S. Drynda *et al.*, "Molecular discrimination of responders and nonresponders to anti-TNFalpha in rheumatoid arthritis therapy by Etanercept," *Arthritis Research & Therapy*, vol. 10, p. R50, May 2008.
- [28] L. Shi, L. H. Reid *et al.*, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, September 2006.
- [29] J. S. Moray, J. C. Ryan, and F. M. Van Dolah, "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR," *Biological Procedures Online*, vol. 8, no. 1, pp. 175–193, December 2006.
- [30] R. D. Canales, Y. Luo *et al.*, "Evaluation of DNA microarray results with quantitative gene expression platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1115–1122, September 2006.
- [31] R. A. Irizarry, B. Hobbs *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, April 2003.
- [32] Affymetrix Inc, "Statistical algorithms description document. whitepaper. part no. 701137 rev. 3," Tech. Rep., 2002.
- [33] R. A. Irizarry, Z. Wu, and H. A. Jaffee, "Comparison of Affymetrix GeneChip expression measures," *Bioinformatics*, vol. 22, no. 7, pp. 789–794, July 2006.
- [34] J. Seo and E. P. Hoffman, "Probe set algorithms: is there a rational best bet?" *BMC Bioinformatics*, vol. 7, no. 395, August 2006.
- [35] S. D. Pepper, E. K. Saunders *et al.*, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, no. 273, July 2007.
- [36] M. Eisenstein, "Microarrays: Quality control," *Nature*, vol. 442, pp. 1067–1070, August 2006.
- [37] M. Grabe, *Measurement Uncertainties in Science and Technology*. New York: Springer Press, 2005.
- [38] P. Boutros, "Systematic evaluation of the microarray analysis pipeline," in *Proceedings of the First 11th MGED Meeting: 1-4 September 2008; Riva del Garda*, G. Sherlock, Ed. MGED, 2008, pp. 16–27.
- [39] H.-C. Liu, C.-Y. Chen *et al.*, "Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 570–579, August 2008.
- [40] K. Shedden, W. Chen *et al.*, "Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data," *BMC Bioinformatics*, vol. 6, no. 26, 2005.
- [41] H. Parkinson, M. Kapushesky *et al.*, "ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D868–D872, January 2009.

A validation method for fuzzy clustering of gene expression data

Thanh Le¹, Katheleen J. Gardiner²

¹Department of CSE, University of Colorado Denver, Denver, CO, USA

²Department of Pediatrics, University of Colorado Denver, Denver, CO, USA

Abstract - Clustering is a key process in data mining for revealing structure and patterns in data. Fuzzy C-means (FCM) is a popular algorithm using a partitioning approach for clustering. One advantage of FCM is that it converges rapidly. In addition, using fuzzy sets to represent the degrees of cluster membership of each data point provides more information regarding relationships within the data than do alternative approaches that use crisp clustering. However, a limitation of FCM is that it requires initial specification of the number of clusters and subsequent validation of this number. Here, we propose a Bayesian method for fuzzy clustering validation using the fuzzy partition. We show that this method outperforms popular fuzzy cluster indices on both artificial and real biological datasets.

Availability: The supplementary documents and the method software are at <http://ouray.ucdenver.edu/~tnle/fzble>.

Keywords: fuzzy c-means; Bayesian; cluster index

1 Introduction

Cluster analysis groups data points based on their similar properties and can help to discover patterns and correlations in large datasets. Successful clustering maximizes both the compactness of data points within a cluster and the discrimination between clusters. Fuzzy C-Means (FCM, Bezdek 1981) is a popular algorithm that uses a partitioning approach with fuzzy cluster boundaries and fuzzy sets that associate each data point with one or more clusters. An advantage of FCM is that it converges rapidly, however, like most partitioning clustering algorithms, it depends strongly on the initial parameters and requires estimation of

the number of clusters. While for some initial values, FCM may converge to a global optimum, for others, it may get stuck in a local optimum. In addition, during the clustering process, the optimization of the compactness and separation of a fuzzy partition may be inconsistent with the optimal number of clusters in the dataset. For these reasons, final clustering results require validation to assess how good the fuzzy partition is, if better fuzzy partitions exist, and, when not known a priori, the optimal number of clusters in the dataset.

Several cluster validity index functions have been proposed. Bezdek [1] measured performance using partition entropy and the overlap of adjacent clusters. Fukuyama and Sugeno [2] combined the FCM objective function with the separation factor, while Xie and Beni [3], integrated the Bezdek index [1] with the cluster separation factor. Rezaee et al. [4] combined the compactness and separation factors, and Pakhira et al. [5] combined the same two factors where the separation factor was normalized. Recently, Rezaee [6] proposed a new cluster index in which the two factors are normalized across the range of possible numbers of clusters.

Here, we propose a fuzzy clustering cluster index that uses the fuzzy partition and the distance matrix between cluster centers and data points. Instead of compactness and separation, our cluster index is based on a Bayesian model and a log-likelihood estimator. With the use of both the possibility model and the probability model to represent the data distribution, our method is appropriate for artificial data where the distribution follows a standard model, as well as for real datasets, in particular, gene expression data, that lack a standard distribution. We show that our method outperforms popular cluster indices on both artificial and biological datasets.

2 Fuzzy C-Means and popular cluster indices

2.1 Fuzzy C-Means algorithm

Fuzzy C-Means (FCM) is an unsupervised clustering algorithm that has been applied successfully to numerous problems involving feature analysis. Its applications include biological data analysis, in particular, gene expression data.

This work was supported by the National Institutes of Health (HD056235), the Linda Crnic Institute (KG) and the Vietnamese Ministry of Education and Training (TL).

Thanh Le is a doctoral student in the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217-3364, USA (email: tnlmail@yahoo.com).

Katheleen J. Gardiner is a professor in the Department of Pediatrics; the Intellectual and Developmental Disabilities Research Center; and the Computational Biosciences, Human Medical Genetics and Neuroscience Programs, University of Colorado Denver, Aurora, CO 80045, USA (phone: 303-724-0572; email: katheleen.gardiner@ucdenver.edu).

Given a dataset $X = \{x_i \in \mathbb{R}^p, i=1..n\}$, where $n>0$ is the number of data points and $p>0$ is the dimension of the data space of X , let $c, c \in \mathbb{N}, 2 \leq c \leq n$, be the number of clusters in X . Denote $V = \{v_k \in \mathbb{R}^p, k=1..c\}$ as the set of center points of c clusters in the fuzzy partition; $U = \{u_{ki} \in [0,1], i=1..n, k=1..c\}$ as the partition matrix, where u_{ki} is the membership degree of the data point x_i to the k^{th} cluster, and

$$\sum_{k=1}^c u_{ki} = 1, i = 1..n. \quad (1)$$

The clustering problem is to determine the values of c and V such that:

$$J(X | U, V) = \sum_{i=1}^n \sum_{k=1}^c u_{ki} \|x_i - v_k\| \rightarrow \min, \quad (2)$$

where $\|x-y\|$ is the distance between the data points x and y in \mathbb{R}^p , defined using Euclidean distance as:

$$\|x - y\|^2 = \sum_{i=1}^p (x^i - y^i)^2. \quad (3)$$

By using fuzzy sets to assign data points to clusters, FCM allows adjacent clusters to overlap. It thus provides more information on the relationships of data points. In addition, by using a fuzzifier factor, $m, 1 \leq m < \infty$, in its objective function (4), the clustering model from FCM is more flexible in changing the overlap regions among clusters.

$$J(X | U, V) = \sum_{i=1}^n \sum_{k=1}^c u_{ki}^m \|x_i - v_k\| \rightarrow \min, \quad (4)$$

The following is a solution of (4) with respect to (1):

$$v_k = \frac{\sum_{i=1}^n u_{ki}^m x_i}{\sum_{i=1}^n u_{ki}^m}, \quad (5)$$

$$u_{ki} = \left(\frac{1}{\|x_i - v_k\|^2} \right)^{\frac{1}{1-m}} / \sum_{j=1}^c \left(\frac{1}{\|x_i - v_j\|^2} \right)^{\frac{1}{1-m}}. \quad (6)$$

FCM uses an iteration process to estimate the solution of (5) and (6). This process is iterated until convergent where

$$\exists \varepsilon_u > 0, T > 0: \forall t > T,$$

$$\|U_{t+1} - U_t\| = \max_{k,i} \left\{ \|u_{ki}(t+1) - u_{ki}(t)\| \right\} < \varepsilon_u. \quad (7)$$

$$\text{Or, } \exists \varepsilon_v > 0, T > 0: \forall t > T,$$

$$\|V_{t+1} - V_t\| = \max_k \left\{ \|v_k(t+1) - v_k(t)\| \right\} < \varepsilon_v. \quad (8)$$

While FCM can converge quickly, it is unable to determine the optimal number of clusters in the dataset.

2.2 Cluster validation indices

- (i) To determine if the fuzzy partition is valid, traditional cluster indices use two criteria, compactness, which measures the closeness of cluster elements typically using the variance. Because variance indicates how different the members are, a low value of variance is an indicator of closeness, and (ii)
- (ii) Separation, which computes the “distance” between two different clusters, e.g., the distance between representative objects of two clusters. This measure has been widely used due to its computational efficiency and its effectiveness for hyper sphere-shaped clusters.

2.2.1 PC index

The partition coefficient (PC) index was proposed by Bezdek [1] as in (9). It indicates the average relative amount of shared membership between pairs of fuzzy subsets in U , by combining into a single number, the average content of pairs of fuzzy algebraic products. The index values range from $[1/c, 1]$.

$$V_{PC} = \frac{1}{n} \sum_{k=1}^c \sum_{i=1}^n u_{ki}^2. \quad (9)$$

An optimal cluster number c can be found by solving,

$$V_{PC}(c_{opt}) = \max_{2 \leq c \leq n} \{V_{PC}(c)\}$$

2.2.2 PE index

The partition entropy (PE) index was proposed by Bezdek [1] as

$$V_{PE} = \frac{1}{n} \sum_{k=1}^c \sum_{i=1}^n u_{ki} \times \log_a(u_{ki}), \quad (10)$$

where a is the base of the logarithm. According to [1], the limitation of the PE can be attributed to its apparent monotonicity and to an extent, to the heuristic nature of the rationale underlying its formulation. An optimal cluster number c can be found by solving $V_{PE} \rightarrow \min$.

2.2.3 FS index

The Fukuyama-Sugeno cluster index (FS) was proposed by Fukuyama and Sugeno [2] as

$$V_{FS} = J - \sum_{k=1}^c \sum_{i=1}^n u_{ki}^m \|v_k - \bar{v}\|^2, \quad (11)$$

where, $\bar{v} = \sum_{k=1}^c v_k / c$. An optimal number of clusters can be found by solving $V_{FS} \rightarrow \min$.

2.2.4 XB index

The XB index was proposed by Xie and Beni as in (12). The numerator indicates the compactness of the fuzzy partition, while the denominator indicates the strength of the separation between clusters. A good partition produces a small value for the compactness, and well-separated $\{v_i\}$ will produce a high value for the separation. An optimal c therefore is found by solving $V_{XB} \rightarrow \min$.

$$V_{XB} = \frac{\sum_{k=1}^c \sum_{i=1}^n u_{ki}^m \times \|x_i - v_k\|^2}{n \times \min_{k,l} \|v_k - v_l\|^2} \quad (12)$$

2.2.5 CWB index

The Compose Within and Between scattering (CWB) index was proposed by Rezaee et al. [4].

$$V_{CWB} = \alpha \text{Scat}(c) + \text{Dis}(c), \quad (13)$$

where α is a weighting factor equal to $\text{Dis}(c_{\max})$. The average scatter is defined as

$$\text{Scat}(c) = \frac{\sum_{k=1}^c \|\sigma(v_k)\|}{c \times \|\sigma(X)\|}, \quad (14)$$

where $\|x\| = (x^T x)^{1/2}$, $\sigma(X) = \frac{1}{n} \sum_{i=1}^n \|x_i - \bar{x}\|^2$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

The Dis function is defined as

$$\text{Dis}(c) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^c \left(\sum_{i=1}^n \|v_k - x_i\| \right)^{-1}, \quad (15)$$

where $D_{\min} = \min_{k,l} \|v_k - v_l\|$ and $D_{\max} = \max_{k,l} \|v_k - v_l\|$. The Scat() function indicates the average of the scattering variation within the clusters. A small value for this term indicates a compact partition. The Dis() function indicates the total scattering separation between the clusters, it is influenced by the geometry of the cluster centroids and increases with the number of clusters. An optimal number of clusters c is found by solving $V_{CWB} \rightarrow \min$.

2.2.6 PBMF index

The PBMF index is a fuzzy version of the PBM index proposed by Pakhira, Bandyopadhyay and Maulik [5] as

$$V_{PBMF} = \left(\frac{1}{c} \frac{E_1}{J} D_c \right)^2, \quad (16)$$

$$E_1 = \sum_{i=1}^n u_{ii} \|x_i - \bar{x}\|, \quad (17)$$

where $D_c = \max_{k,l} \|v_k - v_l\|$. The value of V_{PBMF} decreases as the number of clusters c increases. An optimal number of clusters can be found by solving $V_{PBMF} \rightarrow \max$.

2.2.7 BR index

The cluster index of Rezaee B. (BR) [6] uses both the compactness and separation criteria normalized across clustering partitions using possible numbers of clusters in a given range. The index is defined as

$$V_{BR} = \frac{\text{Sep}(c)}{\max_c \{\text{Sep}(c)\}} + \frac{J(c)}{\max_c \{J(c)\}}, \quad (18)$$

where $\text{Sep}(c) = \frac{2}{c(c-1)} \sum_{k=1}^c S_{\text{rel}}(v_k, v_l)$.

The similarity $S_{\text{rel}}(\cdot)$ of two fuzzy sets is defined as

$$S_{\text{rel}}(v_k, v_l) = \sum_{i=1}^n S(x_i : v_k, v_l) \times h(x_i), \quad (19)$$

where $S(x_i : v_k, v_l) = \min(u_{ki}, u_{li})$,

$$h(x_i) = - \sum_{k=1}^c u_{ki} \times \log_a(u_{ki}).$$

Because V_{BR} is a sum of compactness and separation factors, the smaller it is, the better the fuzzy partition is. An optimal number of clusters c therefore can be found by solving $V_{BR} \rightarrow \min$.

3 The proposed validation method

3.1 The proposed validation method (fzBLE)

Instead of compactness and separation factors, we propose a validation method (fzBLE) that is based on a log likelihood estimator with a fuzzy based Bayesian model. Each fuzzy clustering solution is modeled with $\theta = \{U, V\}$, where V represents the cluster centers and, U is the partition matrix representing the membership degrees of the data points to the clusters. The likelihood of the clustering model and the data is measured as

$$L(\theta | X) = L(U, V | X) = \prod_{i=1}^n P(x_i | U, V) = \prod_{i=1}^n \sum_{k=1}^c P(v_k) \times P(x_i | v_k). \quad (20)$$

The log likelihood estimator is then computed as

$$\log(L) = \sum_{i=1}^n \log \left(\sum_{k=1}^c P(v_k) \times P(x_i | v_k) \right) \rightarrow \max. \quad (21)$$

An optimal number of clusters is obtained by solving (21).

3.2 Possibility to probability transformation

Because our clustering model is possibility-based, before applying equations (20) and (21), a transformation of possibility to probability is needed. Given a fuzzy clustering model $\theta = \{U, V\}$, according to [7], u_{ki} is the possibility that $v_k = x_i$. If θ is a proper fuzzy partition, then there exists some x^* such that $U_k(x^*) = 1$, $k=1..c$, and U_k is a normal possibility distribution. Assume P_k is the probability distribution of v_k on X where $p_{k1} \geq p_{k2} \geq p_{k3} \geq \dots \geq p_{kn}$. We associate with P_k a possibility distribution U_k on X [7] such that u_{ki} is the possibility of x_i where

$$\begin{aligned} u_{kn} &= n \times p_{kn} \\ u_{ki} &= i(p_{ki} - p_{k,i+1}) + u_{k,i+1}, \quad i = n-1, \dots, 1. \end{aligned} \quad (22)$$

Reversing (22), we obtain the transformation of a possibility distribution to a probability distribution. Assume that U_k is ordered the same way with P_k on X : $u_{k1} \geq u_{k2} \geq u_{k3} \geq \dots \geq u_{kn}$.

$$\begin{aligned} p_{kn} &= u_{kn} / n \\ p_{ki} &= p_{k,i+1} + (u_{ki} - u_{k,i+1}) / i. \end{aligned} \quad (23)$$

P_k is an approximate probability distribution of v_k on X , and $p_{ki} = P(x_i | v_k)$. If U_k is a normal possibility distribution then $\sum p_{ki} = 1$.

3.3 Data distributions

Using the value of P_k , we can estimate the variance σ_k , the prior probability $P(v_k)$ and the normal distribution of v_k .

$$\sigma_k = \sum_{i=1}^n p_{ki} \|x_i - v_k\|^2, \quad (24)$$

$$P(v_k) = \frac{\sum_{i=1}^n P(x_i | v_k)}{\sum_{i=1}^c \sum_{i=1}^n P(x_i | v_i)}, \quad (25)$$

$$P_n(x_i | v_k) = \left((\sum \pi)^{1/n} \times \sigma_k \times e^{-\frac{\|x_i - v_k\|^2}{2\sigma_k^2}} \right)^{-1}. \quad (26)$$

In real datasets, for a cluster v_k , the data points usually come from different random distributions. Because they cluster in v_k , they tend to follow the normal distribution estimated as in (26). This idea is based on the Central Limit Theorem. We therefore integrate the probabilities computed in (23) and (26) for the probability of the data point x_i given cluster v_k as

$$P^*(x_i | v_k) = \max\{P(x_i | v_k), P_n(x_i | v_k)\}. \quad (27)$$

Equation (27) better represents the data distribution, particularly in real datasets. The fzBLE method is based on (21) with (25) and (27).

3.4 fzBLE and FCM combination

fzBLE can be used with the standard FCM algorithm to search for the optimal number of clusters for a dataset using a cluster range.

- Input:
 - The data to cluster $X = \{x_i\}$, $i=1..n$
 - Cluster range $[c_{min}, c_{max}]$
- Output: An optimal fuzzy partition solution,
 - c_{opt} : Optimal number of clusters
 - $V = \{v_i\}$, $i=1..c$: Cluster centers
 - $U = \{u_{ki}\}$, $i=1..n, k=1..c$: Partition matrix

Steps

1. Set $c_{opt} = c_{min}$
2. For each value of c in $[c_{min}, c_{max}]$
 - Generate a fuzzy partition using FCM
 - Validate the partition using fzBLE
 - If the current partition is better than the optimal one then, set $c_{opt} = c$
3. Return $\{c_{opt}, U, V\}$ an optimal solution.

4 Experimental results

To evaluate fzBLE, we generated 84 artificial datasets using the method in [8]. Datasets are distinguished by the dimensions and cluster number, and we generated $(3-2+1) \times (9-3+1) = 14$ dataset types. For each type, we generated 6 datasets, for a total of $6 \times 14 = 84$. For real datasets, we used the Iris, Wine and Glass datasets from the UC Irvine Machine Learning Repository [9], and the gene expression datasets, Yeast [13], Yeast-MIPS [14, 15] and RCNS [10]. These datasets contain classification information, useful for comparing cluster indices. We compared performance of fzBLE with the cluster indices from PC, PE, FS, XB, CWR, PBMF and BR [1-6]. The compactness factor (CF) of the FCM algorithm is also recorded in the results.

4.1 Artificial datasets

For each artificial dataset, we ran the standard FCM algorithm five times with m set to 2.0 and the partition matrix initialized randomly. In each case, the best fuzzy partition was then selected to run fzBLE and the other cluster indices to search for the optimal number of clusters between 2-12 and to compare this with the known number of clusters. We repeated the experiment 20 times and averaged the performance of each method. Table 1 shows the fraction of correct predictions. fzBLE and PBMF outperform other approaches, while CF is the least effective.

Table 1

Fraction of correct cluster predictions on artificial datasets

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
3	1.00	0.42	0.42	0.42	0.42	1.00	1.00	0.83	0.00
4	1.00	0.92	0.92	0.92	0.83	1.00	1.00	1.00	0.00
5	1.00	0.75	0.75	0.83	0.75	0.83	1.00	1.00	0.00
6	1.00	0.92	0.83	0.92	0.58	0.58	1.00	0.92	0.00
7	1.00	0.83	0.83	0.83	0.67	0.58	1.00	0.67	0.00
8	1.00	1.00	0.92	1.00	0.92	0.67	1.00	0.83	0.00
9	1.00	0.92	0.67	0.92	0.67	0.33	1.00	0.83	0.00

Table 2

Validation method performance on the Iris dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-763.0965	0.9554	0.0977	-10.6467	0.0203	177.1838	12.3280	1.1910	0.9420
3	-762.8034	0.8522	0.2732	-9.3369	0.1292	213.4392	17.7131	1.0382	0.3632
4	-764.8687	0.7616	0.4381	-7.4821	0.2508	613.2656	14.4981	1.1344	0.2665
5	-770.2670	0.6930	0.5703	-8.2331	0.3473	783.4697	13.6101	1.0465	0.1977
6	-773.6223	0.6549	0.6702	-7.3202	0.2805	904.3365	12.3695	1.0612	0.1542
7	-774.4740	0.6155	0.7530	-6.8508	0.2245	1029.7342	11.2850	0.9246	0.1262
8	-774.8463	0.6000	0.8111	-6.9273	0.3546	1635.3593	10.5320	0.8692	0.1072
9	-780.1901	0.5865	0.8556	-6.6474	0.3147	1831.5705	9.9357	0.7653	0.0905
10	-781.7951	0.5765	0.8991	-6.0251	0.2829	2080.3339	9.3580	0.7076	0.0787

Table 3

Validation method performance on the Wine dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-926.4540	0.9264	0.1235	-113.0951	0.1786	3.9100	1.3996	2.0000	61.1350
3	-924.0916	0.8977	0.1764	-104.9060	0.2154	3.2981	0.9316	1.4199	39.3986
4	-932.8377	0.8607	0.2525	-139.9144	0.5295	6.6108	0.6306	1.1983	33.7059
5	-929.6146	0.8225	0.3281	-126.5746	0.5028	6.9001	0.4700	1.0401	28.4741
6	-928.8121	0.8066	0.3669	-118.4715	0.6173	9.2558	0.3706	0.9111	25.3451
7	-930.6451	0.7988	0.3874	-120.3128	0.6465	10.3803	0.2972	0.7629	23.1742
8	-932.0462	0.7993	0.3917	-124.7999	0.6459	11.0836	0.2471	0.6392	21.4411
9	-932.1902	0.7929	0.4120	-122.8396	0.6367	11.8373	0.2100	0.5801	19.9154
10	-935.0478	0.7909	0.4217	-130.9089	0.6270	11.9941	0.1773	0.5252	18.9891

Table 4

Validation method performance on the Glass dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-1135.6886	0.8884	0.1776	0.3700	0.7222	6538.9311	0.3732	1.9817	0.5782
3	-1127.6854	0.8386	0.2747	0.1081	0.7817	4410.3006	0.4821	1.5004	0.4150
4	-1119.2457	0.8625	0.2515	-0.0630	0.6917	3266.5876	0.4463	1.0455	0.3354
5	-1123.2826	0.8577	0.2698	-0.1978	0.6450	2878.8912	0.4610	0.8380	0.2818
6	-1113.8339	0.8004	0.3865	-0.2050	1.4944	5001.1752	0.3400	0.8371	0.2430
7	-1116.5724	0.8183	0.3650	-0.2834	1.3802	5109.6082	0.3891	0.6914	0.2214
8	-1127.2626	0.8190	0.3637	-0.3948	1.4904	7172.2250	0.6065	0.5916	0.2108
9	-1117.7484	0.8119	0.3925	-0.3583	1.7503	8148.7667	0.3225	0.5634	0.1887
10	-1122.1585	0.8161	0.3852	-0.4214	1.7821	9439.3785	0.3909	0.4926	0.1758
11	-1121.9848	0.8259	0.3689	-0.4305	1.6260	9826.4211	0.3265	0.4470	0.1704
12	-1135.0453	0.8325	0.3555	-0.5183	1.4213	11318.4879	0.5317	0.3949	0.1591
13	-1138.9462	0.8317	0.3556	-0.5816	1.4918	14316.7592	0.6243	0.3544	0.1472

Table 5

Validation method performance on the Yeast dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-2289.8269	0.9275	0.1172	-85.1435	0.2060	8.3660	1.2138	2.0000	133.0734
3	-2296.4502	0.9419	0.0983	-157.2825	0.2099	4.7637	0.6894	1.0470	94.6589
4	-2305.3369	0.9437	0.1000	-191.7664	0.2175	4.0639	0.5575	0.7240	74.7629
5	-2289.3070	0.9087	0.1648	-187.1073	1.0473	13.6838	0.4087	0.6722	65.9119
6	-2296.3098	0.8945	0.1939	-196.6711	0.9932	13.8624	0.3050	0.6170	60.8480
7	-2296.6017	0.8759	0.2299	-198.2858	1.0558	15.4911	0.2434	0.5686	56.1525
8	-2299.4225	0.8634	0.2526	-201.7688	1.0994	16.9644	0.2050	0.5132	51.2865
9	-2299.3653	0.8453	0.2871	-205.1489	1.2340	20.2532	0.1741	0.4819	48.0737
10	-2302.7581	0.8413	0.2992	-208.5687	1.1947	20.7818	0.1512	0.4533	45.9442
11	-2300.3294	0.8325	0.3186	-209.4023	1.1731	21.1525	0.1307	0.4272	43.6600
12	-2307.5701	0.8290	0.3272	-213.4658	1.2245	23.0389	0.1157	0.4040	42.1594
13	-2310.7819	0.8270	0.3354	-215.2463	1.3036	25.4062	0.1016	0.3847	40.8654

Table 6

Validation method performance on the YEAST-MIPS dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-1316.4936	0.9000	0.1625	25.4302	0.3527	16.7630	0.7155	1.9978	81.0848
3	-1317.3751	0.9092	0.1615	-32.8476	0.2981	10.1546	0.8032	1.2476	58.2557
4	-1304.0374	0.8216	0.3252	-39.4858	2.5297	39.8434	0.5400	1.3218	48.6275
5	-1308.6776	0.8279	0.3216	-54.4979	2.4245	34.9963	0.3620	0.9558	41.9671
6	-1309.9191	0.8211	0.3460	-59.8918	2.3511	35.4533	0.2691	0.8291	38.5468
7	-1315.3692	0.8139	0.3654	-65.4866	2.3562	38.8797	0.2423	0.7252	36.0906
8	-1315.1479	0.8062	0.3918	-67.6774	2.4958	43.9502	0.1966	0.6712	34.1387
9	-1321.2280	0.8109	0.3874	-72.3197	2.2854	41.2112	0.1664	0.6072	32.3289
10	-1324.1578	0.8158	0.3847	-74.7867	2.0433	37.6154	0.1395	0.5588	30.9686

Table 7

Validation method performance on the RCNS dataset

#c	fzble	PC	PE	FS	XB	CWB	PBMF	BR	CF
2	-580.0728	0.9942	0.0121	-568.7972	0.0594	5.5107	4.2087	1.1107	177.8094
3	-564.1986	0.9430	0.0942	-487.6104	0.4877	4.1309	4.2839	1.6634	117.9632
4	-561.0169	0.9142	0.1470	-430.4863	0.9245	6.1224	3.3723	1.3184	99.1409
5	-561.7420	0.8900	0.1941	-397.0935	1.3006	9.4770	2.6071	1.1669	88.5963
6	-552.9153	0.8695	0.2387	-300.6564	2.5231	20.6496	1.9499	1.1026	84.0905
7	-556.2905	0.8707	0.2386	-468.3121	2.1422	21.0187	2.8692	0.7875	57.5159
8	-555.3507	0.8925	0.2078	-462.0673	1.7245	20.0113	2.5323	0.5894	52.0348
9	-558.8686	0.8863	0.2192	-512.4278	1.6208	22.4772	2.6041	0.5019	45.9214
10	-565.8360	0.8847	0.2241	-644.1451	1.1897	21.9932	3.4949	0.3918	33.1378

4.2 Real datasets

The Iris, Wine and Glass datasets contain 3, 3 and 6 clusters, respectively. For the Iris dataset, only fzBLE and PBMF detected the correct number of clusters (Table 2). For the Wine and Glass datasets, only fzBLE and CWB, and only fzBLE, respectively, detected the correct number of clusters (Tables 3 and 4).

4.3 Biological datasets

4.3.1 Yeast

The Yeast dataset [13] reports expression levels of yeast genes throughout two cell cycles at 17 time points spaced at 10-minute intervals. Each of the 384 differentially expressed genes was labeled with one of the five cell cycle phases where their expression changed. We ran the FCM algorithm with m set to 1.17 [12] and used the clustering partition to test all methods as in previous sections. Table 5 shows that only fzBLE detected the correct number of clusters (5) in Yeast dataset.

4.3.2 Yeast-MIPS

The Yeast-MIPS dataset [14] is a subset of the Yeast dataset [14]. It contains 237 genes belonging to four functional categories: DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins [15]. We ran the FCM algorithm using the same parameters as with Yeast dataset. The results in Table 6 show that only fzBLE detected the four clusters in the Yeast-MIPS dataset.

4.3.3 RCNS

The RCNS (Rat Central Nervous System) dataset contains expression levels of 112 genes measured at nine time points during rat central nervous system development [10]. Wen et al. [11] preprocessed the dataset using a normalization method and scaling across adjacent axes to generate a 112x17 dataset so that Euclidean distance can be applied. The FITCH software was used to detect 6 clusters with biological relevance. Dembélé and Kastner [12] used the FCM algorithm varying the number of clusters and reported that 6 is the optimal number. We ran fzBLE and the other cluster indices on the dataset clustering partition found by the standard FCM algorithm using the Euclidean metric for distance measurement. Table 7 shows that again only fzBLE detected the correct number of clusters.

5 Conclusions

We have presented a novel method, fzBLE to evaluate results of fuzzy partitioning by the standard FCM algorithm. fzBLE is novel in that it uses the log likelihood estimator with a Bayesian model and the possibility, rather than the probability, distribution model of the dataset from the fuzzy partition. By using the Central Limit Theorem, fzBLE effectively represents distributions in real datasets. Results have shown that fzBLE performs effectively on both artificial and real datasets. In future work, we will integrate this method with optimization algorithms, to develop new clustering algorithms that can effectively support clustering analysis on real datasets.

6 References

- [1] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981.
- [2] Y. Fukuyama, M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method", in: Proc. Fifth Fuzzy Systems Symp., 1989, pp. 247–250.
- [3] X.L. Xie, G. Beni, "A validity measure for fuzzy clustering", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 13, pp. 841–847, 1991.
- [4] M.R. Rezaee, B.P.F. Lelieveldt, J.H.C. Reiber, "A new cluster validity index for the fuzzy c-mean", Pattern Recognition Lett, Vol. 19, pp. 237–246, 1998.
- [5] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, "Validity index for crisp and fuzzy clusters", Pattern Recognition, Vol. 37, pp. 481–501, 2004.
- [6] B. Rezaee, "A cluster validity index for fuzzy clustering", Fuzzy Sets and Systems, Vol. 161, pp. 3014–3025, 2010.
- [7] M.C. Florea, A.L. Jusselme, D. Grenier, E. Bosse, "Approximation techniques for the transformation of fuzzy sets into random sets", Fuzzy Sets and Systems, Vol. 159, pp. 270–288, 2008.
- [8] L. Xu, M.I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures", Neural Computation, Vol. 8, pp. 129–151, 1996.
- [9] A. Frank and A. Asuncion, (2010) "Machine Learning Repository", [Online], <http://archive.ics.uci.edu/ml>.
- [10] R. Somogyi, X. Wen, W. Ma, J.L. Barker, "Developmental kinetic of GLAD family mRNAs parallel neurogenesis in the rat", Journal of Neurosciences, Vol. 15, pp. 2575–2591, 1995.
- [11] X. Wen, S. Fuhrman, G.S. Michaels, G.S. Carr, D.B. Smith, J.L. Barker, R. Somogyi, "Large scale temporal gene expression mapping of central nervous system development". Proc of the National Academy of Science USA, Vol. 95, 1998, pp. 334–339.
- [12] D. Dembele, P. Kastner, "Fuzzy C-Means method for clustering microarray data", Bioinformatics, Vol. 19, pp. 973–980, 2003.
- [13] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, R.W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle", Mole Cell, Vol. 2, pp. 65–73, 1998.
- [14] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, W.L. Ruzzo, "Model based clustering and data transformations for gene expression data", Bioinformatics, Vol. 17, pp. 977–987, 2001.
- [15] H. W. Mewes, J. Hani, F. Pfeiffer, D. Frishman, "MIPS: A database for protein sequences and complete genomes", Nucleic Acids Research, Vol. 26, pp. 33–37, 1998.

Robust SVD Method for Missing Value Estimation of DNA Microarrays

Fen Qin, Joseph Collins, and Jeonghwa Lee

Department of Computer Science, Shippensburg University, Shippensburg, PA, U.S.A.

Abstract—A majority of DNA microarray datasets contain missing or corrupt values and it is critical to estimate these values accurately. These missing values are most often attributed to insufficient experimental resolution or the presence of foreign objects on the experimental slide's surface. To improve existing missing value estimation algorithms, this paper introduces and investigates the scalable singular value decomposition (SSVD) solver, which is an improvement upon the Jacobi singular value decomposition (SVD) approach. Experiments were conducted on a study comparing SSVD to the Jacobi and QR SVD methods against several legitimate microarray datasets. The robustness of SSVD is verified by subjecting it to several test cases containing 1–20% of missing values. For nearly all of the test cases across all configurations of missing value percentages, SSVD provides more accurate recovery results than Jacobi and SQ SVD. These numerical results strongly suggest SSVD is a robust and scalable solver.

Keywords: Microarrays, missing value estimation, singular value decomposition

1. Introduction

Deoxyribonucleic acid (DNA) microarray analysis is the study of large scale gene expression experiments, which grants researchers insight into solving many pertinent biological questions [12] including cancer classification, identifying the effects of specific gene therapies and exploring the unknown gene function [11]. The microarray data generated by gene expression experiments is presented as one large matrix consisting of genes ordered by rows and experimental conditions by columns [3]. Even though DNA microarray analysis is an emerging and powerful tool for researchers to utilize, the data produced by microarray experiments is typically not complete. Missing or corrupt data is most commonly attributed to insufficient resolution, image corruption or foreign objects such as dust or scratches on the surface of the experimental slide [8]. Incomplete microarray data is undesirable because complete datasets are a prerequisite for existing gene expression data analysis algorithms. If a microarray dataset contains missing values, then researchers are unable to properly draw conclusions about the gene expression experiments.

There are several gene expression data analysis algorithms available for missing value recovery, including the

singular value decomposition imputation (SVDimpute) [7], weighted k -nearest neighbors imputation (KNNimpute) [10], least squares imputation (LSimpute) [2], local least squares imputation (LLSimpute) [9] and dynamic local least squares imputation (DLLSimpute) [6]. Among these methods, LLSimpute has been suggested to be an efficient recovery method for microarray datasets. The solving force behind LLSimpute is the orthogonal-triangular decomposition algorithm powered by QR factorization, denoted the QR singular value decomposition (SVD) method. When the SVD routine of the aforementioned algorithms is changed, the overall accuracy of missing value estimation may improve.

Based on its ease of implementation, QR SVD is typically the solution of choice for software written in MATLAB. Within the MATLAB package, there exists a number of approaches that these gene expression data analysis algorithms commonly utilize to calculate the inverse of a matrix. The approaches for which we are concerned, referred to as the standard MATLAB solvers, are $inv(A)$ and $pinv(A)$, which form the explicit inverse and the Moore-Penrose pseudoinverse of a square matrix, respectively. Algorithms from articles [4], [5], denoted the Jacobi SVD method, propose an improved SVD algorithm driven by an advanced matrix inverse approach.

This paper introduces a new SVD algorithm, referred to as the scalable singular value decomposition (SSVD) solver, which is a further improvement upon the Jacobi SVD implementation, designed to improve the accuracy of missing value estimation methods. In order to compare these solvers in a fair and unbiased manner, each solver was tested with four complete microarray datasets. Completing these microarray datasets was achieved by recovering the missing values in each dataset using LLSimpute with the SSVD solver. Test cases were then created by randomly inducing missing values of various percentages into these artificially completed datasets. The experimental results for each test case were achieved by swapping out the SVD solver within LLSimpute.

This paper is organized as follows: Section 2 gives a concise introduction to the implementation of LLSimpute, the implementation of an SVD solver and the improvements introduced by SSVD. In Section 3, numerical experiments of SSVD versus Jacobi and QR SVD are presented to highlight the improved accuracy and scalability of SSVD. Concluding remarks are made in Section 4.

2. Local Least Squares Imputation

2.1 Selecting the Target Gene

$G \in R^{m \times n}$ denotes a gene expression data matrix with m genes and associated n experiments. Assume $m \gg n$. In the matrix G , a row $g_i^T \in R^{1 \times n}$ represents the i -th gene of n experiments as

$$G = \begin{pmatrix} g_1^T \\ \vdots \\ g_m^T \end{pmatrix} \quad (1)$$

and each missing value location at the i -th gene and j -th experiment will be represented as

$$G(i, j) = g_i(j) = \begin{pmatrix} g_{1,1} & \cdots & g_{1,j} & \cdots & g_{1,n} \\ \vdots & & \vdots & & \vdots \\ g_{i,1} & \cdots & g_{i,j} & \cdots & g_{i,n} \\ \vdots & & \vdots & & \vdots \\ g_{m,1} & \cdots & g_{m,j} & \cdots & g_{m,n} \end{pmatrix},$$

where $i \in (1, 2, \dots, m)$ and $j \in (1, 2, \dots, n)$. For consistency, we assume that all missing value estimation algorithms discussed throughout this paper consider the first position of the first gene to be a missing value, i.e.

$$G(1, 1) = g_1(1) = \beta,$$

where this first gene is selected as the target gene.

2.2 Missing Value Recovery Using SVD

The k -nearest neighbor genes of the target gene are selected where $1 < k < m$ and $k > n$. The matrix $A \in R^{k \times (n-1)}$ and vector $b \in R^{k \times 1}$ are formed from these k -nearest neighbor genes. The vector $w \in R^{1 \times (n-1)}$ is formed from the target gene. Local least squares methods solve the following equation:

$$\min_x \|A^T x - w\|_2, \quad (2)$$

where solving Eq. (2) is equivalent to solving

$$\min_x \|A^T x - w\|_2^2. \quad (3)$$

By the definition of the inner product, we have

$$\min_x \|A^T x - w\|_2^2 = \min_x (A^T x - w)^T (A^T x - w). \quad (4)$$

Eq. (4) is equivalent to

$$\frac{\partial}{\partial x_j} \sum_{i=1}^{n-1} (A^T x - w)_i^2 = 2 \sum_{j=1}^{n-1} (A^T x - w)_j^T A_j^T = 0, \quad (5)$$

$$j = 1, 2, \dots, k,$$

where $(A^T x - w)_i$ is the i -th component of the column vector. Eq. (5) comes down to the critical point of $\|A^T x - w\|_2^2$, where

$$\sum_{j=1}^{n-1} A_j (A^T x - w)_j = 0, \quad j = 1, 2, \dots, k, \quad (6)$$

and a vector form

$$\begin{pmatrix} A_1 (A^T x - w)_1 \\ A_2 (A^T x - w)_2 \\ \vdots \\ A_{n-1} (A^T x - w)_{n-1} \end{pmatrix} = A (A^T x - w) = 0. \quad (7)$$

The right hand side of Eq. (7) transforms into

$$\begin{pmatrix} g_1^T \\ g_{s_1}^T \\ \vdots \\ g_{s_k}^T \end{pmatrix} = \begin{pmatrix} \beta & w_1 & w_2 & \cdots & w_{n-1} \\ b_1 & A_{1,1} & A_{1,2} & \cdots & A_{1,n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ b_k & A_{k,1} & A_{k,2} & \cdots & A_{k,n-1} \end{pmatrix},$$

$$g_{s_i}^T = (b_i \quad A_{i,1} \quad A_{i,2} \quad \cdots \quad A_{i,n-1}),$$

where

$$g_1^T = (\beta \quad w_1 \quad w_2 \quad \cdots \quad w_{n-1}),$$

$$\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} g_{s_1}(1) \\ \vdots \\ g_{s_k}(1) \end{pmatrix},$$

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{k,1} & A_{k,2} & \cdots & A_{k,n-1} \end{pmatrix},$$

and g_1^T is a gene with a missing value (depicted as β in the first location of g_1^T) and $g_{s_i}^T, i = 1, 2, \dots, k$, are the k -nearest neighbor gene vectors for g_1^T . Solutions of Eq. (7) involve the generalized inverse. If matrix A is invertible, we have

$$A A^T x = A w \quad (8)$$

and the solution

$$x = (A A^T)^{-1} A w = (A^T)^\dagger w, \quad (9)$$

where $(A^T)^\dagger = (A A^T)^{-1} A$, and $(A^T)^\dagger$ is the pseudoinverse of A^T . The missing value (β) can then be solved as follows:

$$\beta = \sum_{i=1}^{n-1} x_i b_i = b^T (A^T)^\dagger w. \quad (10)$$

Table 1: Error comparison of the standard MATLAB solvers vs. Jacobi SVD

Test Case	α	$\ AA^\dagger A - A\ _\infty$	$\ A^\dagger AA^\dagger - A^\dagger\ _\infty$	$\ (AA^\dagger)^T - AA^\dagger\ _\infty$	$\ (A^\dagger A)^T - A^\dagger A\ _\infty$	$\ Ax - b\ _\infty$
$inv(A)$	n/a	5.8272E+00	4.1592E+20	1.2275E+02	3.1764E+13	5.9128E+04
$pinv(A)$	n/a	7.1710E-06	7.5940E+08	4.7020E-04	9.3850E-04	1.6750E-04
Jacobi SVD	1.00	1.1430E-05	3.7480E+07	1.4290E-04	1.0210E-03	2.3480E-04

Table 2: Error comparison of Jacobi SVD vs. SSVD

Test Case	α	$\ AA^\dagger A - A\ _\infty$	$\ A^\dagger AA^\dagger - A^\dagger\ _\infty$	$\ (AA^\dagger)^T - AA^\dagger\ _\infty$	$\ (A^\dagger A)^T - A^\dagger A\ _\infty$	$\ Ax - b\ _\infty$
Jacobi SVD	0.55	3.6440E-10	2.8770E-02	1.9820E-08	1.5680E-08	9.6300E-05
	0.60	6.5320E-10	2.6639E+00	1.1930E-07	6.9380E-08	4.0300E-05
	0.65	6.6790E-09	2.7378E+02	4.5570E-07	5.0710E-07	1.4780E-05
	0.70	2.3680E-08	1.7679E+03	2.1580E-06	3.4820E-06	5.8480E-06
	0.75	4.1550E-07	1.6235E+04	3.1180E-05	2.6530E-05	1.0030E-05
SSVD	0.55	2.9390E-10	1.9540E-02	3.4340E-09	2.3480E-08	9.6300E-05
	0.60	1.3730E-09	3.9960E-01	4.9080E-09	3.9740E-08	4.0310E-05
	0.65	9.6320E-09	1.6243E+01	2.0710E-07	6.5240E-08	1.4770E-05
	0.70	2.7090E-08	6.9364E+02	1.0470E-06	1.0320E-06	5.9650E-06
	0.75	1.4630E-07	1.5730E+04	2.2050E-06	1.1560E-05	5.0670E-06

2.3 Improvement of the SVD Solver

The result obtained by the previous method to calculate the pseudoinverse of the matrix $A \in R^{m \times n}$,

$$A^\dagger = V \begin{bmatrix} \Sigma_{r_A}^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T = V_{r_A} \Sigma_{r_A}^{-1} U_{r_A}^T, \quad (11)$$

does not satisfy the 4 Moore-Penrose equations [1]:

$$\begin{aligned} AA^\dagger A &= A, \\ A^\dagger AA^\dagger &= A^\dagger, \\ (AA^\dagger)^T &= (AA^\dagger), \\ (A^\dagger A)^T &= (A^\dagger A). \end{aligned}$$

As a result, if the size of the matrix is increased, the number of computational errors is also increased—that is, the SVD results become less accurate. There are five different ways to test the accuracy of the pseudoinverse:

$$\begin{aligned} &\|AA^\dagger A - A\|_\infty, \\ &\|A^\dagger AA^\dagger - A^\dagger\|_\infty, \\ &\|(AA^\dagger)^T - AA^\dagger\|_\infty, \\ &\|(A^\dagger A)^T - A^\dagger A\|_\infty, \\ &\|Ax - b\|_\infty. \end{aligned}$$

In this paper, the Hilbert matrix,

$$A = H_{200 \times 200} = \left(\frac{1}{i+j+1} \right)_{200 \times 200}, \quad (12)$$

is used to benchmark a solver's robustness, scalability and accuracy. As shown in Table 1, $pinv(A)$ does not satisfy $A^\dagger AA^\dagger = A^\dagger$ and $inv(A)$ does not satisfy any of the 4 Moore-Penrose equations, because—in both cases—their respective errors are too large.

The Jacobi SVD method is an improvement upon the commonly implemented QR SVD solver. The procedure for the Jacobi SVD solver is as follows:

$$A = Q \begin{bmatrix} R_{m \times n} \\ 0_{(m-n) \times n} \end{bmatrix}, \quad (13)$$

where $R_{m \times n}$ is the upper triangular matrix. If $r_A < n$, $R_{m \times n}^T$ is further decomposed by QR SVD to get

$$\begin{aligned} R_{m \times n}^T &= P \begin{bmatrix} \tilde{R}_{r_A \times r_A} \\ 0 \end{bmatrix}, \\ A &= Q \begin{bmatrix} \tilde{R}_{r_A \times r_A} & 0 \\ 0 & 0 \end{bmatrix} P^T. \end{aligned} \quad (14)$$

Jacobi SVD is then used to solve $R_{m \times n}$ or $\tilde{R}_{r_A \times r_A}$. From Eq. (13)-(14) we have

$$R_{m \times n} = U_R \Sigma_{r_A} V_R^T, \quad A = Q_{r_A} U_R \Sigma_{r_A} V_{R,r_A}^T, \quad (15)$$

or

$$\tilde{R}_{m \times n} = U_{\tilde{R}} \Sigma_{r_A} V_{\tilde{R}}^T, \quad A = Q_{r_A} U_{\tilde{R}} \Sigma_{r_A} V_{\tilde{R},r_A}^T P_{r_A}^T, \quad (16)$$

and the pseudoinverse (A^\dagger) is

$$A^\dagger = V_R \Sigma_{r_A}^{-1} U_R^T Q_{r_A}^T \quad \text{or} \quad A^\dagger = P V_{\tilde{R}} \Sigma_{r_A}^{-1} U_{\tilde{R}}^T Q_{r_A}^T. \quad (17)$$

As seen in Table 1, the error associated with the Jacobi SVD solver for $A^\dagger AA^\dagger = A^\dagger$ is smaller than the results produced by the standard MATLAB solutions, yet this value is still too large to be acceptable.

SSVD is a further improvement upon the Jacobi SVD solver, aiming to reduce the overall size of these errors. Since U and V from Eq. (11) are orthogonal matrices, they do not lead to an error, which means the computational errors are

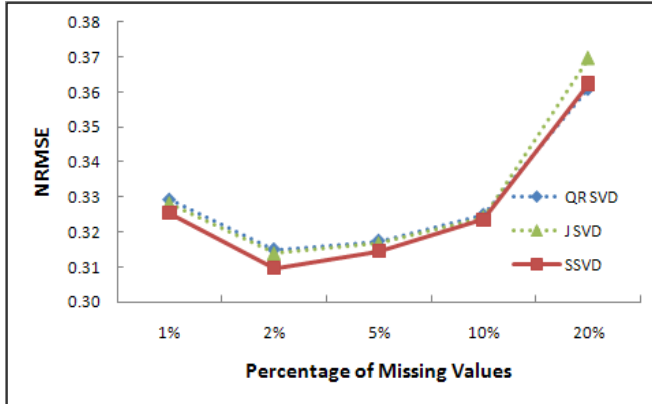


Fig. 1. NRMSE result for the YO microarray dataset

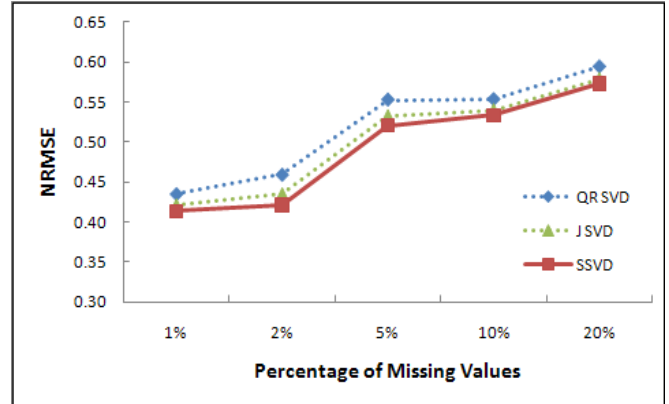


Fig. 2. NRMSE result for the CU microarray dataset

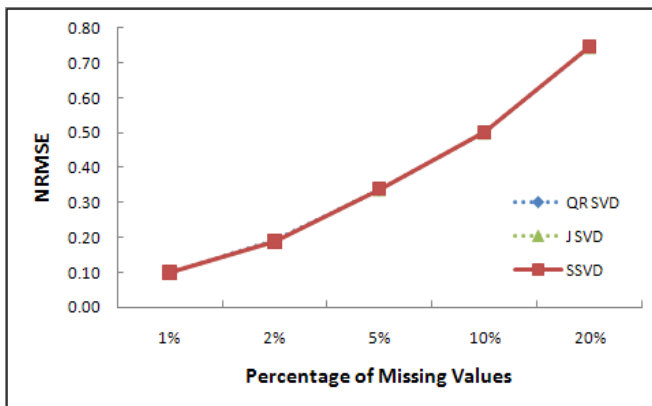


Fig. 3. NRMSE result for the RO microarray dataset

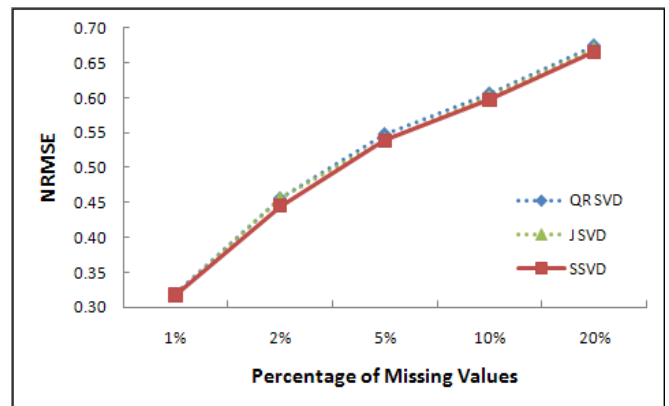


Fig. 4. NRMSE result for the SP microarray dataset

Table 3: NRMSE comparison of QR SVD, Jacobi SVD and SSVD

Test Case	SVD Solver	1%	2%	5%	10%	20%
YO.Calcineurin/Crzlp	QR SVD	0.3292	0.3149	0.3174	0.3248	0.3610
	Jacobi SVD	0.3282	0.3139	0.3167	0.3242	0.3696
	SSVD	0.3254	0.3097	0.3145	0.3235	0.3624
CU.Growth-regulator	QR SVD	0.4349	0.4591	0.5526	0.5531	0.5938
	Jacobi SVD	0.4214	0.4351	0.5318	0.5393	0.5779
	SSVD	0.4134	0.4209	0.5208	0.5334	0.5727
RO.Cellline	QR SVD	0.0988	0.1907	0.3390	0.5021	0.7467
	Jacobi SVD	0.0991	0.1901	0.3391	0.5017	0.7463
	SSVD	0.0988	0.1882	0.3391	0.5010	0.7457
SP.Alpha	QR SVD	0.3186	0.4557	0.5481	0.6064	0.6753
	Jacobi SVD	0.3183	0.4556	0.5398	0.6010	0.6702
	SSVD	0.3175	0.4454	0.5387	0.5974	0.6657

produced by $\Sigma_{r_A}^{-1}$. It is important to note that matrix Σ_{r_A} is dependent on the accuracy of the system executing its implementation; if the singular values $\sigma_i < \epsilon$ (ϵ is the error bound), the system will set it to zero, then Eq. (11) will contain computational errors.

Suppose the singular values of matrix A are $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{r_A}$, then

$$\Sigma_{r_A}^{-1} = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \dots, \frac{1}{\sigma_{r_A}}, 0, \dots, 0\right). \quad (18)$$

Since the cumulative error is $\epsilon_i, i = 1, 2, \dots, r_A$, the singu-

lar values of matrix A will become $\sigma_i + \varepsilon_i, i = 1, 2, \dots, r_A$. Therefore,

$$\Sigma_{r_A}^{-1} = \text{diag}\left(\frac{1}{\sigma_1 + \varepsilon_1}, \frac{1}{\sigma_2 + \varepsilon_2}, \dots, \frac{1}{\sigma_{r_A} + \varepsilon_{r_A}}, 0, \dots, 0\right), \quad (19)$$

where the maximum computing error is

$$\varepsilon(\Sigma_{r_A}^{-1}) = \text{diag}\left(\dots, \frac{|\varepsilon_1|}{\sigma_1|\sigma_1 + \varepsilon_1|}, \frac{|\varepsilon_2|}{\sigma_2|\sigma_2 + \varepsilon_2|}, \dots, \frac{|\varepsilon_k|}{\sigma_k|\sigma_k + \varepsilon_k|}, 0, \dots, 0\right). \quad (20)$$

If $\sigma_i + \varepsilon_i$ is close to zero, the error is significantly magnified, hence it should be set to zero. Eq. (19) becomes as follows:

$$\Sigma_{r_A}^{-1}(\alpha) = \text{diag}\left(\frac{1}{\sigma_1 + \varepsilon_1}, \frac{1}{\sigma_2 + \varepsilon_2}, \dots, \frac{1}{\sigma_k + \varepsilon_k}, 0, \dots, 0\right), \\ \sigma_i + \varepsilon_i \leq \text{eps}^\alpha, i = k + 1, \dots, r_A. \quad (21)$$

Experimental results using Eq. (21) are shown in Table 2. When $\alpha = 0.75$, the errors associated with SSVD are much smaller than those presented in Table 1. When $\alpha = 0.55$, the pseudoinverse of Hilbert matrix ($A = H_{200 \times 200}$) does satisfy the 4 Moore-Penrose equations, yet the solution of $\|Ax - b\|_\infty$ for $\alpha = 0.55$ is worse than $\alpha = 0.75$, which is due to the removal of more singular values that are close to zero. Therefore, α cannot be too small, and the empirically chosen value of $\alpha = 0.75$ is used in this research.

3. Numerical Results

The normalized root mean squared error (NRMSE) was used to measure the accuracy of the results from the test cases.

$$\text{NRMSE} = \sqrt{\frac{\text{mean}[(\gamma_{\text{estimated}} - \gamma_{\text{known}})]^2}{\text{std}[\gamma_{\text{known}}]}}, \quad (22)$$

where $\gamma_{\text{estimated}}$ are the estimations for missing values, and γ_{known} are the known values. The mean and the standard deviation are calculated over missing values in the whole dataset.

The NRMSE test results of Eq. (22) for various percentages (1%, 2%, 5%, 10% and 20%) of missing values for QR SVD, Jacobi SVD and SSVD are presented in Table 3. For an overwhelming majority of test cases, the SSVD method generates more accurate recovery results than QR SVD, and, for all test cases, SSVD consistently performed better than Jacobi SVD. Only in the YO.Calcineurin/Crzlp test case for 20% of missing values and the RO.Cellline test case for 5% of missing values did QR SVD outperform our proposed SSVD solver; however, the difference in performance for the latter test case is so insignificant that it may be regarded as an equal level of performance between the two solvers.

The graphical representations of the results from Table 3 are located in Fig. 1 through Fig. 4. Note that Jacobi SVD is referred to as J SVD within the legend of a figure. Each figure is oriented with the experimental NRMSE results in

the y-axis and the various percentages of missing values in the x-axis. A data trend favoring the lower end of the NRMSE scale is favorable because this represents a series of experimental results with smaller levels of erroneous estimations. The scale of each figure is not consistent, thus is insufficient to gauge the performance between datasets based solely on the distance of the separation between their respective data trend lines. These figures further illustrate the improvement in accuracy associated with the SSVD solver when tested with the YO.Calcineurin/Crzlp, CU.Growth-regulator, RO.Cellline and SP.Alpha microarray datasets, respectively.

Fig. 2 depicts the most exciting results of the four microarray datasets. The difference in accuracy between these solvers for these specific test cases is quite substantial, granting significant increases in the accuracy of missing value estimation. Fig 1. and Fig 4. show SSVD's typical outcome, which is a marginal increase in accuracy with respect to Jacobi and QR SVD. Fig. 3 illustrates the worst-case scenario for SSVD—the level of accuracy is nearly equal, yet slightly better, to that of the QR SVD solver.

4. Conclusion

We have successfully developed a scalable solver for estimating the missing values of DNA microarray datasets. For nearly all the test cases across all configurations of missing value percentages, SSVD provides more accurate recovery results than Jacobi and QR SVD. The numerical results presented in this paper strongly suggests that SSVD is a robust, scalable and accurate solver. One would be safe to assume that the benefits from SSVD may be realized in many other disciplines and not those limited to missing value estimation.

References

- [1] A. Albert, *Regression and the Moore-Penrose pseudoinverse*, Academic Press, INC, New York, 1972.
- [2] T. H. BZ, B. Dysvik and I. Jonassen, *LSimpute: accurate estimation of missing values in microarray data with least squares methods*, Nucleic Acids Res 32 (3) (2004) e34.
- [3] Z. Cai, M. Heydari and G. Lin, *Iterated local least squares microarray missing value imputation*, Journal of Bioinformatics and Computational Biology 4 (5) (2006) 935–957.
- [4] Z. Drmac and K. Veselic, *New Fast and Accurate Jacobi SVD Algorithm I*, SIAM Journal on Matrix Analysis and Applications, 29 (4) (2008) 1322–1342.
- [5] Z. Drmac and K. Veselic, *New Fast and Accurate Jacobi SVD Algorithm II*, SIAM Journal on Matrix Analysis and Applications, 29 (4) (2008) 1343–1362.
- [6] Q. Fen and J. Lee, *Dynamic Methods for Missing Value Estimation for DNA Sequences*, International Conference on Computational & Information Sciences (ICIS), 2010 442–445.
- [7] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown and D. Botstein, *Imputing missing data for gene expression arrays*, Technical Report, Division of Biostatistics, Stanford University, 1999.
- [8] P. Johansson and J. Hakkinen, *Improving missing value imputation of microarray data by using spot quality weights*, BMC Bioinformatics 7 (2006) 306.

- [9] H. Kim, G. H. Golub and H. Park, *Missing value estimation for DNA microarray gene expression data: local least squares imputation*, *Bioinformatics* 21 (2) (2005) 187–198.
- [10] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman, *Missing value estimation methods for DNA microarrays*, *Bioinformatics* 17 (6) (2001) 520–525.
- [11] D. Yoon, E. Lee and T. Park, *Robust imputation method for missing values in microarray data*, *BMC Bioinformatics* 8 (Suppl 2):S6 (2007).
- [12] A. Zien, J. Fluck, R. Zimmert and T. Lengauer, *Microarrays: how many do you need?*, *Journal of Computational Biology* 10 (3-4) (2003) 653–667.

Finding Biomarkers for Non-Small Cell Lung Cancer Diagnosis with Novel Data Mining Techniques

Quoc-Nam Tran[†], Lamar (Texas State) University, USA.

Abstract—Non-small cell lung carcinoma (NSCLC) is the most common cause of worldwide cancer premature death with a very low survival rate of 8%-15%. Patients with an early stage diagnosis can have up to four times the survival rate of 40%-55%. Hence, discovering cost-effective biological markers that can be used to improve the diagnosis and prognosis of the disease is an important clinical challenge.

Significant progress has been made to address this challenge. Some sets of biomarkers were identified in the last few years ranging from 5-gene signatures to 133-gene signatures. Since datasets of gene-expression profiles typically have tens of thousands of genes for just few hundreds of patients, this type of datasets will create many technical challenges impacting the accuracy of the diagnostic prediction. A typical molecular sub-classification method for lung carcinomas would have a low predictive accuracy of 68%-71%.

In this paper, we present a new data mining method that finds genetic markers and uses the markers to predict with up to 100% accuracy whether a patient has NSCLC and the sub-type of cancer in case the patient has NSCLC. Our method overcomes many challenges arose from datasets of gene-expression profiles. The new method discovers novel genetic changes that occur in lung tumors using gene-expression profiles. We discovered that a small set of nine gene-signatures (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5) from the dataset of 12,600 gene-expression profiles of NSCLC acts like an inference basis for NSCLC lung carcinoma and hence can be used as genetic markers. This very small and *previously unknown set of biological markers* gives an almost perfect predictive accuracy for the diagnosis of the disease.

While proteins encoded by some of these gene-signatures (e.g., JAG1 and MAPRE2) have been showed to involve in the signal transduction of cells and proliferative control of normal cells, specific functions of proteins encoded by other gene-signatures have not yet been determined. Therefore, this work opens new questions for structural and molecular biologists about the role of these gene-signatures for the disease.

Keywords-Mining gene-expression profiles in bioinformatics, lung cancer, diagnosis.

I. INTRODUCTION

In the last several years, one in four deaths in the United States is due to cancer, which makes cancer a major public health problem in the United States as well as many other parts of the world [1, 2]. Currently, cancer is a leading cause of death in the United States, second only to cardiovascular diseases. Last year, 1.48 million people were diagnosed with cancer, and 562,340 people died from cancer. The top five most common cancer-related deaths were due to lung, breast, prostate, colorectal and pancreatic cancer. Together, these five diseases accounted for over 50% of all cancer deaths in the United States in 2009. Lung cancer alone, with NSCLC as the most common cause of worldwide cancer premature death, killed over 160,000 people, more than the other four cancers put together. The disease has a very low survival rate of 8%-15%. Meanwhile, the survival rate for patients with early-stage disease increases to 40%-55% after surgery. That said, discovering cost-effective biological markers that can be used to improve the diagnosis and prognosis of the disease is an important clinical challenge [3].

NSCLC is sub-categorized as adenocarcinomas, squamous cell carcinomas, and large-cell carcinomas, of which adenocarcinomas are the most common [4]. The histopathological sub-classification of lung adenocarcinoma is challenging. For example, in one study independent lung pathologists agreed on lung adenocarcinoma sub-classification in only 41% of cases [5]. In another study, proportional hazard models identified an optimal set of 50 prognostic mRNA transcripts using a 5-fold cross-validation procedure. This signature was tested in an independent set of 36 squamous cell lung carcinomas (SCC) samples and achieved 84% specificity and 41% sensitivity with an overall predictive accuracy of 68%

[†] Supported in part by NSF award CCF-0917257.

[6]. Combining the SCC classifier with their adenocarcinoma prognostic signature gave a predictive accuracy of 71% in 72 NSCLC samples.

Multiple techniques have evolved over the past few years allow rapid measurement of gene expression and simultaneous high-throughput measurement of thousands of genes from several hundred samples. Different parts of the gene-protein relationship can be measured such as messenger RNA levels, protein expression and cellular metabolic activity. Some of the available genomic technologies include gene expression arrays, serial analysis of gene expression, single-nucleotide polymorphism analysis, and high-throughput capillary sequencing [3].

Gene-expression array analysis methodologies developed over the last few years have demonstrated that expression data can be used in a variety of class discovery or class prediction biomedical problems including those relevant to tumor classification [7, 8, 9, 10]. Data mining and statistical techniques applied to gene expression data have been used to address the questions of distinguishing tumor morphology, predicting post treatment outcome, and finding molecular markers for disease [11, 12, 13, 14].

However, gene expression profiles present many challenges for data mining both in finding differentially expressed genes, and in building predictive models because the datasets are highly multidimensional (12,600 dimensions in our study) and contain a small number of records (197 records in our study). Although microarray analysis tool can be used as an initial step to extract most relevant features, one has to avoid overfitting the data and deal with the very large number of dimensions of the datasets. The current challenges in analyzing gene-expression profiles, is illustrated in a method recently published in the Journal of Experimental & Clinical Cancer Research in July 2009 [15] where it used prior knowledge with support vector machine-based classification in diagnosis of lung cancer. The authors of [15] reported an accuracy of 98.51%-99.06% for their classification algorithm using 5 marker genes on a dataset of 31 malignant pleural mesothelioma (MPM) and 150 lung adenocarcinomas. Even though the method in [15] can differentiate between MPM and lung adenocarcinomas with high accuracy, it gives an accuracy of 70% when we added other types of NSCLC lung cancer including adenocarcinomas, squamous cell lung carcinomas and pulmonary carcinoids into consideration. Other researchers also limited themselves in differentiate two sub-types of NSCLC lung cancer such as between adenocarcinomas and squamous cell lung carcinomas.

This paper aims at a novel data mining method that finds cost-effective genetic markers and uses the markers to differentiate with very high accuracy *all sub-types of NSCLC lung cancer*. Comparing with a recent publication [16] in that the author uses currently available data mining techniques in Weka to find biomarkers for NSCLC lung cancer, we found that our new method finds significantly more cost-effective genetic markers and provides more accurate sub-classification of NSCLC lung cancer. Comparison with SAM [17], a popular method for significance analysis of microarrays, is also provided in Section III.

Among the nine gene-signatures found by our new method (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5), proteins encoded by some of these gene-signatures (e.g., JAG1 and MAPRE2) have been showed to involve in the signal transduction of cells and proliferative control of normal cells [18]. It has also been found that MAPRE2 is highly expressed in pancreatic cancer cells, and seems to be involved in perineural invasion [19]. However, specific functions of proteins encoded by other gene-signatures have not yet been determined. Hence, this work opens new questions for structural and molecular biologists about the role of these gene-signatures for the disease.

II. A NEW DATA MINING METHOD FOR SIGNIFICANT GENES SELECTION & SUB-CLASSIFICATION

Before presenting our new algorithm for finding genetic markers and predicting NSCLC lung cancer, we will address the challenges one has to overcome while working with gene-expression profile datasets. Basic information about Gini indexes and classification algorithms can be found in many data mining books [20, 21, 22].

A. Solving the bias due to the order of classes

The first challenge that arose from the gene-expression datasets is the bias due to the order of cancer types or classes in data mining's terminology. Let's consider a

Range/Class	C_1	C_2	C_3
R_1	4	6	30
R_2	6	30	4
R_3	0	4	16

Table I
BIAS DUE TO THE ORDER OF CLASSES

simple example of expression profiles for a gene A in Table I where the gene dataset D has $d = 100$ elements

and three classes. The gene expression values were partitioned into three ranges. Clearly, the cancer types or classes can be labeled in any order. When this gene is ranked by current microarray analysis methodologies, for example by calculating the Gini index $gini_A(D) = \sum_{i=1}^m \frac{|R_i|}{d} \cdot gini(R_i)$, the first two rows contribute equally to the Gini index because $gini(R_i) = 1 - \sum_{j=1}^n p_{i,j}^2$ where $p_{i,j} = \frac{|C_{i,j}|}{|R_i|}$ is the relative frequency of class C_j in R_i , and $|\cdot|$ is the notation for cardinality [23]. We have the same problem when entropy is calculated instead of the Gini index. That said, when one just considers the probability distribution without taking into account the order of the classes, the first two partitions of expression profiles will contribute equally. Clearly, the two partitions should contribute differently because Partition R_1 says that 75% of patients with gene expression values within this range are classified into Class C_3 while Partition R_2 says that 75% of patients with gene expression values within this range are classified into Class C_2 . Hence, in order to have a robust gene selection method, one has to differentiate the partitions with different class orders because they have different amount of information.

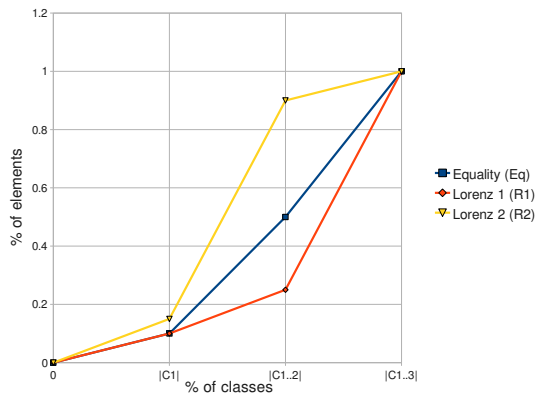


Figure 1. Lorenz curves

To solve this problem, we generalized the well known Lorenz curves, a common measure in economics to gauge the inequalities in income and wealth. In Figure 1, we illustrate how modified Lorenz curves and modified Gini coefficients are calculated. The Equality Polygon (Eq) is defined based on the percentages of elements in $|C_1|$, $|C_{1..2}| = |C_1| + |C_2|$, \dots , $|C_{1..n}| = \sum_{j=1}^n |C_j|$ at x -coordinates $0, 1/n, 2/n, \dots, 1$, where n is the number of classes and $|C_1| \leq |C_2| \leq \dots \leq |C_n|$. The Lorenz polygon of a partition, say R_i , is defined based on the percentage of elements in $|C_{i,1}|$, $|C_{i,1}| + |C_{i,2}|$, \dots , $\sum_{j=1}^n |C_{i,j}|$ at x -coordinates $0, 1/n, 2/n, \dots, 1$.

The Gini coefficient of a partition, say R_i , is defined as $(\int_0^1 L(R_i) \cdot dx - \int_0^1 Eq \cdot dx) / \int_0^1 Eq \cdot dx$. One can easily see that the partitions with different class orders are now differentiated.

B. Solving the bias due to the order of gene expression values

Another technical challenge for microarray analysis methodologies comes from the order of discretized gene expression values. Let's consider another simple example

Class/Range	C_1	C_2	C_3	Class/Range	C_1	C_2	C_3
R_1	3	0	0	R_1	3	0	0
R_2	0	100	0	R_2	4	0	0
R_3	4	0	0	R_3	0	100	0
R_4	0	0	5	R_4	0	0	5

Table II
BIAS DUE TO THE ORDER OF GENE EXPRESSION VALUES

of gene-expression profiles for two genes in Table II with three classes. The gene expression values were discretized into four ranges. In contrast to the previous challenge, the ranges of gene-expression values do follow some order. When this genes are ranked by current microarray analysis methodologies, for example by calculating the Gini index of gene A using dataset D $gini_A(D) = \sum_{i=1}^m \frac{|R_i|}{d} \cdot gini(R_i)$ where $d = |D|$, the two genes would have the same rank. Clearly, the gene-expression profiles on the right hand side of Table II have a more harmonic distribution with respect to the rows in comparison with the gene on the left. That said, these two genes should be ranked differently.

To solve this problem, we generalized the Gini coefficients by taking into account the splitting status and the Gini ratio. The splitting status of D with respect to the attribute A is calculated as

$$split_A(D) = 1 - \sum_{i=1}^m \left(\frac{|R_i|}{d}\right)^2.$$

The Gini ratio of D with respect to the attribute A is defined as $LorenzGini(A) = \Delta gini(A) / split_A(D)$, where $\Delta gini(A) = gini(D) - gini_A(D)$ and $gini(D) = 1 - \sum_{j=1}^n \left(\frac{|C_j|}{d}\right)^2$.

Furthermore, to take into account the gene expression profiles with different value orders, the Gini coefficient is calculated as $gini_A(D) = \sum_{i=1}^m \frac{|R_i|}{d} \cdot \delta(i) \cdot gini(R_i)$, where $\delta(i)$ is the sum of the normalized distances between the row i and rows $i-1, i+1$. The coefficient $\delta(i)$ is used as a weight to emphasize a row when it is close to its neighbors.

C. New Algorithm

Input: A gene-expression profiles dataset D with up to 34,000 dimensions.

Output: A small subset of genes as genetic markers and a prediction model for NSCLC lung cancer

Step1: Discretize the gene-expression profile values.

Step2: Select genetic markers by using the genes with highest ranking LorenzGini.

Step3: Build the prediction model to classify patients using the genetic markers.

A threshold can be used for controlling the number of significant genes for genetic markers. The splitting status of dataset D with respect to a gene A can be calculated as a by-product when the reduction in impurity of D with respect to the attribute A is calculated. Therefore, the time complexity and space complexity of the algorithm are the same as the complexities of Gini index algorithm.

Our method has been implemented in Maple and Weka [24, 25]. In the next section, we will present our experiment with a dataset of gene-expression profiles of NSCLC from the mRNA expression profiles.

Notice that our new method works for any dataset with ≥ 2 classes. For any number of classes, even when the number of classes is equal to 2, the new method is completely different with other microarray analysis methodologies.

III. EXPERIMENTATION

A. mRNA Materials

To test and validate our algorithm, we extract the gene-expression profiles of NSCLC from the mRNA expression profiles in [26] in that a total of 203 snap-frozen lung tumors ($n=186$) and normal lung ($n=17$) specimens were used to create the dataset. Of these, 125 adenocarcinoma samples were associated with clinical data and with histological slides from adjacent sections. The 203 specimens include histologically defined lung adenocarcinomas ($n=139$), squamous cell lung carcinomas ($n=21$), pulmonary carcinoids ($n=20$), and normal lung ($n=17$) specimens. Total RNA extracted from samples was used to generate cRNA target, subsequently hybridized to human U95A oligonucleotide probe arrays according to standard protocols. As the result, we obtained a dataset of 12,600 gene-expression profiles for 197 patients.

B. Finding genetic markers

Using the algorithm described in the previous section, we select 250 genes with the highest LorenzGini indexes. To further reduce the size of the gene subsets and to improve the prediction accuracy, we evaluate different combinations of genes to identify an optimal subset in terms of accuracy for the Bayesian Net classification. The gene subsets to be evaluated are generated using different subset search techniques. We use Best First and Greedy search methods in the forward and backward directions. Greedy search considers changes local to the current subset through the addition or removal of genes. For a given parent set, a greedy search examines all possible child subsets through either the addition or removal of genes. The child subset that shows the highest goodness measure then replaces the parent subset, and the process is repeated. The process terminates when no more improvement can be made. Best First search is similar to greedy search in that it creates new subsets based on the addition or removal of genes to the current subset with the ability to backtrack along the subset selection path to explore different possibilities when the current path no longer shows improvement. To prevent the search from backtracking through all possibilities in the gene space, a limit is placed on the number of non-improving subsets that are considered. In our evaluation we chose a limit of five.

The algorithm returns a set of nine genes (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5) from the dataset of 12,600 gene-expression profiles of NSCLC. We exploit this small set of genes to differentiate all sub-types of NSCLC lung cancer.

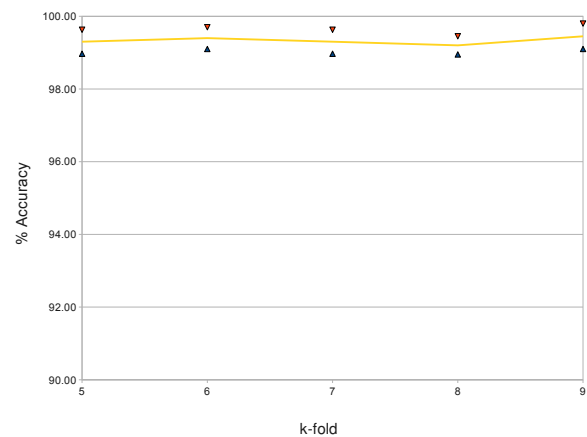


Figure 2. Accuracy of sub-classifications with standard deviations

To build the classification model, we used Bayesian Network (BayesNet), which is structured as a combination

of a directed acyclic graph of nodes and links, and a set of conditional probability tables. Nodes represent features or classes, while links between nodes represent the relationship between them. Conditional probability tables determine the strength of the links. There is one probability table for each node (feature) that defines the probability distribution for the node given its parent nodes. If a node has no parents the probability distribution is unconditional. If a node has one or more parents the probability distribution is a conditional distribution, where the probability of each feature value depends on the values of the parents.

Figure 2 shows the averaged accuracies of the gene expression profile classification using Bayesian Net classification together with their standard deviations. To test the accuracy of classification models, we use k -fold cross validation, which is a common method for estimating the error of a model on benchmark medical data sets. The reason for using this testing approach is that when a model is built from training data, the error on the training data is a rather optimistic estimate of the error rates the model will achieve on unseen data. The aim of building a model is usually to apply the model to new, unseen data—we expect the model to generalize to data other than the training data on which it was built. Another reason for using this testing approach is that the available medical data sets are small and no test data set is available. It is well-known that k -fold cross-validation is very useful for this type of data sets.

For a reliable evaluation of the accuracy, we test the classification algorithm for many values of k . More precisely, we test for $k = 5..9$. For each value of k , the data set D is randomly divided into k subsets D_1, D_2, \dots, D_k . We leave out one of the subsets $D_i, i = 1..k$ each time for being used as a test data set for cross validation. The remaining subset $\cup_{j \neq i} D_j$ is used to build the model. The cross validation accuracy computed for each of the k test samples are then accumulated to give the k -fold estimate of the cross validation accuracy. To ease the effects of the random partitions on the data set, this whole process is repeated 10 times with different random seeds and the results are then averaged to give the estimated accuracy of the comparing algorithms in Figure 2.

During the validation process, all patients with lung adenocarcinomas were correctly predicted, all patients except one with squamous cell lung carcinomas were correctly predicted, all patients with pulmonary carcinoids were correctly predicted, and all patients with normal lung specimens were correctly predicted. The only false prediction for random seed 1 was a patient with

squamous cell lung carcinomas but incorrectly predicted as adenocarcinomas. As we can see, this very small set of genes gives an almost perfect predictive accuracy for the diagnosis of the disease. When the number of genes is further reduced or increased, the accuracy starts to decline. That said, this set of nine genes acts like an inference basis for NSCLC lung carcinoma and hence can be used as genetic markers.

C. Comparing with other gene selection methods

We now investigate the classifying accuracy of the significant genes generated by LorenzGini with respect to the size of the reduced microarray datasets. Comparing with a recent publication [16] in that the author uses currently available data mining techniques in Weka to find biomarkers for NSCLC lung cancer, we found that our new method finds significantly more cost-effective genetic markers and provides more accurate sub-classification of NSCLC lung cancer. We also compare our method with SAM using the same dataset for NSCLC lung cancer. SAM combines t-test and permutations to calculate a False Discovery Rate to provide a subset of genes that are considered significant [17]. Using SAM, we select four sets of 50, 100, 150, 200 and 250 most significant genes by using the parameter values of 0.556, 0.458, 0.4188, 0.383 and 0.3568, respectively.

We then use the Bayesian Net classification in Weka to check the accuracy of the most significant gene sets generated by LorenzGini and SAM [25]. Besides our fresh implementation of LorenzGini algorithms, simple converters were written to connect SAM and Weka. For a reliable evaluation of the accuracy, we test the classification algorithm for many values of k as specified in our validation plan.

Figure 3 shows the accuracy of the gene expression profile classification using Bayesian Net algorithm on SAM's gene sets and on LorenzGini's gene sets with 50 genes. As we can see, the classifying accuracy has been improved with the LorenzGini's gene selections. We also observed that the accuracy of the gene expression profile classification using Bayesian Net algorithm on SAM's gene sets declined when the number of genes is reduced to 50 and below. In contrast, the accuracy of the gene expression profile classification using LorenzGini's gene sets is stable even when the number of genes is reduced to 9, which has the highest accuracy. This observation is also true for other classification methods.

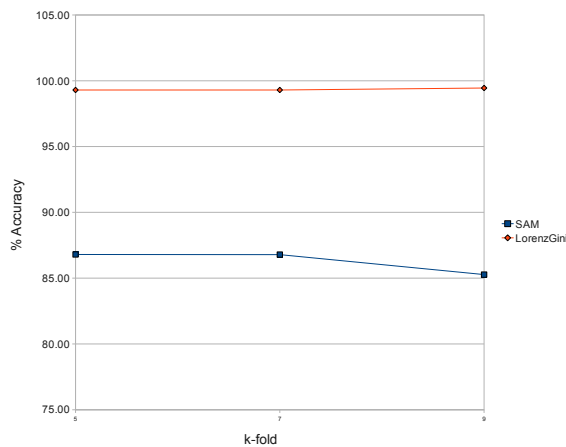


Figure 3. SAM's & LorenzGini's gene sets classified by Bayesian Net

IV. CONCLUSION

We presented a method that can find cost-effective biological markers as quantifiable measurements for an almost perfect predictive accuracy of NSCLC lung cancers. As cancers are complicated, one can only predict the status using a combination of many genes. The genes we discovered as genetic markers (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5) are different with previously known results. Furthermore, proteins encoded by some of these gene-signatures (e.g., JAG1 and MAPRE2) have been showed to involve in the signal transduction of cells and proliferative control of normal cells while specific functions of proteins encoded by other gene-signatures have not yet been determined. Therefore, this work opens new questions for structural and molecular biologists about the role of these gene-signatures for the disease.

REFERENCES

- [1] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun, "Cancer statistics, 2007," *CA Cancer J Clin*, vol. 57, pp. 43–66, 2007.
- [2] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA Cancer J Clin*, vol. 59, pp. 225–249, 2009.
- [3] S. Singhal, D. Miller, S. Ramalingam, and S.-Y. Sun, "Gene expression profiling of non-small cell lung cancer," *Lung cancer*, vol. 60, no. 3, pp. 313–324, 2008.
- [4] J. D. Watson, "The human genome project: past, present, and future," *Science*, vol. 248, pp. 44–49, 1990.
- [5] F. S. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, pp. 286–290, 2003.
- [6] B. Cox, T. Kislinger, and E. A., "Integrating gene and protein expression data: pattern analysis and profile mining," *Methods*, vol. 35, no. 3, pp. 303–314, 2005.
- [7] A. Butte, "The use and analysis of microarray data," *Nature Review Drug Discovery*, vol. 1, no. 12, pp. 951–960, 2002.
- [8] G. Piatetsky-Shapiro and P. Tamayo, "Microarray data mining: Facing the challenges," *SIGKDD Explorations*, vol. 5, no. 2, 2003.
- [9] S. Ramaswamy and T. R. Golub, "DNA microarrays in clinical oncology," *Journal of Clinical Oncology*, vol. 20, pp. 1932–1941, 2002.
- [10] P. Tamayo and S. Ramaswamy, "Cancer genomics and molecular pattern recognition," in *Expression profiling of human tumors: diagnostic and research applications*, M. Ladanyi and W. Gerald, Eds. Humana Press, 2003.
- [11] W. Dalton and S. Friend, "Cancer biomarkers—an invitation to the table," *Science*, vol. 312, no. 5777, pp. 1165–1168, 2006.
- [12] T. J. Yeatman, "Predictive biomarkers: Identification and verification," *J Clin Oncol*, vol. 27, no. 17, pp. 2743–2744, 2009.
- [13] K. Shedden, J. Taylor, S. Enkemann, M. Tsao, T. Yeatman, W. Gerald, S. Eschrich, I. Jurisica, T. Giordano, D. Misek, A. Chang, C. Zhu, S. D., S. Hanash, F. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V. Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V. Seshan, M. Meyerson, R. Kuick, K. Dobbin, T. Lively, J. Jacobson, and D. Beer, "Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study," *Nat Med*, vol. 14, pp. 822–827, 2008.
- [14] B. Kim, H. J. Lee, H. Y. Choi, Y. Shin, S. Nam, G. Seo, D.-S. Son, J. Jo, J. Kim, J. Lee, J. Kim, K. Kim, and S. Lee, "Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data," *Cancer Res*, vol. 67, pp. 7431–8, 2007.
- [15] P. Guan, D. Huang, M. He, and B. Zhou, "Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method," *J Exp Clin Cancer Res*, vol. 28, no. 103, pp. 1–7, 2009.
- [16] N.-P. Tran, "Using data mining techniques for improving non-small cell lung cancer classification," *Journal of Computing Sciences in Colleges*, 2010, accepted for publication.
- [17] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 5116–5121, 2001.
- [18] F. J., A. Chung, H. Xu, J. Zhu, H. Outtz, J. Kitajewski, Y. Li, X. Hu, and L. Ivashkiv, "Autoamplification of notch signaling in macrophages by tlr-induced and rbp-j-dependent induction of jagged1," *J Immunol*, vol. 185, no. 9, pp. 5023–31, 11 2010.
- [19] A. I., G. S., D. T., K. T., B. K., G. N.A., F. H., and K. J., "The microtubule-associated protein mapre2 is involved in perineural invasion of pancreatic cancer cells," *Int J Oncol*, vol. 35, no. 5, pp. 1111–6, 2009.
- [20] J. Han and K. Micheline, *Data Mining: Concepts and Techniques*, 2nd ed. Morgan Kaufmann, 2006.
- [21] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, 2008.
- [22] N. Ye, Ed., *The Handbook of Data Mining*. Lawrence Erlbaum Associates, 2003.
- [23] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, 1984, monterey, CA.
- [24] B. W. Char, K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt, *Maple V Language Reference Manual*. Springer Verlag, 1991.
- [25] "http://www.cs.waikato.ac.nz/ml/weka," 2009.
- [26] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 24, pp. 13 790–13 795, 2001.

Identification of Pseudo-Periodic Gene Expression Profiles

Li-Ping Tian¹, Li-Zhi Liu², and Fang-Xiang Wu^{2,3*}

¹School of Information, Beijing Wuzi University,

No.1 Fuhe Street, Tongzhou District, Beijing, P.R. China

²Department of Mechanical Engineering, ³Division of Biomedical Engineering,
University of Saskatchewan, 57 Campus Dr., Saskatoon, SK S7N 5A9, CANADA

*Corresponding author: faw341@mail.usask.ca

Abstract— Time-course gene expression profiles associated with periodic biological processes should appear periodic. However, because of inherent problems with the experimental protocols measured gene expression data are actually pseudo-periodic, not exactly periodic. Therefore, identifying pseudo-periodically expressed gene from their time-course data could help understand the molecular mechanism of periodic biological processes. This paper proposes a method for identifying pseudo-periodic gene expression profiles. In the proposed method, a pseudo-periodic gene expression profile is modeled by a linear combination of trigonometric and exponential functions in time plus a Gaussian noise term. A two-step parameter estimation method is employed for estimating parameters in the model. On the other hand, non-pseudo periodic gene expression profiles are model by a constant plus a Gaussian noise term. The statistic F-testing is used to make a decision if a gene is pseudo-periodically expressed or not. Three biological datasets were employed to evaluate the performance of the proposed method. The results show that the proposed method can effectively identify pseudo-periodically expressed genes.

Keywords: time-course gene expression profiles, pseudo-periodically expressed gene, parameter estimation, F-testing

I. INTRODUCTION

DNA microarray experiments have been employed to produce gene expression profiles at a series of time points. Such time-course gene expression data provides a dynamic snapshot of most (if not all) of the genes

related to the biological development process. The analysis of such time-course gene expression data is helpful in understanding the mechanism of their associated biological process. Many time-course gene expression datasets have been collected from periodic biological processes. For periodic biological process, Furthermore, identifying periodically expressed gene from their time-course expression data could help understand the molecular mechanism of those biological processes [1,2].

In past decade, a number of methods have been proposed to identify periodically expressed genes. The discrete Fourier transform method is the earliest method for identifying periodically expressed genes [1, 2]. However, microarray experiments typically generate short time-course data. As pointed in [3], the frequency resolution by the discrete Fourier transform is often not adequate for resolving periodicities of interest. Recently periodic (trigonometric) functions are used to model periodic gene expression data.

There are typically two ways to match the models with data. In one way, many models with known parameters are created, and searching datasets is performed to find the expression profiles which match well with some of created models. For example, Authors in [4] proposed a method called CORRCOS which generates 101000 periodic synthetic models with different frequencies and phases. Each gene expression profile is compared to each of these 101000 models. The cross-correlation is used to measure the similarity between the synthetic model and gene expression profiles. The frequency and phase of the model most similar to the expression profile is assigned to the corresponding gene. Although it can identify periodically expressed gene, CORRCOS is too time-consuming and the cross-correlation is not real metric. Authors in [3] developed another algorithm named RAGE for detecting periodically expressed genes. Like

CORRCOS, RAGE is a synthetic model-based method. RAGE first estimates the frequency of expression profiles using autocorrelations of both the synthetic model and gene data. Then, RAGE generates a number of models with the estimated frequency over a variety of phases. The similarity between the synthetic model and gene expression profile is measured by a real metric called Hausdorff distance. Compared with CORRCOS, RAGE is less time-consuming [3].

These methods lack the statistical analysis. Wichert et al [5] proposed a statistical method to identify periodically expressed genes from their time-course gene expression profiles. The method models gene expression profiles also as sine functions. Instead of estimating nonlinear parameters (frequency) in the model, they used the Fisher g-test to find the best frequency. Based on Fisher g-test, several similar methods were also developed for identifying [6,7,8]. However, a recent research [9] concludes that the Fisher g-test is poor if the time-course data is short and/or that data length is not an integer number of periods. In [9], the data length is said to be short if it is less than 40 data points. By this criterion, most gene expression profiles are too short to use Fisher g-test. In addition, it is hard in practice to obtain gene expression profiles with an integer number of periods as the period might be unknown before collecting the data.

In another way, models with unknown parameters are employed and unknown parameters are estimated based on the data such that the models with estimated parameters match well with the data. However, it is challenging to estimate parameters which are nonlinear in a model such as trigonometric function. Recently we proposed a two-step parameter estimation method to estimate all parameters in trigonometric function models from gene expression profiles [10, 11]

In principle, expression profiles associated with periodic processes should appear periodic. However, because of inherent problems with gene expression experimental protocols [1,12, 13], measured gene expression data are actually pseudo-periodic, not exactly periodic. In this paper, a method is proposed for identifying pseudo-periodic gene expression profiles. In the proposed method, a pseudo-periodic gene expression profile is modeled by a linear combination of trigonometric and exponential functions in time plus a Gaussian noise term. This model is more complex than the one in [10, 11]. A new two-step parameter estimation method is employed for estimating parameters in the model. On the other hand, non-pseudo periodic gene expression profiles are modeled by a constant plus a Gaussian noise term. The statistic F-testing is used to make a decision if a gene is

pseudo-periodically expressed or not. Three biological datasets were employed to evaluate the performance of the proposed method.

II. METHODS

In this section, we first propose the model for pseudo-periodic gene expression profiles and then describe a two-step parameter estimation method for the proposed model. Finally a hypothesis testing is described to make a decision whether a gene expression profile is pseudo-periodic or not.

2.1 Model for pseudo periodic gene expression profiles

Let $x(t)$ ($t=1,2,\dots, m$) be a time-course gene expression profile generated from a periodic biological process, where m is the number of time points at which gene expression is measured. In this study, we always shift the mean of gene expression profiles to 0. To model pseudo-periodic gene expression profile, we adopt the linear combination of trigonometric and exponential functions plus a Gaussian noise term as follows:

$$x(t) = e^{\alpha t} [a \cos(\omega t) + b \sin(\omega t)] + \varepsilon(t) \quad (1)$$

where a and b are the coefficients of sine and cosine function, respectively; α is the decrease (increase) rate; ω is the frequency of periodic expression data; and $\varepsilon(t)$ represent random errors. This study assumes that the errors have a normal distribution independent of time with the mean of 0 and the variance of σ^2 . When $\alpha=0$, model (1) becomes

$$\begin{aligned} x(t) &= a \cos(\omega t) + b \sin(\omega t) + \varepsilon(t) \\ \text{or } x(t) &= A \sin(\omega t + \Phi) + \varepsilon(t) \end{aligned} \quad (2)$$

which are widely used to generate the synthetic periodic gene expression profiles [1-9]. However, because of inherent problems with gene expression experimental protocols we believed that model (1) is more reasonable.

Given a time-course gene expression profile $x(t)$ ($t=1, 2, \dots, m$), estimating parameters a , b , α and ω in model (1) is a nonlinear estimation problem as α and ω is nonlinear in the model. Nonetheless, our observation is that noise-free model (1)

$$x(t) = e^{\alpha t} [a \cos(\omega t) + b \sin(\omega t)] \quad (3)$$

can be viewed as the general solution of a following second order ordinary differential equation

$$\ddot{x}(t) + 2\alpha\dot{x}(t) + \gamma^2 x(t) = 0 \quad (4)$$

where $\gamma^2 = \omega^2 + \alpha^2$ and equation (4) is independent of a and b. Note that α and γ^2 are linear in equation (4) while a and b are linear in model (1). Therefore, we propose the following two-step parameter estimation methods to estimate parameters a, b, α and ω in model (1):

Step1: Based on equation (4), use linear least squares method to estimate parameters α and γ^2 , thus α and ω . In detail, let

$$X2 = \begin{bmatrix} \ddot{x}(1) \\ \vdots \\ \ddot{x}(l) \end{bmatrix}, \text{ and } X1 = \begin{bmatrix} \dot{x}(1) & x(1) \\ \vdots & \vdots \\ \dot{x}(l) & x(l) \end{bmatrix}$$

then by the least squares method, α and γ^2 are estimated as

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\gamma}^2 \end{bmatrix} = -(X1^T X1)^{-1} X1^T X2 \quad (5)$$

and thus ω is estimated

$$\hat{\omega} = \sqrt{\hat{\gamma}^2 - \hat{\alpha}^2} \quad (6)$$

As time-course gene expression data are discrete, the first and second derivatives $\dot{x}(t)$ and $\ddot{x}(t)$ are estimated by the central finite difference formula, respectively, as follows

$$\dot{x}(t) = \frac{x(t+1) - x(t-1)}{2\Delta} \quad \text{for } t=2, \dots, m-1 \quad (7)$$

$$\ddot{x}(t) = \frac{x(t+1) + x(t-1) - 2x(t)}{\Delta^2} \quad \text{for } t=2, \dots, m-1 \quad (8)$$

where Δ is time difference between two consecutive gene expression data points. From equations (7) and (8), $l=m-2$. Note that equations (7) and (8) are for evenly spaced time-course data. For unevenly spaced time-course data, equation (7) and (8) should be replaced by a modified formula which can be found in any numerical method textbooks. If the value of $\hat{\gamma}^2 - \hat{\alpha}^2$ calculated by (5) for a gene is negative, this gene will be judged not to be periodically expressed.

Step2: Substitute the estimated values of α and ω in Step 1 into equation (1). Apply the least squares method to model (1) to estimate parameters a and b. In detail, let

$$X = \begin{bmatrix} x(1) \\ \vdots \\ x(m) \end{bmatrix}, \text{ and } A = \begin{bmatrix} \cos(\Delta\hat{\omega}) & \sin(\Delta\hat{\omega}) \\ \vdots & \vdots \\ \cos(m\Delta\hat{\omega}) & \sin(m\Delta\hat{\omega}) \end{bmatrix}$$

by the linear least squares method, a and b are estimated as

$$\beta = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (A^T A)^{-1} (A^T X) \quad (9)$$

2.2 Hypothesis testing

To determine if a gene is pseudo-periodically expressed, we test the null hypothesis of

$$H_0: x(t) = \varepsilon(t) \quad (10)$$

versus the alternative hypothesis of

$$H_a: x(t) = e^{a t} [a \cos(\omega t) + b \sin(\omega t)] + \varepsilon(t) \quad (1)$$

In terms of the following F-statistic

$$F = \frac{m-2}{2} \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} - 1 \right) \quad (11)$$

where $\hat{\sigma}_0^2$ is the estimated variance of white noise in model (10) and is calculated as

$$\hat{\sigma}_0^2 = \frac{1}{m-1} X X^T \quad (12)$$

and $\hat{\sigma}_1^2$ is the estimated variance of white noise in model (1) and is calculated as

$$\hat{\sigma}_1^2 = \frac{1}{m-1} [X^T - A^T \beta]^T [X - A^T \beta] \quad (13)$$

As noise terms in both model (1) and (10) are normal white noise, F-statistic (11) follows the F-distribution with the degrees of freedom (2, m-2), according to statistics theory. When the value of F-statistic is large enough (greater than a threshold), model (10) is rejected, i.e., the gene expression profile exhibits periodic behaviour, and otherwise the gene expression profile appears white noises. According to degrees of freedom (i.e., the length of time-course data m) and a significance level (typically, 0.01, 0.05, 0.1, 0.2, or the like) specified by a user, the threshold value can be determined from F-distribution table or by using a standard MatLab function `icdf('f', 1- α , 2, m-2)`, where α is the significance level. If a significance level associated with a gene is smaller than the preset significant level, the genes are judged to be pseudo-periodic, and otherwise it is not.

III. EXPERIMENTAL RESULTS AND DISCUSSION

This study employs the following three biological datasets to investigate the performance of the proposed method.

Eluration-synchronized gene expression data of the yeast (ELU): Spellman et al. [1] studied the mitotic cell division cycle of yeast and monitored more than 6000 genes of yeast (*Saccharomyces cerevisiae*) at 14 equally-spacing time points in the eluration-synchronized experiment. Genes with missing data were excluded in this study. The resultant dataset contains the expression profiles of 5766 genes.

Alpha-synchronized gene expression data of the yeast (ALPHA): Spellman et al. [1] studied the mitotic cell division cycle of yeast and monitored more than 6000 genes of yeast (*Saccharomyces cerevisiae*) at 18 equally-spacing time points in the alpha-synchronized experiment. Genes with missing data were excluded in this study. The resultant dataset contains 4489 expression profile of 4489 genes.

Bacterial cell cycle (BAC): This dataset contains gene expression measurements during the bacterial cell cycle division process for about 3000 predicted open reading frames, representing about 90% of all bacterium *Caulobacter crescentus* genes [2]. The measurements were taken at 11 equally-space time points over 150 minutes. Genes with missing data were excluded in this study. The resultant dataset contains the expression profile of 1593 genes.

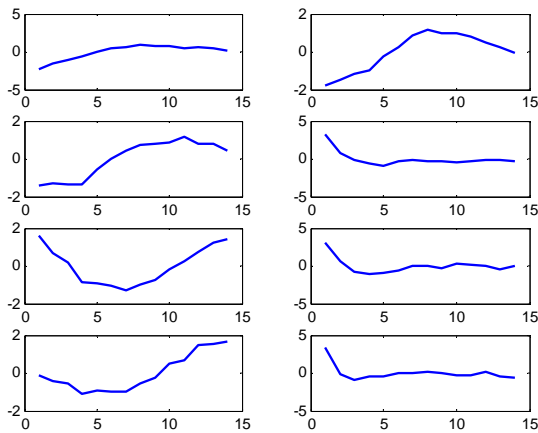


Figure 1. 8 gene profiles identified to be pseudo-periodically expressed in ELU dataset

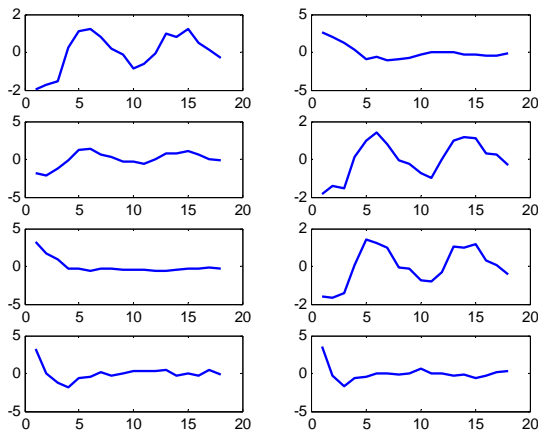


Figure 2. 8 gene profiles identified to be pseudo-periodically expressed in ALPHA dataset

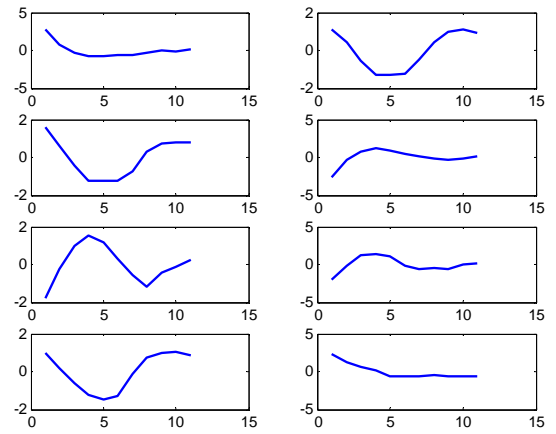


Figure 3. 8 gene profiles identified to be pseudo-periodically expressed in BAC dataset

The proposed method is applied to these three datasets. Figures 1-3 show the 8 gene profiles identified to be pseudo-periodically expressed from these datasets, respectively. From these figures, we can see these gene expression profiles appear pseudo-periodic. Most of gene expression profiles look more periodic, in whose models the values of the decrease (increase) rate α is small. Others look less, in whose models the values of the decrease (increase) rate α is dominant.

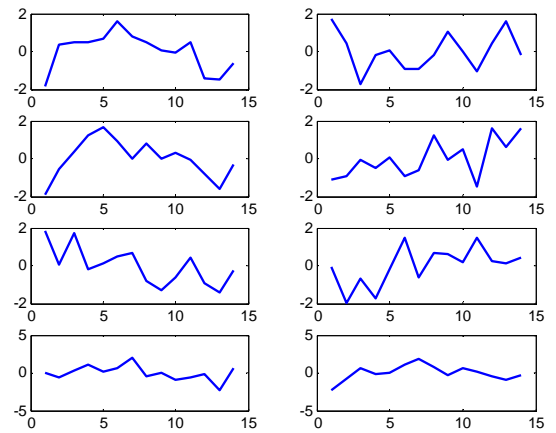


Figure 4. 8 gene profiles identified not to be periodically expressed in ELU dataset

Figures 4-5 shows show the 8 gene profiles identified to be non-pseudo-periodically expressed from ELU and ALPHA datasets (Figure for BAC is omitted because of space limitation), respectively. These gene expression profiles really look random noises.

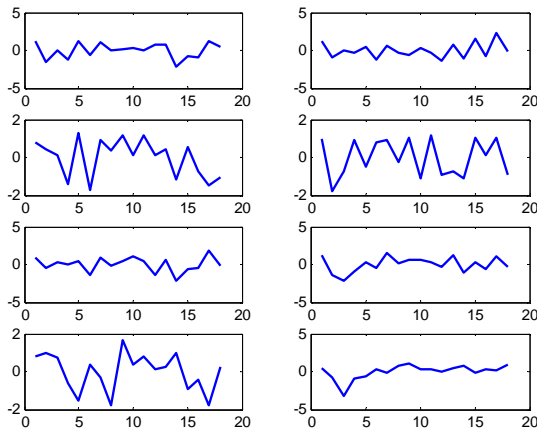


Figure 5. 8 gene profiles identified not to be periodically expressed in ALPHA dataset

IV. CONCLUSION AND FUTURE WORK

The linear combination of trigonometric and exponential functions has proposed to model pseudo-periodic gene expression profiles. A two step linear least squares method is proposed to estimate all model parameters. In addition, the proposed method uses F-test to determine if a gene expression profile appears pseudo-periodic or not. Computational experiments on three biological datasets have showed that the proposed method can effectively identify periodically expressed genes from their time-course expression profiles.

In this paper, the performance of the propose method is evaluated by manually checking some of results, for example, showing the profiles identified to be pseudo-periodic or those identified not to be pseudo-periodic. In the future, more objective criteria should be used to evaluate from both bioinformatic and biological view of points. In addition, this paper does not evaluate the proposed method on gene expression profiles. Another direction of feature work is to perform cluster analysis of gene expression data based proposed models.

Acknowledgment: This study was supported by Base Fund of Beijing Wuzi University and Fund for Beijing Excellent Team for Teaching Mathematics through the first author and by Natural Science and Engineering Research Council of Canada (NSERC) through the second and third authors.

REFERENCES

[1] Spellman, PT, et al.: Comprehensive identification of cell cycle-regulated genes of the Yeast *Saccharomyces cerevisiae* by microarray

hybridization. *Molecular Biology of the Cell* 9 (1998) 3273-3297

- [2] Laub, MT, et al: Global analysis of the genetic network controlling a bacteria cell cycle. *Science* 290(2000) 2144-2148
- [3] Langmead CJ, Yan A K, McCung C R, and Donald B. R.: Phase-independent Rhythmic analysis of genome-wide expression patterns, *Proceedings of the 6th Annual International Conference on Computational Biology* (2002) 1-11
- [4] Harmer S, et al.: Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* 290 (2000) 2110-2113
- [5] Wichert S, Fokianos K and Strimmer K: Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* 20 (2004) 5-20
- [6] Chen J: Identification of significant period genes in microarray gene expression data. *BMC bioinformatics* 6 (2005) 286
- [7] Glynn EF, Chen J, and Mushegian AR: Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle Periodograms. *Bioinformatics* 22 (2006) 310-316
- [8] Chen J and Chang KC: Discovering Statistically significant period Gene expression. *International Statistical Review* 76 (2008) 228-246
- [9] Liew AWC, et al.: Statistical power of fisher test for the detection of short periodic gene expression profiles. *Pattern Recognition* 42 (2009) 549-556
- [10] FX Wu (2010): Identification of Periodically Expressed Genes from Their Time-Course Expression Profiles, *ISBRA10*(short paper): 12-15
- [11] LP Tian, LZ Liu, FX Wu (2010): Parameter estimation method for Periodical Gene Identification, *iCBBE2011*, accepted
- [12] Duggan D.J., et al. (1999) Expression profiling using cDNA microarrays. *Natural Genetics* 21(Sup1): 10-14.
- [13] Eisen M.B. and Brown, P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol* 303: 179-205.

Gene Selection using Multidimensional False Discovery Rate

A. Moussa¹, M. Maouene¹, and B. Vannier²

¹LTI Laboratory, National School of Applied Sciences Abdelmalek Essaadi University Tangier, Morocco

²IPBC, University of Poitiers, Poitiers, France

Contact Author: Ahmed Moussa ; amoussa@ensat.ac.ma

Abstract - This paper proposes our algorithm for gene selection in microarray data analysis comparing conditions with replicates. Based on background noise computation in replicate array, this algorithm uses the global False Discovery Rate based on 'Between' group and 'Within' group comparisons of replicates to select the set of differential expressed genes. This method uses two types of statistics that lead to improve the selection procedure when confronted to very high background noise. Using simulated datasets and the well known Latin square data, the behavior of the proposed method is compared to results of some algorithms.

Keywords: Gene Selection; Replicates; False Discovery Rate; Local and global FDR.

1 Introduction

The most basic question one can ask in a transcriptional profiling experiment is which genes' expression levels changed significantly [1]. Answering this question involves many considerations. There may be two experimental conditions or many, the conditions may be independent or related to each other in some way, or there may be many different combinations of experimental variables. In each of these situations, the main goal is to identify genes expressed above background levels (absolute analysis), and/or that are differentially expressed (DE) between conditions of interest. In this work we are interested to genes that are DE between replicated conditions.

A standard statistical test to detect significant changes between repeated measurements of a variable in two groups is the t-test; It can be generalized to multiple groups via the ANOVA F-statistic [2]. Variations on the t-test statistic for microarray analysis are abundant [3, 4, and 5].

For microarray studies focusing on finding sets of predictive genes, a simple method proposed by [6] computes the probability that a given gene identified as differentially expressed is a false positive by means of 'false discovery rate' (FDR). A permutation-based approximation of this method, assuming that each gene is an independent test, is implemented in the Significant Analysis of Microarray (SAM) program [3].

The variation present in microarray data poses the challenge of determining whether differences between expression measurements are caused by biological difference, or by technical variations. The best way to address this question is to use replicates for each condition studied. There are two primary types of replicates: technical and biological. Technical replicates involve taking one sample from the same source tube and analyzing it across multiple conditions (multiple microarrays). Biological replicates are different samples measured across multiple conditions (multiple samples). The use of replicates offers three major advantages:

- Replicates can be used to measure variations in the experiment so that statistical tests can be applied to evaluate differences. This property will be more explored in this paper.
- Averaging across replicates increases the precision of gene expression measurements and allows the detection of smaller changes to be detected. As the number of replicates increases, both the detectable difference from background and the detectable fold change decrease [7].
- Replicates can be compared to detect outlier results (that may occur) due to aberrations within the arrays, the samples, or the experimental procedures. The presence of outlier sample can have a severe impact on the interpretation of data. Most array platforms have internal controls to detect various problems in an experiment. However, internal controls can not identify all issues.

Multiple studies have shown that fold change on its own is an unreliable indicator [7]. If multiple measurements (i.e. replicates) exist for each gene within each condition, the measurement of variations can be estimated [8].

2 Local and Global FDR

Noting V the random variable representing the number of false discoveries and R the number of significant results obtained from a particular multiple testing procedure, [6] defined the FDR by :

$$FDR = E(V / R) \text{ if } R > 0, \text{ and } 0 \text{ otherwise} \quad (1)$$

The positive FDR (pFDR) defined by [9] (for $R > 0$), is:

$$pFDR = Pr(H=0/T \in \Gamma) = \frac{\pi_0 P_r(T \in \Gamma / H=0)}{P_r(T \in \Gamma)} \quad (2)$$

where H is the variable such as H = 0 if the null hypothesis H_0 is true, H = 1 if the alternative hypothesis H_1 is true, $\pi_0 = Pr(H = 0)$ is the probability of not being modified and T is the test statistic used for all tested hypotheses. pFDR and FDR are asymptotically equivalent and, in the following, we will note FDR for both of them.

Data provided from microarray in gene expression analysis can be considered as composed of two subpopulations of genes, those for which the null hypothesis is true (unmodified genes or non DE genes), and those for which the alternative hypothesis is true (modified genes or DE genes). Let $p_i, i = 1, \dots, m$ be the *P-values* calculated for the m tested hypotheses. Let *P* be the random variable for which the *P-values* are the observations and let *f* be the marginal probability density function (pdf) of *P*. Denote f_0 the conditional pdf of *P* under the null hypothesis and f_1 the conditional pdf of *P* under the alternative hypothesis. Then:

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p) \quad (3)$$

In this setting, the local false discovery rate is:

$$fdr(p) = \pi_0 \frac{f_0(p)}{f(p)} \quad (4)$$

The local fdr can be interpreted as the expected proportion of false positives if genes with observed statistic are declared DE. Alternatively, it can be seen as the posterior probability of a gene being non-DE.

The main problem is the π_0 estimation. One solution assumes that the marginal distribution of the P-values arises from a beta-uniform mixture distribution. The model parameters are estimated using the maximum-likelihood method [10]. However, the widely estimator for π_0 is the one proposed by [11]. Using a tuning parameter $\lambda \in [0,1]$, π_0 is estimated by:

$$\pi_0^e = \frac{\#\{p_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)} \quad (5)$$

In [12], the local fdr is generalized to multidimensional fdr for more one P statistic. For example in the two dimensional case, we can use two different statistics P_1 and P_2 that capture different aspect of the information contained in the data. The obtained fdr-2D can be expressed as:

$$fdr2D(p_1, p_2) = \pi_0 \frac{f_0(p_1, p_2)}{f(p_1, p_2)} \quad (6)$$

An already established graphical display for studying the trade-off between effect size and significance is the volcano plot of \log_{10} -P-values versus fold changes [13], corresponding to:

$$p_{1i} = \text{mean}(x_{i1}) - \text{mean}(x_{i2}) \text{ and } p_{2i} = -\log_{10} P\text{-value}_i \quad (7)$$

where $\text{mean}(x_{i1})$ is gene-wise group mean.

In multidimensional case, the global FDR is the average of the local fdr for all used statistics. This FDR is a useful relationship for characterizing a collection of genes declared DE by local methods. Suppose R is a rejection region such that all genes with multidimensional statistics $p \in R$ are called DE. The global FDR associated with genes in R is [12]:

$$FDR(R) = E(fdr(p)/R) \quad (8)$$

This means that the global FDR of gene lists found by fdr2D can be computed by simple averaging of the reported local fdr values, and consequently, fdr2D can be compared easily with other procedures in terms of its implied global FDR.

Please use the styles contained in this document for: Title, Abstract, Keywords, Heading 1, Heading 2, Body Text, Equations, References, Figures, and Captions. Do not add any page numbers and do not use footers and headers (it is ok to have footnotes).

3 Method Description

3.1 Between and Within Group Comparisons

Consider the example where we have to compare two experiments (Traited # Control) with three replicates. For the available microarrays, we can process in term of statistics, to two types of comparisons: 'Between' group comparisons that concern chips providing from the two samples "Fig.1". And 'Within' group comparison that concern chips inside biological or technical replicates "Fig.3".

For each set of comparison, a multidimensional fdr2D, based on statistics of equation 7 may be computed. These statistics can be summarized in two volcano plots where the first one represents results of 'Between' group comparison "Fig.3": in this plot the significance correspond to the average ($-\log_{10}$ P-value) across all the 'Between' groups comparison and the average Signal Log-Ratio (SLR) obtained from average fold change across all the 'Between' group comparison. And the second one show the same statistics related to the 'Within' group comparison "Fig.4". This latter informs about the experiments background noise [14]. In fact, gene stimulated in 'within' group comparisons inform about amplitude and act of experimental background noise. When this noise is very low, all genes SLR are falling around 0 in this plot.

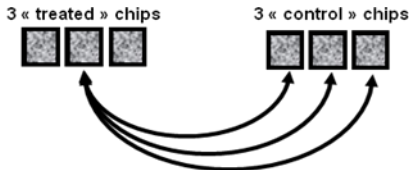


Figure 1: 'Between' group comparisons

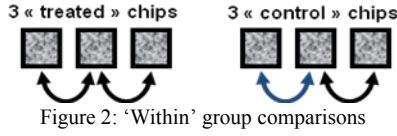


Figure 2: 'Within' group comparisons

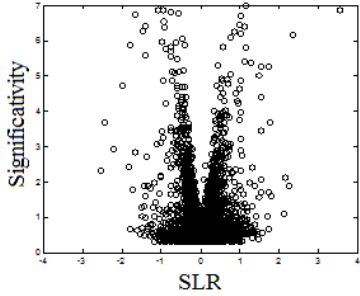


Figure 3: volcano plot of 'between' group comparison

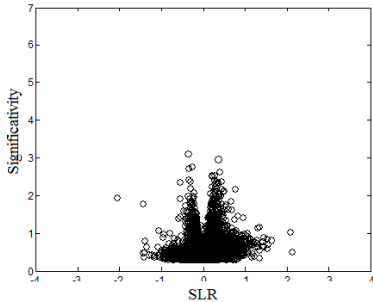


Figure 4: volcano plot of 'within' group comparison

3.2 Local fdr and Replicates

To illustrate our procedure, we use first the local fdr as described in section I. For the two sets of comparison we use the same statistics and the same null hypothesis $H = 0$. In this context the local fdr for 'Between' group comparison and 'within' group comparison are :

$$fdr^b(p) = \pi_0^b \frac{f_0^b(p)}{f(p)} \text{ and } fdr^w(p) = \pi_0^w \frac{f_0^w(p)}{f(p)}$$

Without loss of generality the expression:

$$FDR = \frac{fdr^w(p)}{fdr^b(p)} = \frac{\pi_0^w f_0^w(p)}{\pi_0^b f_0^b(p)} \quad (9)$$

interpreted as the expected proportion of false positives if genes with observed statistic are declared DE, is the common local FDR with the same null hypothesis[15].

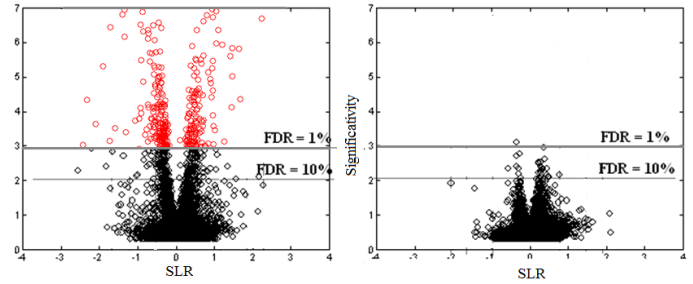


Figure 5 : volcano plot for 'between' and 'within' group comparison with the same null hypothesis.

The FDR of equation 9 changes from 0 to 1 according to the cutoff fixed by the analyst. Each FDR-cutoff value correspond to one value of significativity ($-\log_{10}(\text{P-value-cutoff})$). But in certain case, especially when the 'Within' group comparison presents a high degree of noise, this curve may not be straight monotonous and two FDR-cutoff values can corresponds to the same significativity "Fig.6". This not advisable behavior is corrected by a curve smoothing (FDR versus Significativity) with a monotonic quadratic function, where the smoothed curve guarantees the FDR uniqueness versus significativity correspondence "Fig.6".

The proposed method works well when the noise observed in 'within groups' comparison is moderate. But when the background noise is high, the FDR is not well informative, and it is very difficult to find the appropriate function to extrapolate the curve FDR versus Significance. Thus, to improve the method we used two statistics ($-\log_{10} \text{Pvalue}$ and SLR) to generalize this concept to the global FDR.

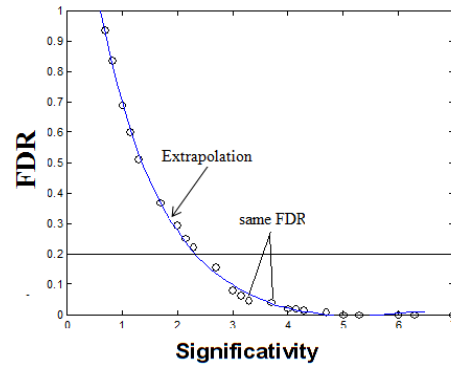


Figure 6 : smoothing the FDR vs Significativity plot

3.3 Global fdr-2D and Replicates

This solution introduces the SLR information in the selection method [16].As explained in the last section we use the local FDR for both the significance and fold change statistics. The use of two different statistics that test the same null hypothesis, but have different power against t-statistics and fold changes, comparable with the proposal made by [16], is another possibility. Thus, this method takes into account the information provided by both signals and replicates and gives a best estimate of background noise in microarray.

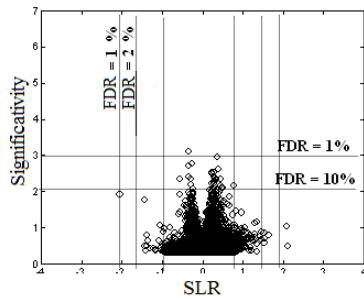


Figure 7 : FDR corresponding to the null hypothesis applied to SLR

In the selection step, the method uses conjointly FDR for significance and FDR for SLR. This Global FDR, which uses replicates as a background adjustment is called in the next “global FDR-2D”, and is expressed exactly by the equation 8.

The gene selection procedure proposed here run as follows:

- 1- Establish a curve, as in “Fig.6” for the studied example using a global FDR-2D values set.
- 2- Curve “FDR-2D versus significance” smoothing
- 3- Assignment of the cutoff value and search a corresponding FDR-2D in the curve (FDR-2D) cutoff

Selection of DE Genes with $FDR-2D < FDR-2D \text{ cutoff}$

4 Results and Discussions

4.1 Simulated Dataset

We assume 10 000 genes per array with a proportion of truly non-DE genes $\pi_0 = 0.95$ throughout, and compare two independent groups with $n=4$ arrays per group. We further assume that the log expression values are also normally distributed in each group.

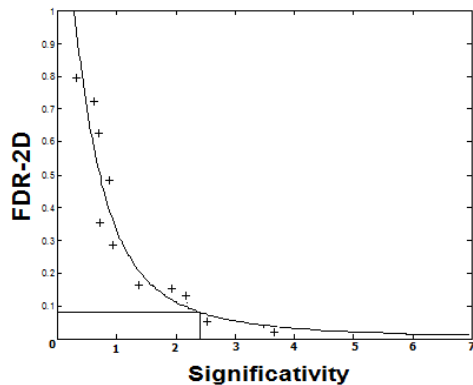


Figure 8 : smoothing the FDR vs Significance for the simulated dataset

We have compared results of this gene selection method to :Significance Analysis of Microarray (SAM)[3], Controlling the fdr (Benjamini method) [6], and Multidimensional local fdr [12]

In the comparison, we use three values of $FDR-2D^{\text{cutoff}}$ e.g. 1%, 5% and 8% “Table I”.

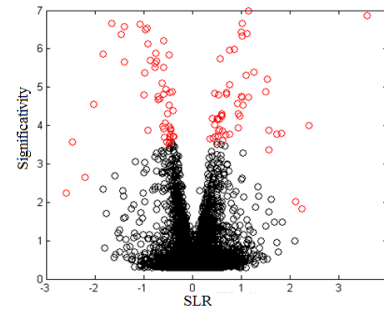


Figure 9 : Gene selected by the global $FDR-2D^{\text{cutoff}}=5\%$

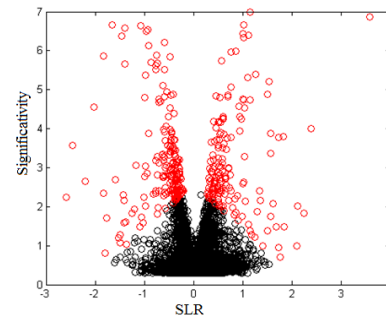


Figure 10 : $FDR-2D^{\text{cutoff}}=8\%$

TABLE I. RESULT OF SIMULATED DATASET

FDR Value	TDR			Percentage of spike detected		
	1%	5%	8%	1%	5%	8%
Method 1	58.20	52.30	44.50	72.36	76.45	66.33
Method 2	68.56	45.50	35.56	66.15	70.83	67.98
Method 3	96.17	95.26	97.11	88.32	80.64	73.37
Proposed Method	95.23	93.45	91.48	89.21	81.70	75.77

4.2 Real Dataset

The proposed method was used to analyze spiked-in genes arrayed in a Latin square. In this publicly available set, 112 yeast genes and 14 human genes are cloned. Each of the labeled genes were pooled into groups and diluted to concentrations of 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 pM. In every microarray experiment, 14 groups of genes in 14 different concentrations were hybridized to the microarray in the presence of a complex background of expressed human genome (30 Mb) and several control genes. For this Latin square design, 14 groups of experiments with 3 replicates for each experiment, giving a total of 42 experiments. The concentrations of the 14 in vitro transcript (IVT) groups in the first experiments are 0, 0.25, 0.5, . . . , 1024 pM, their concentrations in the second experiments are 0.25, 0.5, . . . , 1024, 0 pM, and so on [17].

The selection method proposed in this work has been applied to the Latin Square dataset. The main objective is to select a set of genes according to pre-defined P-value and compare the result with the 42 spiked-in genes. Result

summarized in “Table I” compare the results of this new selection gene method to those used in the last section for evaluating the performance of this algorithm thought simulated dataset.

TABLE II. RESULT OF REAL DATASET

FDR Value	TDR			Percentage of spike detected		
	1%	5%	8%	1%	5%	8%
Method 1	50.39	45.36	29.87	58.26	66.45	67.35
Method 2	65.44	66.21	69.52	67.6	68.84	75.38
Method 3	58.59	60.49	68.11	74.32	78.26	80.36
Proposed Method	60.58	62.47	67.21	75.65	80.25	85.46

Table 1 regroup results of four gene selection methods applied on statistical parameter of simulated dataset. The best percentage of spike detected was found by the global fdr-2D algorithm. Method3 and FDR-2D have the best percentage of spike detected. These results confirm the good behavior of the two methods in the case of simulated data. This conclusion is confirmed where the proposed algorithm have been confronted to complex data like Latin Square. In fact, in table II, the proposed method and the method 3 gives a good result of detected spike.

All of these results confirm on the one hand the good behavior of the proposed algorithm in the gene selection problem. on the other hand, it proof that when taking into account replicates of arrays by mean of the ‘within’ group comparison, the method allows good detection of modulations for weakly expressed genes and eliminates false positives.

References

- [1] W. Liu, R. Mei, X. Di, T. Ryder, E. Hubbel, S. Dee, T. Webster, C. Harrington, M. Ho, J. Baid, S. Smeekens, “Analysis of high density expression microarray with signed-rank calls algorithms”, *Bioinformatics*, vol. 18, N°. 12, 2002.
- [2] J.-H. Zar, “Biostatistical Analysis” Prentice-Hall , Upper Saddle River, NJ, 663, 1999.
- [3] V. G. Tusher, R. Tibshirani, G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response”, *Proc. Nat. Acad. Sci. USA* 98, pp. 5116–5121, 2001
- [4] T. R. Golub, et al, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”, *Science* vol. 286, pp.531–537, 1999.
- [5] F. Model, P. Adorjan, A. Olek, C. Piepenbrock, “Feature selection for DNA methylation based cancer classification” *Bioinformatics* vol. 17, N°. 1, pp. 157–164, 2001.
- [6] Y. Benjamini, Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing”, *Journal of the Roy. Stat. Soc.* vol. B 57, pp.289–300,1995.
- [7] M. Newton, C. Kendzioriski, C. Richmond, F. Blattner, K. Tsui, “On differential variability of expression ratios: improving statistical inference about gene expression changes from microarraydata”, *Journal of Comparative, Biol.*, vol. 8, pp.37-52, 2001.
- [8] W. Pan, J. Lin, C. Le, “How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach”, *Genome Biolo.* Vol. 3, N°.5, 2002.
- [9] J. -D. Storey, “A direct approach to false discovery rates”. *Journal of the Roy. Stat. Soc. Serie B*, vol. 64,pp.479-498, 2001.
- [10] S. Pounds, W. Morris, “Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values”, *Bioinformatics*, vol.19, pp.1236-1242, 2003.
- [11] J.-D. Storey,R. Tibshirani, “Statistical significance for genome-wide studies”, *Proc. Natl Acad. Sci. USA*, 100, pp. 9440–9445, 2003.
- [12] A. Ploner, S. Calza, A. Gusnanto, Y. Pawitain “Multidimensional local false discovery rate for microarray studies” , *Bioinformatics* vol. 22 N°5, pp.556–565, 2006.
- [13] R.-D.Wolfinger, et al. “Assessing gene significance from cDNA microarray expression data via mixed models”, *Journal of Comput. Biol.*, vol. 8, pp.625–637, 2001.
- [14] G.-A. Churchill, “Fundamental of experimental design fo cDNA microarray”, *Nature genetics supplement* 32, pp.490- 490, 2002.
- [15] A. Moussa, M. Maouene, B. Vanier, “Multidimensional Method For Gene Selection” *Proceeding of the 5th World Congres on Celular and Molecular Biology*, Indor, India, 6-9 November 2009
- [16] Y. H. Yang, et al., “Identifying differentially expressed genes from microarray experiments via statistic synthesis”, *Bioinformatics*, vol. 21, pp. 1084–1093, 2005.
- [17] W. Liu, and all. , “Analysis of high density expression microarray with signed-rank calls algorithms” *Bioinformatics*, vol. 18, N°.12, pp.1593-1599, 2002.

Knowledgments

This work was supported in part by the Moroccan National Center of Scientific and Technical Research of Grant ScVie 03/10.

Letting all have a say: A novel method for microRNA RT-PCR normalization

R. Qureshi¹, A. Sacan¹

¹Center for Integrated Bioinformatics, School of Biomedical Engineering, Science and Health Systems, Drexel University, 3120 Market Street, Philadelphia, PA 19104, USA

Abstract - *MicroRNAs (miRNAs) are short non-coding RNA molecules. MicroRNAs regulate mRNA transcript levels and translation. miRNA expression is measured by microarray or real-time polymerase chain reaction (RT-PCR). The findings of RT-PCR data are limited by the normalization techniques. Some commonly used endogenous controls are differentially expressed in cancer, making them inappropriate internal controls.*

We show that RT-PCR data contains a systematic bias resulting in large variations in the Cycle Threshold (CT) values of the low-abundant miRNA samples. This observation is illustrated on a microRNA dataset obtained from primary cutaneous melanocytic neoplasms. We propose a new data normalization method that considers all available microRNAs as endogenous controls. A weighted normalization approach is utilized to allow contribution from all microRNAs, weighted by their empirical stability. We show that through a single control parameter, this method is able to emulate other commonly used normalization methods and thus provides a more general approach.

Keywords: microRNA, PCR, normalization

1 Introduction

MicroRNAs (miRNAs) are short non-coding RNA sequences that average 22 nucleotides in length [1-3]. These class of RNAs are distinct from other short sequence RNA types such as siRNA and snRNA, The first RNA of this class was identified in *C. Elegans* in 1993 [4]. However, miRNAs were not recognized as a special class of RNAs until a decade ago [5]. To date, all animal and plant species have been found to express miRNAs [6]. At this time approximately 1000 miRNA sequences have been identified in the human microribonucleome [7]. miRNA sequences are highly evolutionarily conserved among mammals [4,8-12]. Approximately 80% of miRNA genes occur in intronic regions of the genome [13-14]. miRNAs are involved in many biological processes by influencing the regulation of their target genes, generally resulting in down-regulation. There are two postulated methods by which miRNAs act on their target genes. If the miRNA binds with an mRNA transcript and they exhibit high complementarity, it will cause the degradation of the mRNA. If the miRNA binds with incomplete complementarity then it causes translational repression of the mRNA. In plants the primary mechanism of action of miRNAs mRNA transcript degradation, while in

animals, translational repression is more common [6]. An estimated 60% of mammalian mRNAs are targeted by one or more miRNAs [10, 12].

miRNAs have been discovered to play a role in many diseases and pathologies [2,10,13,15-16]. The role of miRNAs in cancer has been examined and several miRNAs have been found to regulate tumor-related genes [1-3,10,13,17-19]. In fact, more than half of all miRNA genes are located in cancer-associated regions of the genome or in fragile sites [3,13]. As a result, therapeutic applications of miRNAs are being investigated. Furthermore, due to the link between many miRNAs and cancer, these RNA molecules are being investigated as potential cancer biomarkers. The fact that some miRNAs can be found extracellularly and maintain their stability in the extracellular environment facilitates their usage as biomarkers [10].

There are two main tools used to quantify the expression of miRNAs: microarrays and real-time polymerase chain reaction (RT-PCR). RT-PCR returns the number of cycles that the samples underwent before they were detected, reported as a value known as the Cycle Threshold (CT). The CT values vary logarithmically with expression levels. There are several methods of normalizing the data and calculating the fold-change of each gene between samples. For convenience, in this presentation miRNA and gene are used interchangeably in the context of RT-PCR. ΔCT values are calculated by subtracting the CT value of the endogenous control for a given sample (or the mean of the CT values of the endogenous controls if more than one exist) from the CT value of the gene for the given sample. In the calculation of ΔCT values we refer to the number subtracted from the raw CT values of each gene as the CT_0 . The $\Delta\Delta CT$ is calculated by subtracting the ΔCT of an experimental sample from a control sample. Fold change is calculated by raising 2 to the power of the negative $\Delta\Delta CT$ value, since CT values are related to the amount of miRNA or gene logarithmically [20]. The relationship between CT, ΔCT , $\Delta\Delta CT$, and Fold Change (FC) are given by the equations below.

$$\Delta CT = CT - CT_0 \quad (1)$$

$$\Delta\Delta CT = \Delta CT - \Delta CT_{control} \quad (2)$$

$$FC = 2^{-\Delta\Delta CT} \quad (3)$$

Theoretically, endogenous controls are selected because they have low variance in their expression levels across samples. In the case of miRNAs, the endogenous controls are typically recommended by the manufacturer of the miRNA kit used in the PCR. Some of the most commonly used endogenous controls are RNU44, RNU48, and U6 [17]. However, the usage of these endogenous controls is problematic, because even though these endogenous controls have stable expression levels in normal tissue samples, they have been found to be differentially expressed in cancerous tissue compared with normal tissue [17].

Directly applying this method can lead to misleading results if the CT values in the data are not normalized. There are several commonly used methods for miRNA normalization, including: quantile normalization, median normalization, and cyclic loess. Quantile normalization involves sorting the expression values of each gene in a given sample in order from least to greatest. This is done for each sample in the study. The vectors of the sorted CT values for each sample are combined into a matrix. The mean of each row of the matrix is calculated. The CT value in each element in each row is replaced with the mean of the entire row. In the case of median quantile normalization the median of the row is used instead of the mean. The CT values in each sample are then rearranged back into their original order. This causes the distribution of CT values across all samples to assume the same shape, which will minimize the variance except for that resulting from the experimental condition beings studied [21-22].

Median normalization shifts the CT values in each sample such that the median CT value of each sample is the same. The median of each plate should be determined, and the medians of all plates should be arranged in a vector and sorted to determine the median of the medians. In each plate the difference between the median of the sample and the overall median should be subtracted from the CT value of each gene [9].

In cyclic loess normalization, pairs of plates are considered. For all pairs of plates the difference of the log of the CT for each gene is represented by M, and the average of each gene of the log of the expression values is represented by A. Then a loess curve is fit by regression of M on A which results in a fitting vector F. The genes in the first sample are adjusted by adding half the F value corresponding to the log of the CT for each gene. In the second sample half the F value is subtracted from the log CT of the gene [9, 21].

One of the main problems with RT-PCR that remains as yet unaddressed by current normalization methods is the systematic bias present within the data. We observe that standard deviation increases as CT values increase. We believe that the most likely cause of this observation is the assumption that the PCR magnification at each cycle is an exact doubling of the expression levels is inaccurate. There seems to be an accumulation of expression-level specific rate-

limiting effect. As a result, a small difference in the size of the initial sample being amplified causes larger variations in the CT values of the less abundant microRNA molecules. Consequently, using endogenous controls, which are usually chosen from highly expressed microRNAs, for normalization becomes inappropriate for the less-abundant microRNAs. One potential solution is to use the mean expression values of all genes in a sample as the endogenous control and calculate ΔCT by subtracting this mean CT value from the CT value of all genes on the plate. However, this approach is not ideal because the mean of the entire plate is sensitive to fluctuating genes as well as undetected genes which have high CT values. As a result, the mean-value normalization method is dominated by the large fluctuations of the less-abundant microRNAs and may cause spurious differential expression levels for otherwise stable microRNAs. In this study, we propose a method of using a weighted mean as an artificial endogenous control to calculate ΔCT values. The standard deviation of a microRNA across all samples is considered as a stability measure and each microRNA is weighted by its stability to generate the artificial endogenous control levels.

2 Methods

The dataset used in this study was obtained from a recently deposited microRNA RT-PCR dataset in the Gene Expression Omnibus (GEO) [23]. The data was from a study by Jukic et al. that examined the difference in miRNA expression profiles in melanocytic neoplasms between young and older adults [1]. Their study examined 10 young adults and 10 older adults and measured the expression of 666 microRNAs. We used the raw CT values measured in their data to compare different approaches to normalizing the data.

We have investigated several normalization methods, including quantile, mean, and median normalization methods, and endogenous controls identified using various stability criteria. In mean and median normalization, the mean and median of all of the genes in a given sample are used as the value for CT_0 . For identification of endogenous controls, we calculate the standard deviation of each microRNA across all samples, and rank them in the order of increasing standard deviation. The CT values of the top-k microRNAs are averaged in each sample to provide the CT_0 values.

A new weighted mean metric is proposed using the standard deviations of the microRNAs as weights. For a given gene, the weighted average is calculated using the following equation:

$$CT_0 = \sum (CT \times \frac{(\frac{1}{STD(CT)})^{wmp}}{\sum_{i=1}^n 1/STD(CT_i)}) \quad (4)$$

where wmp is the weighted mean power, which can be adjusted to shift the dominance between stable and unstable microRNAs, n is the number of genes or microRNAs, and STD is the standard deviation. The weighted mean

calculation involves raising the inverse of the standard deviation of a given gene across all samples to the weighted mean power, which is usually specified as 1, and dividing by the sum of the inverses of the standard deviations for all genes. CT_0 is calculated for each sample by taking the sum of the product of all the raw CT values in the sample and the previous number. When the ΔCT is calculated the CT of each gene is subtracted by the above value. This method gives a higher weight to genes with a lower standard deviation.

3 Experiments and Results

In order to test the hypothesis that increasing CT values magnifies the natural variation between the initial amounts of samples loaded in each well during RT-PCR, we examined the standard deviation of the genes against their mean CT values (Fig. 1). A linear regression fitted to this data clearly shows a trend of increasing standard deviation values for higher CT values. Note that the higher the CT value, the more cycles were required to observe that microRNA, hence the less abundant that microRNA was in the initial loaded sample.

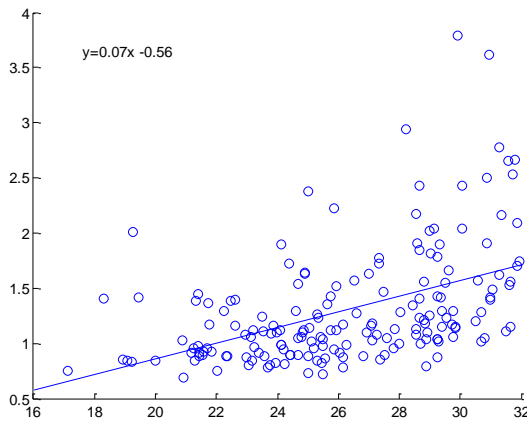


Fig. 1: A plot of the standard deviation vs. expression level fitted to a line.

As expected, the CT values of most genes are well correlated with the mean expression of all the genes. This is illustrated Fig. 2, where we show the expression of the 20 miRNAs that are most correlated with the mean expression. Each tick on the x-axis represents a unique experimental sample.

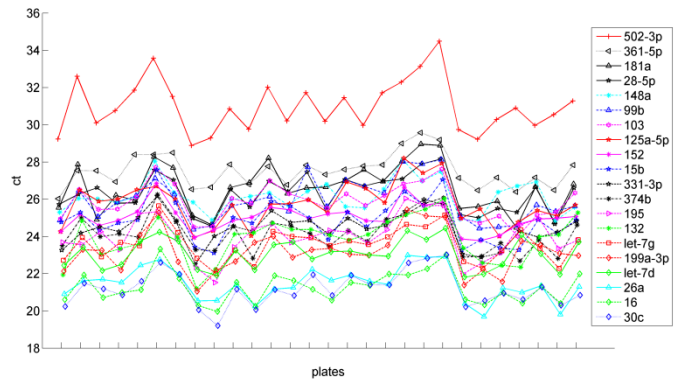


Fig. 2: The 20 miRNAs most correlated with fluctuations in the mean expression value.

The correlation with the mean expression level extends to low-abundant miRNAs. We demonstrate this in Fig. 3, wherein the Pearson correlation coefficient of the fluctuations in each gene with respect to its own average is shown against the fluctuations of the mean expression levels of all genes. The plot shows that a high correlation is observed whether the mean CT values are low or high.

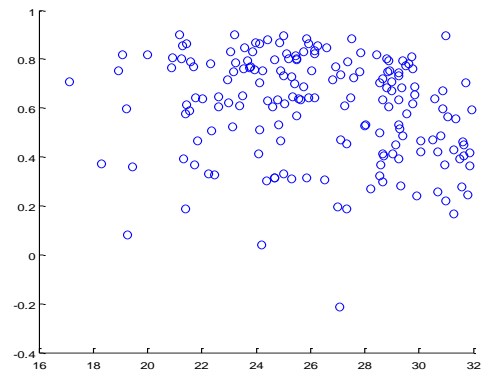


Fig. 3: A plot of the correlations of miRNAs with fluctuations in the mean miRNA CT value.

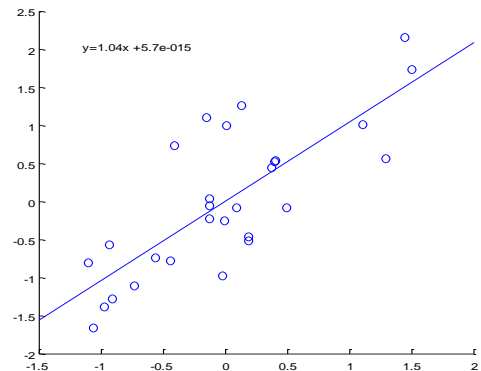


Fig. 4: An example of line fitting.

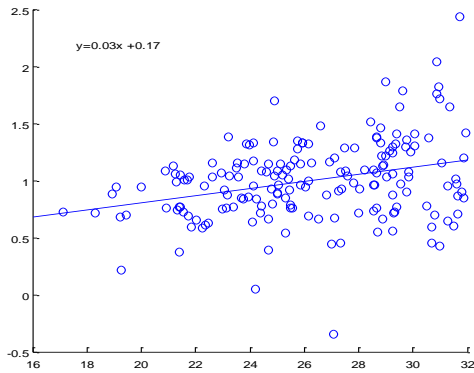


Fig. 5: A plot of the fluctuation ability versus the expression level.

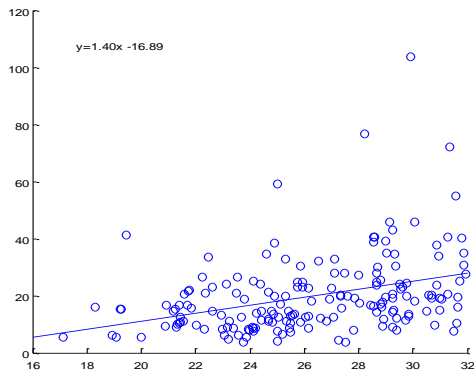


Fig. 6: A plot of the difference ratio versus the expression level.

In order to quantify the "response" of the microRNA levels to the initial loaded sample size, a regression line is fitted to the fluctuation of each gene against the fluctuation of mean expression. In Fig. 4 we demonstrated this for a single miRNA. The slope of the line indicates how sensitive the miRNA is to initial sample size, with larger slope values corresponding to larger variations in response to a small change in sample size. Fig 5 shows the response of each gene against the mean expression level of that gene. We observe that the response is expression level dependent. Highly expressed genes (those with small CT values) are less responsive to changes in the overall mean of the genes, whereas the low-abundant genes are more sensitive to the changes in the overall mean of the genes. Note that, this is not simply a random effect due to low abundant microRNAs being more variable, since the variation is still correlated and is in the same direction of the change in mean expression level. The same observation is made by examining the ratio of the fluctuations in individual genes and in the mean expression level (Fig. 6).

In conclusion, the fluctuations of the low-abundant miRNAs are not random. The changes in their expression levels are correlated well with the overall changes in all miRNAs, which is assumed to be due to different starting sample sizes for the PCR reactions. We see that there is a systematic bias in the CT values that causes the expression levels of the low-abundant miRNAs to be more sensitive to the initial sample sizes.

We then investigated the suitability of our weighted mean metric. In Fig. 7 we display the values for CT_0 for

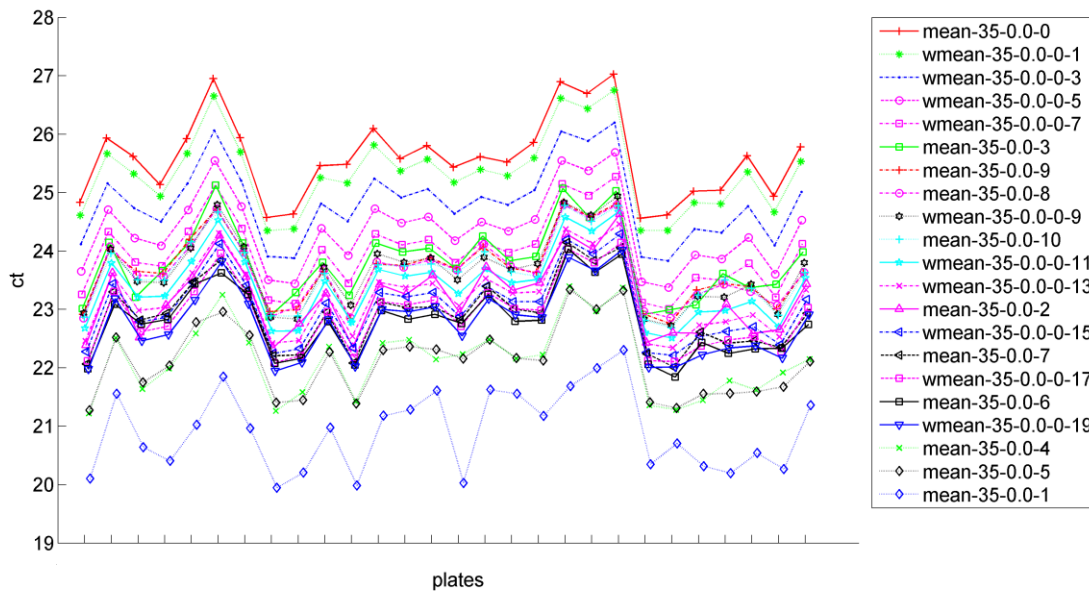


Fig. 7: A comparison of different methods of calculating CT_0 .

several different methods including using the mean of all raw CT values in the uppermost line (top-k = 0), the means of the top-k miRNAs for different values of k, and the weighted mean for different values for the weighted mean power. The plot demonstrates that varying the weighted mean power enables the shifting of the curve upwards or downwards. In Table 1 and Table 2, we compare the resulting means, standard deviations, and geNorm stability values [24] for mean and weighted mean normalizations, respectively. We repeat analysis this for the top 10 genes, with the lowest standard deviation in Table 3. We see slightly higher standard deviations in the weighted mean normalization method compared to the top-k calculations, but the weighted means' CT₀ are determined to be more stable by geNorm (the lower the value the more stable). In Table 3, we see that the best individual miRNAs have a much higher standard deviation and are much less stable than any of the CT₀ calculations using either the top-k miRNAs or the weighted mean. This indicates that it is better to use these values in the $\Delta\Delta\text{CT}$ calculation than any endogenous control.

Table 1: Mean normalization results.

mean normalization			
topk	AVG CT	STD CT	geNorm
0	25.59	0.71	0.23
1	20.92	0.69	0.35
2	23.2	0.64	0.21
3	23.8	0.64	0.19
4	22.13	0.63	0.2
5	22.11	0.61	0.17
6	22.79	0.6	0.18
7	22.91	0.61	0.16
8	23.66	0.59	0.15
9	23.67	0.6	0.16
10	23.61	0.61	0.16

Table 2: Weighted mean normalization results.

weighted mean normalization			
power	AVG CT	STD CT	geNorm
1	25.34	0.69	0.21
3	24.82	0.67	0.18
5	24.35	0.65	0.15
7	23.96	0.64	0.14
9	23.65	0.63	0.13
11	23.41	0.62	0.12
13	23.21	0.62	0.12
15	23.04	0.62	0.12
17	22.89	0.61	0.12
19	22.76	0.61	0.13

Table 3: Results for top 10 endogenous control candidates.

miRNA	AVG CT	STD CT	geNorm
191	20.92	0.69	1.14
744	25.49	0.72	1.17
152	25	0.73	1.12
MammU6	17.12	0.75	1.22
92a	22.03	0.75	1.24
29c	26.15	0.78	1.26
186	23.69	0.78	1.17
671-3p	28.89	0.8	1.29
26b	23.75	0.8	1.19
let-7d	23.07	0.8	1.16

4 Conclusion

We explored the phenomenon whereby differences in the initial sample size of miRNA in an RT-PCR experiment were magnified with increasing CT levels. This was illustrated by the strong correlation of the CT values of the individual miRNAs with the average CT values of all miRNAs and by the increased sensitivity in the CT values of the low-abundant miRNAs to the average CT values. We conclude that the systematic bias in RT-PCR exists in which the fluctuations in the CT are dependent on the expression levels of the particular miRNAs. We further proposed a method of addressing this bias by using the weighted mean instead of an endogenous control in the calculation of ΔCT . We demonstrated that the new normalization method produces lower standard deviations and is more stable than other methods.

Note that, while the power parameter in the weighted mean normalization method provides a convenient way of adjusting how much one wishes to let the less stable microRNAs influence the normalization of other microRNAs, its optimization currently requires enumeration of different values and using the one with the best overall stability. Other criteria, such as significance of the differentially expressed microRNAs can be utilized in this optimization. Furthermore, a separate custom CT₀ value for each microRNAs may be used, such that each microRNA is normalized differently, dependent on its average expression level.

While we have observed a similar bias in other miRNA datasets and have found the new normalization method to give superior results, a large scale comparison of different normalization methods on multiple data sources is currently under way. The utility of the new normalization method in better correlating with microarray quantification methods and in better identifying significantly differentially expressed genes will be demonstrated elsewhere.

5 References

- [1] D Jukic, L Kelly, J Skaf, L Drogowski, J Kirkwood, M Panelli. "MicroRNA profiling analysis of differences between the melanoma of young adults and older adults," *Journal of Translational Medicine*, vol. 8, pp. 27-27, -03-19, 2010.
- [2] S Schmeier, U Schaefer, C MacPherson, V Bajic, "dPORE-miRNA: Polymorphic Regulation of MicroRNA Genes," *PloS One*, vol. 6, pp. 835, 2011.
- [3] Y Han, J Chen, X Zhao, C Liang, Y Wang, L Sun, Z Jiang, Z Zhang, R Yang, J Chen, Z Li, A Tang, X Li, J Ye, Z Guan, Y Gui, Z Cai, "MicroRNA Expression Signatures of Bladder Cancer Revealed by Deep Sequencing," *PloS One*, vol. 6, pp. e18286, 2011.
- [4] R Lee, R Feinbaum, V Ambrose, "The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*," *Cell (Cambridge)*, vol. 75, pp. 843, 1993.
- [5] V. Ambrose, R Lee, "An Extensive Class of Small RNAs in *Caenorhabditis elegans*," *Science (New York, N.Y.)*, vol. 294, pp. 862-864, 2001.
- [6] D Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell (Cambridge)*, vol. 116, pp. 281-297, -01-23, 2004.
- [7] S Griffiths-Jones, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Res.*, vol. 34, pp. D140-D144, 2006.
- [8] B Wang, X Wang, P Howell, X Qian, K Huang, A Riker, J Ju, and Y Xi, "A personalized microRNA microarray normalization method using a logistic regression model," *Bioinformatics*, vol. 26, pp. 228-234, -01-15, 2010.
- [9] Y Rao, Y Lee, D Jarjoura, A Ruppert, C Liu, J Hsu, J Hagan, "A comparison of normalization techniques for microRNA microarray data," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, -01-01, 2008.
- [10] A Etheridge, I Lee, L Hood, D Galas, K Wang, "Extracellular microRNA: A new source of biomarkers," *Mutation Research. Fundamental and Molecular Mechanisms of Mutagenesis*, 2011.
- [11] V. Ambros, "The functions of animal microRNAs," *Nature (London)*, vol. 431, pp. 350-355, -09-16, 2004.
- [12] R Friedman, K Farh, C Burge, D Bartel, "Most mammalian mRNAs are conserved targets of microRNAs," *Genome Res.*, vol. 19, pp. 92-105, -01-01, 2009.
- [13] Y Liang, D Ridzon, L Wong, C Chen, "Characterization of microRNA expression profiles in normal human tissues," *BMC Genomics*, vol. 8, pp. 166-166, -06-12, 2007.
- [14] V Kim, Y Kim, "Processing of intronic microRNAs," *EMBO J.*, vol. 26, pp. 775-783, 2007.
- [15] P Mestdagh, P Vlierberghe, A Weer, D Muth, F Westermann, F Speleman, J Vandesompele, "A novel and universal method for microRNA RT-qPCR data normalization," *GenomeBiology.Com*, vol. 10, pp. R64-R64, -01-01, 2009.
- [16] G Latham, "MicroRNAs and the immune system normalization of MicroRNA quantitative RT-PCR data in reduced scale experimental designs," in *Methods in Molecular Biology (Clifton, N.J.)* Anonymous 2010, pp. 19-31.
- [17] H Gee, F Buffa, C Camps, A Ramachandran, R Leek, M Taylor, M Patil, H Sheldon, G Betts, J Homer, C West, J Ragoussis, A Harris, "The small-nucleolar RNAs commonly used for microRNA normalisation correlate with tumour pathology and prognosis," *Br. J. Cancer*, vol. 104, pp. 1168-1177, 2011.
- [18] C. M. Croce, "Causes and consequences of microRNA dysregulation in cancer," *Cell. Oncol.*, vol. 32, pp. 161-162, 2010.
- [19] X Wang, "A PCR-based platform for microRNA expression profiling studies," *RNA (Cambridge)*, vol. 15, pp. 716-723, -04-01, 2009.
- [20] K Livak, T Schmittgen. "Analysis of relative gene expression data using real-time quantitative PCR and the 2-DDCT method," *Methods*, vol. 25, pp. 402, 2001.
- [21] B. Bolstad, R Irizarry, M Astrand, T Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, pp. 185-193, -01-22, 2003.
- [22] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249-264, APR, 2003.
- [23] R Edgar, M Domrachev, A Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." *Nucleic Acids Res.* 30(1):207-10. 2002.
- [24] J Vandesompele, K De Preeter, F Pattyn, B Poppe, N Roy, A Paepe, F Speleman, "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes," *GenomeBiology.Com*, vol. 3, pp. research0034.1, 2002.

Classification of High-throughput Data Using Correlation-shared Gene Clusters

Pingzhao Hu, Hui Jiang

Department of Computer Science and Engineering, York University,
4700 Keele Street, Toronto, Ontario M3J 1P3, Canada

Abstract - *Molecular predictor is a new tool for disease diagnosis, which uses gene expression to classify the diagnostic category of a patient. The statistical challenge for constructing such a predictor is that there are thousands of genes to predict disease category, but only a small number of samples are available. We explored a correlation-sharing based method to integrate 'essential' correlation structure among genes into the predictor in order that the cluster structure of genes, which is related to diagnostic classes we look for, can have potential biological interpretation. We evaluated performance of the method with other methods using three real examples. Our results show that the approach has the advantage of computational simplicity and efficiency with lower classification error rates than the compared methods.*

Keywords: Correlation-sharing, Principal components, Classification, High-throughput

1 Introduction

With the development of microarrays technology, more and more statistical methods have been applied to the disease classification using microarray gene expression data. Microarray data sets often have a large number of features (genes), but only a very limited number of samples are available, which presents unique challenges to feature selection and predictive modeling. In general, these statistical methods can be divided into two categories: one is the supervised classification methods. For example, Golub et al. developed a "weighted voting method" to classify two types of human acute leukemias [1]. Radmacher et al. constructed a 'compound covariate prediction' to predict the BRCA1 and BRCA2 mutation status of breast cancer [2]. Studies have shown that given the same set of selected features, different classification methods often perform quite similarly and simple methods like diagonal linear discriminant analysis (DLDA) and k nearest neighbor (kNN) normally work remarkably well [3]. Thus, finding the most informative features is a crucial

task in predictive modeling from microarray data [4-5]. Another is the unsupervised clustering approaches, which are usually used to determine gene clusters that are mostly correlated with clinical outcomes [6]. However, the clustering approach is purely exploratory and methods that can be used to assess the significance of the clustering results are required. It has been widely known that most diseases (such as cancer) are 'caused' or influenced by multiple gene variations more often than only a single gene. Traditional microarray-based disease classification approaches use only individual differentially expressed genes as biomarkers to discriminate classes of cancer and normal samples. However, a large proportion of such genes are irrelevant and functional correlations among those genes are ignored. Since the genes with the best discriminative power are likely to correspond to a limited set of biological functions or pathways, it is rational to focus on these key functional expression patterns for disease prediction. This approach may then provide clues as for the types of biological processes that underlie the expression patterns of sets of genes.

Some attempts have been made to integrate the unsupervised gene clustering and the supervised disease classification approaches into a unified classification process. Li et al. developed cluster-Rasch models, in which a model-based clustering approach was first used to cluster genes and then the discretized gene expression values were input into a Rasch model to estimate a latent factor associated with disease classes for each gene cluster [7]. The estimated latent factors were finally used in a regression analysis for disease classification. They demonstrated that their results were comparable to those previously obtained, but the discretization of continuous gene expression levels usually results in a loss of information. Hastie et al. proposed a tree harvest procedure for finding additive and interaction structure among gene clusters, in their relation to an outcome measure [8]. They found that the advantage of the method could not be demonstrated due to the lack of rich samples. Dettling et al. presented a new algorithm to search for gene clusters in a supervised way. The average expression profile of each

cluster was considered as a predictor for traditional supervised classification methods. However, using simple averages will discard information about the relative prediction strength of different genes in the same gene cluster [9]. Yu also compared different approaches to form gene clusters. The resulting information was used for providing sets of genes as predictors in regression [10].

Recently, gene co-expression networks have become a more and more active research area [11-14]. A gene co-expression network is essentially a graph where nodes in the graph correspond to genes, and edges between genes represent their co-expression relationship. The gene neighbor relations (such as topology) in the networks are usually neglected in traditional cluster analysis [13]. One of the major applications of gene co-expression network has been centered in identifying functional modules in an unsupervised way [11-12], which may be hard to distinguish members of different sample classes. Recent studies have shown that prognostic signatures that could be used to classify the gene expression profiles from individual patients can be identified from network modules in a supervised way [14].

In this paper we explored a clustering-based approach for classification of high-throughput gene expression data. Specifically, we first used a seed based approach to identify correlation-shared gene clusters from gene network. Each of these clusters included a differentially expressed gene between sample classes, which was treated as a seed, and a set of other genes highly co-expressed with the seed gene; then we performed principal component analysis (PCA) to extract meta-gene expression profiles; finally a supervised PCA-based logistic regression (LR) model was built to predict disease outcomes. We call the method as CPCLR. The method returned signature components of tight co-expression with good predictive performance. The performance of this method was compared with other state-of-the-art classification methods. We demonstrated that the approach has the advantage of computational simplicity and efficiency with lower classification error rates than the compared classification methods.

The remainder of this paper is organized as follows: Section 2 gives a detailed description of our classification method and briefly discusses other methods to be compared as well as the evaluation strategy; Section 3 presents the results based on six classification methods and three case studies; Section 4 summarizes our findings in the study.

2 Methods

2.1 CPCLR algorithm

CPCLR classification algorithm includes three stages: 1) construct correlation-sharing based gene clusters; 2)

extract meta-gene expression profiles from the constructed clusters using PCA; 3) classify samples using PCA-based LR model. Here we briefly described each of the three stages:

Stage 1: construct correlation-sharing based gene clusters. We modified the correlation-sharing method developed by Tibshirani and Wasserman [15], which was originally proposed to detect differential gene expression. The approach works in the following steps:

A: Compute test statistic $T_i (i = 1, 2, \dots, p)$ for each gene i using the standard t-statistic or a modified t-statistic, such as significance of microarrays (SAM) [16].

B: Select seed genes having larger absolute test statistic values, say top m genes.

C: Find the cluster membership s for each selected seed gene i^* . The cluster assignments can be characterized by a many to one mapping. That is, one seeks a particular encoder $C_r(i^*)$ that maximizes

$$i_s^* = \max_{\{0 \leq r \leq 1\}} \text{ave}_{i \in C_r(i^*)} |T_i| \quad (1)$$

where $C_r(i^*) = \{s : \text{abs}(\text{corr}(x_{i^*}, x_s)) \geq r\}$. The set of genes s for each seed gene i^* is an adaptively chosen cluster, which maximizes the average (*ave*) differential expression signal around gene i^* . The set of identified genes s should have absolute (*abs*) correlation (*corr*) with i^* larger than r . The advantage of the correlation-sharing based clustering method is that the membership in different clusters can be overlapped rather than mutually disjoint.

Stage 2: Principal component analysis of correlation-shared expression profiles: To do this, for each of the seed-based gene cluster, we performed principal component analysis. Specifically, for a given gene cluster with C genes, assume $x^{(j)} = (x_{1j}, x_{2j}, \dots, x_{Cj})^t$ be expression indices of C genes in the j -th sample and t denotes transpose of a vector. Let Σ be covariance matrix of x with dimension $C \times C$. All positive eigenvalue of Σ are denoted by $\lambda_1 > \lambda_2 > \dots > \lambda_C$. The first PC score of the j -th sample is given by $x_j^* = e_1^t x^{(j)}$, where e_1 is the eigenvector associated with λ_1 . Therefore, we can define the super-gene expression profile for N samples in a seed-based gene cluster as $x^* = \{x_1^*, x_2^*, \dots, x_N^*\}^t$. The estimated values for the coefficient e_1^t (eigenvector) of the first PC can be computed using singular value decomposition (SVD) [17]. Briefly, assume X be an $N \times C$ matrix with normalized gene expression values of C genes in a given cluster, so we can express the SVD of X as $X = ULA^T$,

where $U = \{u_1, u_2, \dots, u_d\}$ is a $N \times d$ matrix ($d = \text{rank}(X)$), $L = \text{diag}\{l_1^{1/2}, l_2^{1/2}, \dots, l_d^{1/2}\}$ is a $d \times d$ diagonal matrix where l_k is k -th eigenvalue of $X^T X$, $A = \{e_1, e_2, \dots, e_d\}$ is a $C \times d$ matrix where e_k is eigenvector of associated with λ_k and coefficients for defining PC scores. Magnitude of loadings for the first principal component score can be viewed as an estimate of the amount of contribution from the clustered genes.

Stage 3: Classification using PCA-based logistic regression model: Assume Y is a categorical variable indicating the disease status (such as cancer or no cancer). Here we only focus on binary classification and suppose that $Y=1$ denotes the presence and $Y=0$ indicates the absence of the disease. Therefore, we can have following supervised PCA-based logistic regression model:

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \sum_{i^*}^m \beta_{i^*} PC1_{i^*j} + \varepsilon_j \quad (2)$$

where $p_j = \Pr(Y_j = 1 | PC1_{i^*j}, i^* = 1, 2, \dots, m)$. $PC1_{i^*j}$ is the first principal component score estimated

from the seed gene cluster i^* for sample j and represents the latent variable for the underlying biological process associated with this group of genes. The model was fitted using *GLM* function in stats R package.

2.2 Method Comparisons

We compared the prediction performance of CPCLR with other established classification methods, which include, diagonal linear discriminant analysis (DLDA), logistic regression (LR) model, one nearest neighbor method (1NN), support vector machines (SVM) with linear kernel and recursive partitioning and regression trees (Trees). We used the implementation of these methods in different R packages (<http://cran.r-project.org/>), which are *sma* for DLDA, *stats* for LR, *class* for 1NN, *e1071* for SVM and *rpart* for Trees. Default parameters were used. In the comparison, we selected seed genes using t-test and SAM and evaluated the performance of DLDA, LR, 1NN, SVM and Trees using different number of top seed genes and that of CPCLR using the gene clusters built on the selected seed genes.

2.3 Cross-validation

We performed ten-fold cross-validation to evaluate the performance of these classification methods. The basic principle is that we split all samples in a study into 10 subsets of (approximately) equal size, set aside one of the

subsets from training and carried out seed gene selection, gene cluster construction, extracted super-gene expression profiles and classifier fitting using the remaining 9 subsets. We then predicted the class label of the samples in the omitted subset based on the constructed classification rule. We repeated this process 10 times so that each sample is predicted exactly once. We determined the classification error rate as the proportion of the number of incorrectly predicted samples to the total number of samples in a given study. This 10-fold cross-validation procedure was repeated 10 times and the averaged error rate was reported.

3 Experimental Results

3.1 Real datasets

We applied the CPCLR algorithm and the established classification methods mentioned in Section 2.2 to three microarray data sets. The detailed description of these data sets is shown in Table 1. We got the preprocessed Colon cancer microarray expression data from <http://genomics-pubs.princeton.edu/oncology/>. For prostate cancer and lung cancer microarray data, we downloaded the raw data from gene expression omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) and preprocessed them using robust multi-array average (RMA) algorithm [18].

Table 1: Descriptive characteristics of data sets used for classification

Disease	Groups	No. Samples	No. Genes	Studies
Colon Cancer	Tumor/Normal	40 / 22	2000	[6]
Prostate Cancer	Tumor/Normal	50 / 38	12635	[19]
Lung Cancer	Tumor/Normal	60 / 69	22215	[20]

Tables 2, 3 and 4 listed the prediction performance of different classification methods applied to colon cancer, prostate cancer and lung cancer microarray gene expression data using different number of top seed genes. As we can see, for the colon and lung cancer data sets, CPCLR algorithm has better or comparable classification performance than other well-established classification methods based on different number of top seed genes or significantly differentially expressed genes (Tables 2, 4 and 5). However, for the prostate cancer data, the best performance was observed by using SVM predictors (Table 3). In order to save the time to search for genes which were correlated with a given seed gene and maximized their averaged test statistic value (formula 1), we tested 10

cutoffs of correlation r from 0.5 to 0.95 with interval 0.05. We observed that the averaged correlation of genes in the constructed gene cluster is usually between 0.65 and 0.85 with the number of genes in the clusters from 2 to 60, suggesting the genes in the constructed gene clusters are highly co-expressed.

Table 2: Error rates (%) of six classification methods applied to colon cancer data set

No. Genes	DLDA	1NN	Tree	SVM	LR	CPCLR
5	11.3	21.0	22.6	11.3	11.3	9.7
10	17.7	16.1	29.0	12.9	14.5	9.7
15	12.9	12.9	24.2	14.5	12.9	11.3
20	12.9	16.1	25.8	12.9	14.5	11.3
30	12.9	16.1	19.4	14.5	19.4	12.9

Table 3: Error rates (%) of six classification methods applied to prostate cancer data set

No. Genes	DLDA	1NN	Tree	SVM	LR	CPCLR
5	23.9	26.1	22.7	21.6	22.7	21.6
10	19.3	28.4	31.8	17.0	26.1	19.3
15	22.7	26.1	29.5	26.1	26.1	23.9
20	22.7	25.0	27.3	19.3	21.6	20.5
30	21.6	23.9	29.5	21.6	22.7	21.6

Table 4: Error rates (%) of six classification methods applied to lung cancer data set

No. Genes	DLDA	1NN	Tree	SVM	LR	CPCLR
5	17.0	18.6	20.1	16.2	19.3	17.0
10	14.7	18.6	19.3	17.0	20.1	14.7
15	16.2	20.1	17.8	13.2	17.8	15.5
20	16.2	17.0	19.3	17.8	19.3	15.5
30	12.5	13.2	19.3	14.7	20.1	12.5

We also used SAM [16] to select top seed genes and evaluated the prediction performance following the same procedure as described above. Similar prediction results were also observed as shown in Table 5 for lung cancer data. Overall, the CPCLR method has lower error rate than other being compared classification methods.

In all cases, we found that the simple method, DLDA, works well. Its performance is comparable with the advanced method, such as SVM. We also observed that the performance of the predictors with more genes is not necessary better than that of the predictors with fewer genes. For example, the best performance was obtained with only 5 genes for CPCLR predictors in colon cancer data set (Table 2), 10 genes for SVM predictors in prostate

Table 5: Error rates of six classification methods applied to lung cancer data set (seed genes selected by SAM)

No. Genes	DLDA	1NN	Tree	SVM	LR	CPCLR
5	17.0	19.3	22.5	16.2	18.6	17.8
10	17.0	20.9	19.3	17.8	17.8	15.5
15	14.7	20.1	22.5	14.6	20.1	13.2
20	16.2	18.6	17.8	18.6	17.0	15.5
30	17.8	13.2	19.3	10.1	14.7	10.1

cancer data set (Table 3). For lung cancer data set, the best performance was observed using 30 genes for DLDA and CPCLR predictors (Table 4).

4 Discussions and Conclusions

In this study we investigated a correlation-sharing based method for classification of high-throughput gene expression data. The core idea of the method is to identify 'essential' correlation structure among genes and extract representative features from the correlated gene clusters in a supervised classification procedure. The method takes into account the fact that genes act in networks and the gene clusters identified from the networks act as the features in constructing a classifier. The rationale is that we usually expect tightly co-expressed genes to have a meaningful biological explanation. For example, if gene A and gene B has high correlation, it sometimes hints that the two genes belong to the same pathway or are co-expressed. Instead of using individual genes as predictors in our classification models, we constructed meta-gene expression profiles representing information from each co-expressed gene cluster as predictors to classify disease outcomes. The advantage of this method over other methods has been demonstrated by three real data sets. Our results show that this algorithm is working well for improving class prediction.

5 References

- [1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring"; *Science*, vol. 286, pp. 531-536, 1999.
- [2] M.D.Radmacher, L.M. McShane, R. Simon. "A paradigm for class prediction using gene expression profiles"; *J Comput Biol*, vol. 9, pp. 505-512, 2002.
- [3] S. Dudoit, J. Fridlyand, T.P. Speed. "Comparison of discrimination methods for the classification of tumors

- using gene expression data"; *Journal of the American Statistical Association*, vol. 97, pp. 77-87, 2002.
- [4] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys. "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods"; *Bioinformatics*, vol. 26, pp. 392-398, 2010.
- [5] T. Elizabeth, O. Leonardo, B. Pilar, A. Laura. "Multiclass classification of microarray data samples with a reduced number of genes"; *BMC Bioinformatics*, vol. 12, 59, 2011.
- [6] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays"; *Proc Natl Acad Sci U S A*, vol. 96, pp. 6745-6750, 1999.
- [7] H. Li, F. Hong. "Cluster-Rasch models for microarray gene expression data"; *Genome Biol.*, vol. 2, pp. 0031.1-0031.13, 2001.
- [8] T. Hastie, R. Tibshirani, D. Botstein, P. Brown. "Supervised harvesting of expression trees"; *Genome Biol.*, vol. 2, pp. 0003.1-0003.12, 2001.
- [9] D. detting, P. Bühlmann. "Supervised Clustering of Genes"; *Genome Biol.*, vol. 3, pp. 0069.1-0069.15.
- [10] X. Yu. "Regression methods for microarray data"; Ph.D. thesis, Stanford University, 2005.
- [11] L. Elo, H. Jarvenpaa, M. Oresic, R. Lahesmaa, T. Aittokallio. "Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process"; *Bioinformatics*, vol. 23, pp. 2096-103, 2007.
- [12] A. Presson, E. Sobel, J. Papp, C. Suarez, T. Whistler, M. Rajeevan, S. Vernon, S. Horvath. "Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome"; *BMC Syst Biol.* vol. 2, 95, 2008.
- [13] S. Horvath, J. Dong. "Geometric interpretation of gene coexpression network analysis"; *PLoS Comput Biol.* vol.4, e1000117, 2008.
- [14] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, J. L. Wrana. "Dynamic modularity in protein interaction networks predicts breast cancer outcome"; *Nat Biotechnol.* vol. 27, 199-204, 2009.
- [15] R. Tibshirani, L. Wasserman. "Correlation-sharing for detection of differential gene expression"; *arXiv*, math.ST, math/0608061.
- [16] V. Tusher, R. Tibshirani, G. Chu. "Significance analysis of microarrays applied to the ionizing radiation response"; *Proc Natl Acad Sci USA*, vol. 98, pp. 5116-5121, 2001.
- [17] I.T. Jolliffe. *Principal component analysis*. Springer, New York, 2002.
- [18] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, T. P. Speed. "Summaries of Affymetrix GeneChip probe level data"; *Nucleic Acids Research*, vol. 31, pp. E15, 2003.
- [19] R. O. Stuart, W. Wachsman, C.C. Berry, J. Wang-Rodriguez, L. Wasserman, I. Klacansky, D. Masys, K. Arden, S. Goodison, M. McClelland, Y. Wang, A. Sawyers, I. Kalcheva, D. Tarin, D. Mercola. "In silico dissection of cell-type-associated patterns of gene expression in prostate cancer"; *Proc Natl Acad Sci USA*, vol. 101, pp. 615-620, 2004.
- [20] A. Spira, J.E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M.E. Lenburg, J.S. Brody. "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer"; *Nat Med.*, vol.13, pp. 361-366, 2007.

Markov Model Checking of Probabilistic Boolean Networks Representations of Genes

Marie Llubes¹, Jaime Seguel² and Jaime Ramírez-Vick³

^{1,2} Electrical and Computer Engineering Department, University of Puerto Rico, Mayagüez, Puerto Rico

³ General Engineering Department, University of Puerto Rico, Mayagüez, Puerto Rico

Abstract - *Our goal is to develop an algorithm for the automated study of the dynamics of Probabilistic Boolean Network (PBN) representation of genes. Model checking is an automated method for the verification of properties on systems. Continuous Stochastic Logic (CSL), an extension of Computation Tree Logic (CTL), is a model-checking tool that can be used to specify measures for Continuous-time Markov Chains (CTMC). Thus, as PBNs can be analyzed in the context of Markov theory, the use of CSL as a method for model checking PBNs could be a powerful tool for the simulation of gene network dynamics. Particularly, we are interested in the subject of intervention. This refers to the deliberate perturbation of the network with the purpose of achieving a specific behavior. This is attained by selectively changing the parameters in a node or set of nodes so that the network behavior can be controlled.*

Keywords: Gene Regulatory Network, Probabilistic Boolean Networks, Markov-chain, intervention, model-checking algorithms.

1 Introduction

The genome encodes thousands of genes whose products enable cell survival and numerous cellular functions. The amounts and the temporal pattern in which these products appear in the cell are crucial to the processes of life. A gene regulatory network is the collection of molecular species and their interactions, which together modulate the levels of these gene products. The dynamics due to both internal and external interactions constitute the state of a system. With the aid of Computer Science and Statistics, the study of gene regulatory network dynamics has become more feasible, and several models have been developed to simulate such dynamics. The knowledge of the intrinsic mechanisms that govern the network could provide the means to control its behavior. It is because of this that the development of an automated system capable of effectively simulating the behavior of a gene regulatory network may also provide the knowledge to alter such behavior in order to achieve a particular state of the system

or, on contrary, to prevent or to stop an undesirable behavior. This “guiding” of the network dynamics is referred to as intervention. The power to intervene with the network dynamics has a significant impact in diagnostics and drug design.

Biological phenomena manifest in the continuous-time domain. But, in describing such phenomena we usually employ a binary language, for instance, expressed or not expressed; on or off; up or down regulated. Studies conducted restricting genes expression to only two levels (0 or 1) suggested that information retained by these when binarized is meaningful to the extent that it remains in a continuous domain [2]. This allows gene regulatory networks to be modeled using a Boolean paradigm. The drawback of using this formalism is that the interactions among genes are hard-wired rules. This unrealistic assumption precludes the self-organizing nature of biological systems and, therefore, mischaracterizes their dynamics. Self-organization gives the system robustness in presence of perturbations, showing spontaneous ordered collective behavior. PBNs and Boolean networks share this quality through the existence of attractors and absorbing states, which act as a form of memory for the system.

PBNs, like Boolean networks, are rule-based. But, unlike the latter, they are not inherently deterministic using multiple rules, or “predictors”. This makes PBNs robust in the face of the environmental and biological uncertainty. Markov theory allows us to study the dynamic behavior of PBNs in the context of Markov Chains. They explicitly represent probabilistic relationships between genes, allowing quantification of influence between genes. Because of this, PBNs are better suited than Boolean networks for modeling such systems. Nevertheless, given the exponentially growth in the number of states a gene can be in (2^n states for n genes), answering questions on the best way to reach or avoid particular state(s) may be cumbersome if performed through exhaustion. Model-checking algorithms have the ability of automatically check if a certain condition is met under given specifications. Thus, it could answer questions as the one previously stated efficiently. This would greatly facilitate the intervention or deliberate perturbation of a network to achieve a desired response. This research studies the union between PBNs in

the context of Markov theory and model checking techniques for Continuous-time Markov chains.

2 Model Selection

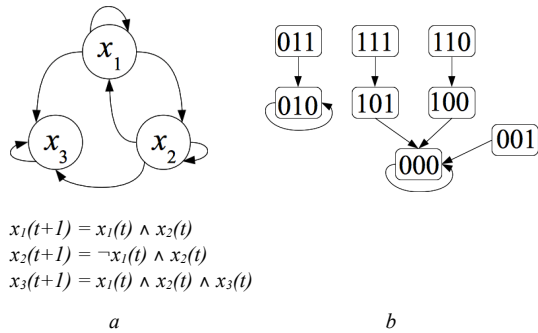
2.1 Boolean Network Model

A Boolean network is a set of Boolean variables whose state is determined by other variables in the network. Formally:

A Boolean network $G(V, F)$ is defined by a set of nodes $V = \{x_1, \dots, x_n\}$, and a list of Boolean functions $F = (f_1, \dots, f_n)$. Each $x_i \in V, i=1, \dots, n$, is a binary variable representing a gene which takes value from $\{0, 1\}$. There are k_i genes assigned to gene x_i , whose value at time t determine the value at time $t+1$ of x_i by means of a Boolean function $f_i \in F$. That is, the mapping $j_k: \{1, \dots, n\} \rightarrow \{1, \dots, n\}, k = 1, \dots, k_i$ determines the “wiring” of gene x_i and we can write [2]:

$$x_i(t+1) = f_i(x_{j_1(i)}(t), x_{j_2(i)}(t), \dots, x_{j_{k_i}(i)}(t)). \quad (1)$$

A network with n genes has 2^n states. Each of these states represents the pattern of expression of the individual genes. Pattern expressions are sometimes called gene activity profiles (GAPs). Some of these GAPs are attractors in the sense that the network flow eventually gets trapped in them. They represent the memory of the system. Attractors may be composed by cycles of states. Figure 1 gives an example of a Boolean network.



$$\begin{aligned} x_1(t+1) &= x_1(t) \wedge x_2(t) \\ x_2(t+1) &= \neg x_1(t) \wedge x_2(t) \\ x_3(t+1) &= x_1(t) \wedge x_2(t) \wedge x_3(t) \end{aligned}$$

Figure 1. Example of Boolean network
a) Boolean network with three nodes
b) State transition diagram

The relationships between genes are determined from experimental data. A coefficient of determination (COD) is used in this endeavor to discover such associations. The COD measures the quality of a predictor in using an observed gene set to infer a target gene set, in the absence of observations. In order to further illustrate this, let x_i be a target gene, which we wish to predict by observing the set of genes $x_{i1}, x_{i2}, \dots, x_{ik}$. Suppose that $f(x_{i1}, x_{i2}, \dots, x_{ik})$ is an optimal predictor of x_i relative to some error measure ε . Let

ε_{opt} be the optimal error achieved by f . Then, the COD for x_i relative to the set $x_{i1}, x_{i2}, \dots, x_{ik}$ is defined as:

$$\theta = \frac{\varepsilon_i - \varepsilon_{\text{opt}}}{\varepsilon_i} \quad (2)$$

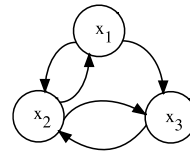
where ε_i is the error of the best (constant) estimate of x_i in the absence of any conditional variables [2].

2.2 PBN Model

The open nature of biological systems brings about a significant uncertainty into the model. One way of coping with this difficulty is to pass the uncertainty to the predictor, by synthesizing a number of good performance predictors. Each one of them contributes its own prediction proportionally to its determinative potential, which is given by the COD. More formally, given genes $V = \{x_1, \dots, x_n\}$, we assign to each x_i a set $F_i = \{f_1^{(i)}, \dots, f_{l(i)}^{(i)}\}$ of Boolean functions representing the “top” predictors for the target gene x_i . Thus, the PBN acquires the form of a graph $G(V, F)$ where $F = (F_1, \dots, F_n)$ [4], and each F_i in F is as previously described. At each point in time or step of the network, a function $f_j^{(i)}$ is chosen with probability $c_j^{(i)}$ to predict gene x_i . Using a normalized COD [2]:

$$c_j^{(i)} = \frac{\theta_j^{(i)}}{\sum_{k=1}^{l(i)} \theta_k^{(i)}} \quad (3)$$

where $\theta_j^{(i)}$ is the COD for gene x_i relative to the genes used as inputs to predictor $f_j^{(i)}$. Figure 2 provides an example of a PBN.



$$\begin{aligned} f_1^{(1)}: x_1(t+1) &= \bar{x}_2(t) & c_1^{(1)} &= 1 \\ f_1^{(2)}: x_2(t+1) &= \bar{x}_1(t) & c_1^{(2)} &= 0.3 \\ f_2^{(2)}: x_2(t+1) &= x_1(t) \wedge x_3(t) & c_2^{(2)} &= 0.7 \\ f_1^{(3)}: x_3(t+1) &= x_1(t) & c_1^{(3)} &= 0.6 \\ f_2^{(3)}: x_3(t+1) &= \bar{x}_1(t) \wedge x_2(t) & c_2^{(3)} &= 0.4 \end{aligned}$$

Figure 2. PBN of three nodes and its predictors

At a given instant in time, the predictors selected for each gene determine the state of the PBN. These predictors are contained on a vector of Boolean functions, where the i^{th} element of that vector contains the predictor selected at that time instant for gene x_i . This is known as a *realization* of the PBN. If there are N possible realizations, then there are N possible vector functions, f_1, f_2, \dots, f_N , each of the form $f_k = (f_{k1}^{(1)}, f_{k2}^{(2)}, \dots, f_{kn}^{(n)})$, for $k = 1, 2, \dots, N, 1 \leq k_i \leq l(i)$ and where $f_{ki}^{(i)} \in F_i$ ($i=1, \dots, n$). In other words, the vector function $\mathbf{f}_k: \{0, 1\}^n \rightarrow \{0, 1\}^n$ acts as a transition function

(mapping) representing a possible realization of the entire PBN. (See Figure 3). Thus, we have the matrix K of realizations:

$$K = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_N \end{bmatrix} \quad (4)$$

Assuming independence of the predictors, $N = \prod_{i=1}^n \mathcal{I}(i)$. Each realization k can be selected with $P_k = \prod_{i=1}^n \mathcal{C}_{k_i}^{(i)}$. The probability of transitioning from state (x_1, \dots, x_n) to (x'_1, \dots, x'_n) is given by [3]:

$$\Pr\{(x_1, \dots, x_n) \rightarrow (x'_1, \dots, x'_n)\} = \sum_{k=1}^N P_k \left[\prod_{i=1}^n \underbrace{(1 - |f_{ki}^i(x_1, \dots, x_n) - x'_i|)}_{\in \{0,1\}} \right] \quad (5)$$

3 Perturbation And Intervention

As an open system, the genome receives inputs from the outside. Such stimuli can either activate or inhibit gene expression; therefore it is necessary for the model of such a system to reproduce this behavior. This is achieved by the inclusion of a realization in the form of a random *perturbation vector* $\gamma \in \{0, 1\}^n$. Lets assume that a gene can get independently perturbed with probability p . Then if $\gamma_i = 1$ the i^{th} gene is flipped, otherwise it is not. For simplicity, we will assume that $\Pr\{\gamma_i = 1\} = E[\gamma_i] = p$ for all $i = 1, \dots, n$ (i.e., independent and identically distributed). Let $x(t) \in \{0, 1\}^n$ be the state of the network at time t . Then, the next state x' is given by:

$$x' = \begin{cases} x \oplus \gamma, & \text{with probability } 1 - (1 - p)^n \\ \mathbf{f}_k(x_1, \dots, x_n), & \text{with probability } (1 - p)^n \end{cases} \quad (6)$$

where \oplus is component-wise addition modulo 2, and f_k is the transition function representing a possible realization of the

entire PBN, $k = 1, 2, \dots, N$ [2]. In presence of perturbation with probability p , the entrances in the state transition matrix are computed by [4]:

$$A(x, x') = \left(\sum_{k=1}^N P_k \left[\prod_{i=1}^n \underbrace{(1 - |f_{ki}^i(x_1, \dots, x_n) - x'_i|)}_{\in \{0,1\}} \right] \right) \times (1 - p)^n + p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')} \times 1_{[x \neq x']} \quad (7)$$

Most relevant to our research is the fact that, when performed in a deliberately way, a perturbation constitutes an intervention. We may introduce a perturbation vector for a set of selected genes for the purpose of achieving a desired state, or moving from an undesirable one, on the network. This can be done by perturbing those genes with greater impact on the global behavior, by perturbing a fewer number of genes, or by reaching the desired state as early as possible. In gene interactions, some genes used in the prediction of a target gene have more impact than others, making them more important, or of higher *influence*, thus, identifying these genes is highly relevant. Similarly, we can determine the sensitivity of a particular gene, defining it as the sum of all *influences* acting upon it. The sensitivity, in turn, defines the particular gene stability and independence. In [2, 4] a method to compute influences and sensitivities is given. One of the main benefits of determining influences and sensitivities of genes is that these allow the identification of vulnerable points in the network, or the ones most likely to affect its entire network if perturbed. Highly influential genes can control the dynamics of the network, making it possible to move to a different basin of attraction when perturbed. This kind of information may provide potential targets when an intervention is needed to obtain a desired state of the system.

4 Model-Checking Algorithms

Given a PBN model of a gene regulatory network, we are interested in knowing (in an automated way) if certain state(s) are reachable under particular conditions, or specifications. This is the verification problem, to which model checking is an instance of. Because these are mathematical problems, we formulate our specifications

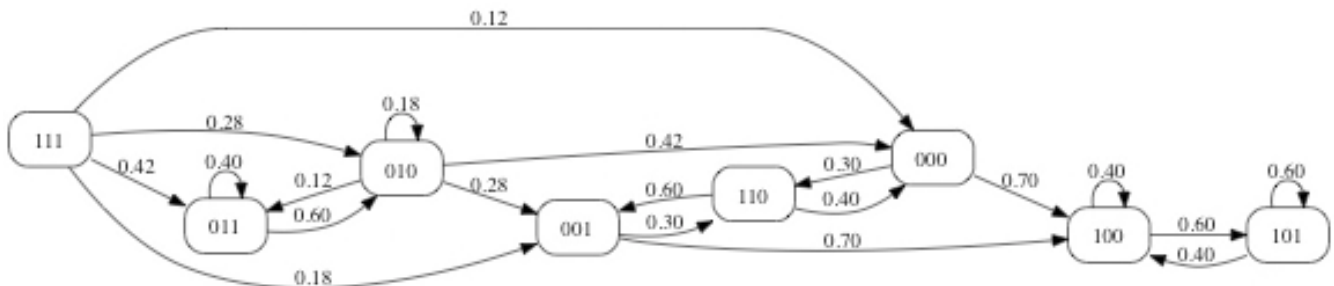


Figure 3. PBN state transition diagram

using mathematical logic. Temporal logics have been crucial in the development of model checking, because of its compact way of expressing correctness properties, and the fact that the Small Finite Model Theorem makes it decidable [5]. Its branching time logic, Computation Tree Logic (CTL) allows us to build compound formulas from the nesting of subformulas. The semantics of temporal logic formulas are defined over a finite transition system (Kripke structure).

The specification of measures of interest over systems is usually done using state-based properties (steady and transient state), due to the difficulty of specifying path-based measures. Continuous Stochastic Logic (CSL) is a probabilistic timed extension of CTL that provides means for specifying measures both state and path-based for Continuous-time Markov chains (CTMC). Numerical methods to model-check CSL over finite-state CTMC are explored in [1].

4.1 Continuous-time Markov chains

The Kripke structure to consider for CSL model checking is a CTMC, where the edges are equipped with probabilistic timing information. Let AP be a fixed, finite set of atomic propositions [1]:

A CTMC M is a tuple (S, \mathbf{R}, L) with S as a finite set of states, $\mathbf{R}: S \times S \rightarrow \mathbb{R}_{\geq 0}$ as the transition matrix, and $L: S \rightarrow 2^{AP}$ as the labeling function.

Each state $s \in S$ corresponds to a GAP of the PBN. \mathbf{R} is the transition probability matrix of the state-transition network. Function L assigns to each state $s \in S$ the set $L(s)$ of atomic propositions $a \in AP$ that are valid in s . We allow self-loops by having $\mathbf{R}(s,s) > 0$. The probability that the transition $s \rightarrow s'$ can be triggered within t time units is $1 - e^{-\mathbf{R}(s,s') \cdot t}$. The probability to move from a state s to state s' within t time units is given by [1]:

$$\mathbf{P}(s, s', t) = \mathbf{R}(s, s') \cdot (1 - e^{-t}) \quad (8)$$

The probability of moving from a nonabsorbing (with at least one transition out of it) state s to s' by a single transition is $\mathbf{P}(s, s') = \mathbf{R}(s, s')$. For an absorbing state s , $\mathbf{P}(s, s') = 0$ for any state s' [1].

For our PBN example (Fig. 2 and 3) the Markov model would have the set of states $S = \{(000), (001), (010), (011), (100), (101), (110), (111)\}$.

\mathbf{R} is an 8×8 matrix containing the transition probabilities between states. $AP = \{x_i \in \{0,1\}, i = 1, \dots, n\}$. $L(x_1 \dots x_n) = \{x_1 \dots x_n, x_1 \dots x_n \in \{0,1\}^n\}$, for instance, $L(011) = (x_1 x_2 x_3 = 011)$. An initial distribution α , which can be a state or set of states, is imposed over the PBN. For this particular case, we assume an initial uniform joint distribution. This means each state has the same chance of being the initial state. Taking $s_0 = (111)$, a possible sequence of transitions, or computation, is $\{(111), (001), (100), (101), (100)\}$.

There are two major types of state probabilities for CTMC:

1. Transient-state probabilities, where the system is observed at a given time instant t :

$$\pi^M(\alpha, s', t) = \Pr_{\alpha} \{ \sigma \in \text{Path}^M \mid \sigma @ t = s' \}$$

2. Steady-state probabilities, where the system is observed when equilibrium has been reached:

$$\pi^M(\alpha, s') = \lim_{t \rightarrow \infty} \pi^M(\alpha, s', t)$$

The two types of measures shown above are state-based. However, we are also interested in the probability on paths through the CTMC obeying particular properties. To the best of our knowledge, suitable mechanisms to measure such properties have not been considered in the literature.

It is worth noting that Binary Decision Diagrams (BDDs), a powerful tool for model checking, are not all that useful in the contexts of PBNs models. What precludes its use is the fact that each state of the PBN, or GAP, contains a string of variables representing genes. As BDDs represents possible transitions for one variable, we would need a BDD for each variable contained in the string. The output would be ramifications of several BDD. As BDDs represents Boolean functions, their values can be directly obtained from the truth table of the predictors.

4.2 Continuous Stochastic Logic

Continuous Stochastic Logic (CSL) provides means to specify state as well as path-based performance and dependability measures for CTMCs in a compact and unambiguous way. This logic is basically a probabilistic timed extension of CTL [1].

Besides the standard steady-state and transient measures, the logic allows for the specification of constraints over probabilistic measures over paths through CTMCs. For instance, we may check the probability of going from state s to state s' within t time units, avoiding or visiting some particular intermediate states. Four types of measures can be identified:

1. *Steady-state* measures: The formula $S_{\leq p}(\Phi)$ imposes a constraint on the probability to be in some Φ state on the long run. For the PBN in the example above, $S_{\geq 0.4}(x_1 \wedge \neg x_2)$ states that there is at least a 40% probability that gene x_1 is expressed and gene x_2 is not expressed when the network reach equilibrium.
2. *Transient* measures: The combination of the probabilistic operator with the temporal operator $\diamond^{[t,t]}$ can be used to reason about transient probabilities. More specifically, $P_{\leq p}(\diamond^{[t,t]} at_{s'})$ is valid in state s if the transient probability at time t to be in state s' satisfies the bound $\leq p$.
3. *Path-based* measures: By the fact that P-operator allows an arbitrary path formula as the argument; much more general measures can be described. An example is the probability of reaching a certain set of states provided that all paths to these states obey certain properties.

4. *Nested* measures By nesting the P and S operators, more complex properties can be specified. These are useful to obtain a more detailed insight into the system's behavior and allow it to express probabilistic reachability that is conditioned on the system being in equilibrium.

The main benefits in using CSL for specifying constraints on measures of interest over CTMCs are[1]:

1. Since the specification is entirely formal, the interpretation is unambiguous. An important aspect of CSL is the possibility of stating performance and dependability requirements over a selective set of paths through a model, which was not possible before.
2. The possibility of nesting steady-state and transient measures provides a means to specify complex, though important measures in a compact and flexible way.

Once we have obtained the model (CTMC M) of the system under consideration, and specified the constraint on the measure of interest in CSL by a formula Φ , the next step is to model check the formula. The model-checking algorithm for CTL that supports the automated validation of Φ over a given state s in M , is adapted to these purposes. The basic procedure is as for model checking CTL: in order to check whether state s satisfies the formula Φ , we recursively compute the set $Sat(\Phi)$ of states that satisfy Φ and, finally, check whether s is a member of that set. For the non-probabilistic state operators, this procedure is the same as for CTL [1].

For the purpose of intervention, it would be necessary to know how likely are certain states to reach a steady-state on the network of genes. This information, and with the use of the influences and sensitivities previously explained, would aid in determining the genes that represent the best candidates for reaching a desired condition. For instance, if we want to verify if a particular state reach a steady-state condition with a certain probability, a very high-level algorithm would look as follows:

Input: PBN, state s , measure m , constraint c

Do:

1. Determine Bottom Strongly Connected Components BSCC of PBN.
2. *If* s isn't in some BSCC
Output "State specified doesn't reach steady state".
Stop.
3. *Else* continue.
4. Compute transition probabilities to state s .
5. Use constraint c to compare computed probabilities with m .
6. *If* constraint is met with some probability p
Output "The condition is met with probability p ".
Stop.
7. *Else*
Output "The system doesn't meet the desired condition".

Stop.

Given the state-explosion problem that characterizes this kind of model, abstraction is crucial. Bisimulation, the technique that guarantees exact abstraction, has a slight variation called lumping. It has been observed that lumping preserves all CSL formulas [1].

5 Future Work

At the moment, we are using CSL for describing some measurements on PBNs constructed with fictitious data. So far, steady-state measurements have been checked. Next, we have to develop algorithms for the particular cases of steady and transient states, as well as for path-based measurements. Then, we will test them with PBNs built from real data. This, of course, belongs to a feedback loop where results will be used to improve the algorithms. Once we are able to verify with certainty particular conditions against real data, we will work on the process of intervention. For this, we need to check the changes on the dynamics due to particular alterations of the parameters using a vector of perturbation.

6 Conclusions

PBNs make an ideal model representation for genetic networks because the robustness that multiple predictors give them. As Kripke structures representing state transitions of a system, CSL can be used as a model-checking algorithm for CTMC, expanding the traditional state-based measures with the use of path-based probabilistic measures. PBNs can be studied in the context of Markov theory, and Markov chains have been widely used to specify system performance and dependability. Because of this, it is our belief that a model-checking algorithm for CSL can be used to study the dynamics of CTMC representations of PBN used to model genetic regulatory networks in an effective way. Avoiding the matrix-based model, such algorithm would mitigate the impact of the analysis of an exponential size network. Intervention on the network would be attainable, due the information gathered thanks to the algorithm's ability of answering questions about the transition system of the PBN.

The breadth of logic topics that this research evolves through is worth remarking. In its most primitive formulation, relationships between genes can be described with the use of logic connectives from propositional logic. Predicate logic is then used for formulating questions on the state and dynamics of the system. Finally, temporal logic is the basis of the model checking algorithms that answers these questions.

7 Acknowledgements

This research is conducted in part thanks to the support of a RISE-NIH scholarship (1R25GM088023-01A1) granted to the first author.

8 References

- [1] Baier, C., Haverkort, B., Hermanns, H. and Katoen, J-P. Model-Checking Algorithms for Continuous-Time Markov Chains. *IEEE Transactions on Software Engineering*. 2003; 29(6):1-18.
- [2] Shmulevich, I., Dougherty, E.R. and Zhang, W. From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *IEEE*. 2002; 90(10).
- [3] Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics (Oxford, England)*. 2002; 18(2):261-74
- [4] Shmulevich, I., Dougherty, E.R. and Zhang, W. Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics (Oxford, England)*. 2002; 18(10):1319-31.
- [5] Emerson, E. Allen. The beginning of Model Checking: A personal perspective. Springer. 2008. Volume 5000; 27-45.

Reliability analysis of Classification of Gene Expression Data

Sujata dash, KMBB, Bhubaneswar, Orissa, India.
B.N. Patra, GIET, Gunupur, Orissa, India.

Abstract

Gene expression data usually contains a large number of genes, but a small number of samples. Feature selection for gene expression data aims at finding a set of genes that best discriminate biological samples of different types. Classification of tissue samples into tumor or normal is one of the applications of microarray technology. When classifying tissue samples, gene selection plays an important role. In this paper, we propose a two-stage selection algorithm for genomic data by combining some existing statistical gene selection techniques and ROC score of SVM and k-nn classifiers. The motivation for the use of a Support Vector Machine is that DNA microarray problems can be very high dimensional and have very few training data. This type of situation is particularly well suited for an SVM approach. The proposed approach is carried out by first grouping genes with similar expression profiles into distinct clusters, calculating the cluster quality, calculating the discriminative score for each gene by using statistical techniques, and then selecting informative genes from these clusters based on the cluster quality and discriminative score. In the second stage, the effectiveness of this technique is investigated by comparing ROC score of SVM that uses different kernel functions and k-nn classifiers. Then Leave One Out Cross Validation (LOOCV) is used to validate the techniques.

Key Words : Fisher Criterion, Golub Signal-to-Noise, Mann-Whitney Rank Sum Statistic, Leave One Out Cross Validation (LOOCV), Support Vector Machine(SVM)

1. Introduction

The problem of cancer classification has clear implications on cancer treatment. Additionally, the advent of DNA microarrays introduces a wealth of genetic expression information for many diseases including cancer. An automated or generic approach for classification of cancer or other

diseases based upon the microarray expression is an important problem. A generic approach to classifying two types of acute leukemia was introduced in Golub et. al.[7]. They achieved good results on the problem of classifying acute myeloid leukemia (AML) versus acute lymphoblastic leukemia (ALL) using 50 gene expressions. Their approach to classification consisted of summing votes for each gene on the test data, and looking at the sign of the sum. In this paper, four statistical techniques include Fisher Criterion, Golub Signal-to-Noise, traditional t-test, and Mann-Whitney Rank Sum Statistic are studied. The objective is to investigate the impact and importance of the gene selection techniques to the tissue classification performance. The effectiveness of this technique is investigated by comparing ROC score of SVM that uses different kernel functions: the dot product, quadratic dot product, cubic dot product and the radial basis function and the k-nn classifiers. The LOOCV is applied to validate the techniques. Results show that a better classification performance can be achieved by the classifiers if genes are first selected prior to the classification task.

2. Background on cDNA Microarrays

A *gene* consists of a segment of DNA which codes for a particular *protein*, the ultimate expression of the genetic information. A *deoxyribonucleic acid* or *DNA* molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. Each *nucleotide* comprises a phosphate group, a deoxyribose sugar, and one of *four nitrogen bases*. The four different bases found in DNA are adenine (A), guanine (G), cytosine (C), and thymine (T). The two chains are held together by hydrogen bonds between nitrogen bases, with base-pairing occurring according to the following rule: G pairs with C, and A pairs with T. While a DNA molecule is built from a four-letter alphabet, proteins are sequences of twenty different types of *amino acids*. The expression of the genetic

information stored in the DNA molecule occurs in two stages: (i) *transcription*, during which DNA is transcribed into *messenger ribonucleic acid* or *mRNA*, a single-stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine; (ii) *translation*, during which mRNA is translated to produce a protein. The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the *genetic code*, which relates nucleotide triplets to amino acids. cDNA microarrays consist of thousands of individual DNA sequences printed in a high density array on a glass microscope slide. The relative abundance of these DNA sequences in two DNA or cDNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. To this end, the two DNA samples or *targets* are labeled using different fluorescent dyes (e.g. a red-fluorescent dye Cy5 and a green-fluorescent dye Cy3), then mixed and hybridized with the arrayed DNA sequences or *probes*. After this competitive hybridization, fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the fluorescence intensity for each spot is indicative of the relative abundance of the corresponding DNA sequence in the two samples (see <http://rana.Stanford.EDU/software/> for more information on the measurement of fluorescence intensities). Microarrays are being applied increasingly in cancer research to study the molecular variations among tumors. This should lead to an improved classification of tumors, which in turn should result in progresses in the prevention and treatment of cancer. An important aspect of this endeavor is the ability to predict tumor types on the basis of gene expression data. We review below a number of prediction methods and assess their performance on the cancer datasets described in Section 3.

3. Gene Selection Technique

3.1 The Fisher Criterion[9], *fisher*, is a measure that indicates how much the class distributions are separated. The coefficient has the following formula:

$$fisher = \frac{(\mu_1 - \mu_2)^2}{(v_1 + v_2)} \quad (1)$$

where μ_i is the mean and v_i is the variance of the given gene in class i (there are two classes in this study, the positive class i.e. the normal tissue

sample and the negative class, i.e. the tumor tissue sample). It gives higher values to genes whose means differ greatly between the two classes, relative to their variances.

3.2 Golub Signal-to-Noise [7] used a measure of correlation that emphasizes the "Signal-to-Noise" ratio, *signaltonoise*, to rank the genes.

$$signaltonoise = \frac{(\mu_1 - \mu_2)}{(\sigma_1 + \sigma_2)} \quad (2)$$

Where μ_i is the mean and σ_i is the standard deviation of the gene in class i .

3.3 Traditional t-test [2], *t-test* assumes that the values of the two tissues variances are equal. The formula is as

$$ttest = \frac{(\mu_1 - \mu_2)}{\sqrt{\left(\frac{v_p}{n_1}\right) + \left(\frac{v_p}{n_2}\right)}} \quad (3)$$

where μ_i is the mean and v_p is the pooled variance,

$$v_p = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) / (n_1 + n_2 - 2), \text{ and } s_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n_i - 1) \quad (4)$$

3.4 The Mann-Whitney Rank Sum Statistic[2], *mann*, has the following formula:

$$mann = n_1 * n_2 * \frac{(n_1 + 1)}{2} - r_1 \quad (5)$$

Where n_i is the sizes of sample i , and r_1 is the sum of the ranks in sample 1.

These techniques are used because they look into the expression profiles of the genes in tumor and normal class [7]. In these techniques, each gene is measured for correlation with the class according to some measuring criteria in the formulas. The genes are ranked according to the score, S , and the top T numbers of genes are selected.

4. The Procedure for Gene Selection and Classification of Gene Expression Data

The procedure for this experiment is shown:

- i. Getting the data.
- ii. Setting the number of genes to be selected, T , the gene selection technique and the classifier. In this experiment, the number of genes to be selected is set to be from 1 to 100.
- iii. Applying LOOCV technique for validation and evaluation purpose, include leaving one sample out in S3.1, selecting genes in S3.2 and S3.3 and training and testing the classifiers from S3.4 to S3.6.
- iv. Calculating the ROC score based on the predicted class.
- v. The process is repeated for another number of genes to be selected, another gene selection technique and another classifier until all combinations are done.

INPUT: Gene expression data matrix, $X = \{x_{11}, \dots, x_{np}\}$ and the class label for each column, $y \in \{-1, 1\}$ where n is the number of genes and p is the number of tissue samples.

- S1. Get the data with p tissues (samples).
- S2. Pre-set the combination: the gene selection technique, the classifier and number of genes to be selected, T , (the experiment run from 1 to 100 genes).

LEAVE ONE OUT CROSS VALIDATION:

- S3. For $i = 1$ to p
 - S3.1 Leave i^{th} sample out.

GENE SELECTION:

- S3.2 Calculate the discriminative score, S , for each gene for the remaining $p-1$ samples, and rank the genes based on the score.
- S3.3 Select top T genes based on the ranked score, S .

CLASSIFICATION:

- S3.4 Train the classifier on the remaining $p-1$ samples by using the selected genes.
- S3.5 Test the trained classifier by using the left out i^{th} sample.
- S3.6 Record the predicted class from S3.5, put back the i^{th} sample.

ROC CALCULATION:

- S4. Calculate the ROC score based on the predicted class and save the ROC score.
- S5. Go to S2 for another number of genes to be selected, another gene selection technique and another classifier, stop if all combinations are done.

OUTPUT: ROC scores for each number of genes to be selected, T and gene selection technique.

5. Tissue Classification

Two classifiers are proposed to evaluate the validity of the selected genes. They are the SVM [1] with different kernels and the k -nn [6].

5.1 Support Vector Machines for Tissue Classification

Different kernel functions, the dot product and radial basis function are used for this experiment [4][5][8][1].

The dot product has the following formula:

$$K(x, y) = (x \cdot y + 1)^d \quad (6)$$

where x and y are the vectors of the gene expression data. The parameter d is an integer which decides a rough shape of a separator. In the case where d equals to 1, a linear classifier is generated, and in the case where d is equal to or more than 2, a nonlinear classifier is generated. In this experiment, when d is equals to 1, it is called the SVM dot product, when d is equals to 2, it is called the SVM quadratic dot product and when d is equals to 3, it is called the SVM cubic dot product.

The radial basis kernel has the following formula:

$$K(x, y) = \exp\left(\frac{-|x - y|^2}{2\sigma^2}\right) \quad (7)$$

where σ is the median of the Euclidean distances between the members and nonmembers of the class.

The main advantages of SVMs are that they are robust to outliers, converge quickly, and find the optimal decision boundary if the data is separable. Another advantage is that the input space can be mapped into an arbitrary high dimensional working space where the linear decision boundary can be drawn. This mapping allows for higher order interactions between the examples and can also find correlations between examples.

SVMs are also very flexible as they allow for a big variety of kernel functions.

5.2 k -nearest neighbor for Tissue Classification

The k -nn classifier is a simple classifier based on a distance metric between the testing samples and the training samples [6]. The main idea of the method is, given a testing sample s , and a set of training tuples T containing pairs of the form (t_i, c_i) where t_i 's are the expression values of genes and c_i is the class label of the tuple. Find k training sample with most similar expression value between t and s , according to a distance measure. The class label with the top voting among the k training sample is assigned to s . The main advantage of k -nn is it has the ability to model very complex target functions by a collection of less complex approximations. It is easy to program and understand. No training or optimization is required for this classifier. It is robust to noisy training data.

6. Result Evaluation Method

ROC score is used to analyze the results for the experiment. ROC score is also the area under the curve (AUC). ROC score is a common way for evaluating classification performance because it takes into account both false negative and false positive errors and it reflects the robustness of the classification. A random classification has a ROC score approaching 0.5 while a perfect classification with no error has a ROC score at 1. In this experiment, for each possible combination of number of genes to be selected, gene selection technique and classifier, the performance varying the number of genes from 1 to 100 are evaluated.

6.1 Results and Discussion

In this section, the impact and importance of gene selection to the classification performance is first studied. This is carried out by comparing the classification performance by using all genes and gene selected by statistical techniques which are mentioned above. After that, the classification performance for each classifier is compared. Finally, based on the classifier with the best classification performance, the effectiveness of each statistical technique to this classifier is discussed.

6.2 Importance of Gene Selection Technique Prior to Tissue Classification

Figure-1 shows the classification performance by using all genes and gene selected by using statistical techniques. The ROC scores recorded for the gene selection techniques in the figure are the average ROC scores for number of genes selected from 1 to 100. From the figure, by using all genes, the best performance is obtained by using SVMs with radial basis function while 1-nn, 2-nn and 5-nn have worst performance. 3-nn and 4-nn are comparable to each other when all genes are used. The performances of the classifiers are improved after genes are selected by gene selection techniques especially for k -nn classifier. This shows the importance of applying gene selection techniques to select informative genes prior to the classification task. Applying gene selection techniques in selecting genes helps in removing a large number of irrelevant genes which improves the classification performance. Since one of the advantages of SVMs is, it is robust to outliers and allows nonlinear classification to be done, gene selection techniques does not give big impact to its performance, but, a better performance still can be obtained after applying gene selection techniques, which can be seen from the figure. One might ask why there is still a need to do gene selection if the classification performance using SVM has little difference while using all the genes in the dataset compare to the selected subset of genes. One reason for this is that selecting subset of genes not only can help biologists to identify the potential genes rather than swimming in the huge dataset, it helps the classifier to build a better and simple rule for classifying future unknown data.

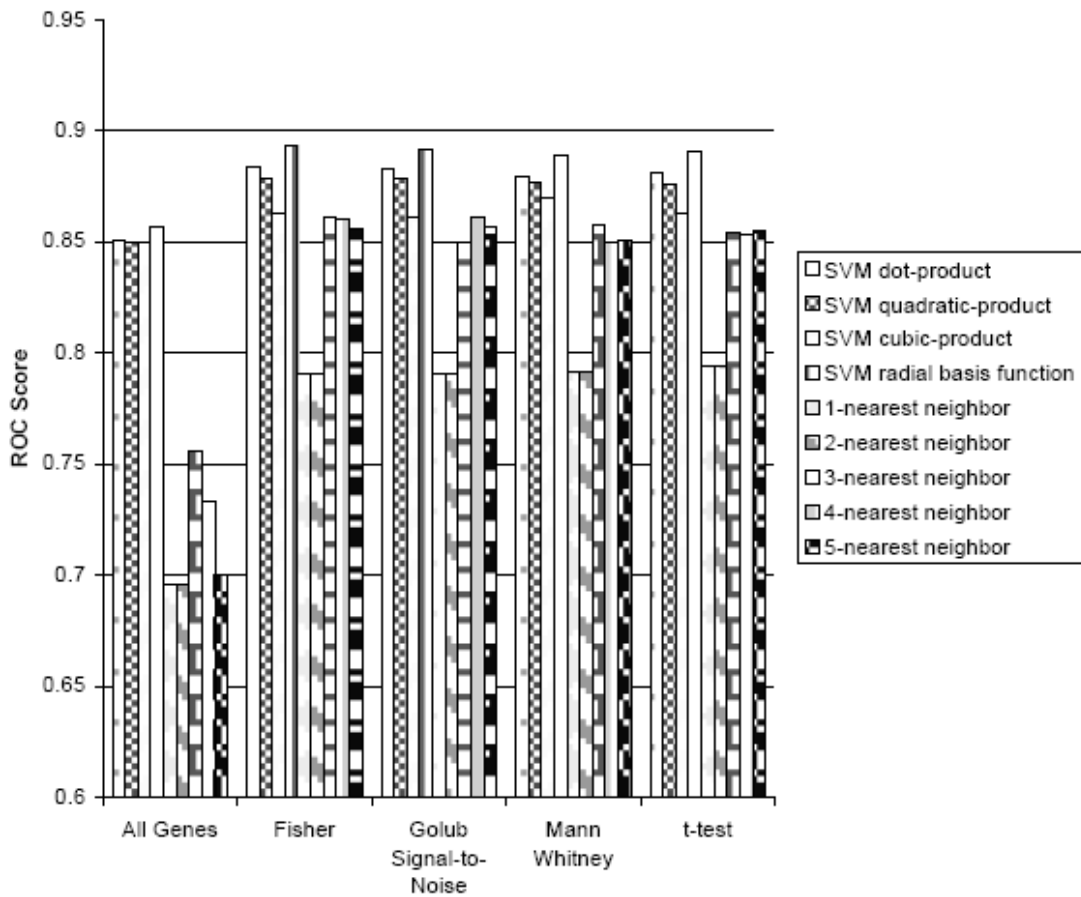


Figure 1: Classification performance by using all genes and genes selected by statistical techniques

This figure shows that a better classification performance can be achieved if genes are first selected by the gene selection techniques. However, which combination of statistical techniques and classifier and how many genes are needed for the best performance? Next section answers this question.

6.3 Classification Performance between Different Classifiers

Table-1 summarizes the performance for each SVM classifier. The ROC scores recorded in the table are the average ROC score over all trials with the number of selected genes from 1 to 100.

SVMs	Fisher	Golub	Mann	t-test
SVM_dot	0.88	0.88	0.88	0.88
SVM_quadratic	0.88	0.88	0.88	0.88
SVM_cubic	0.86	0.86	0.87	0.86
SVM_RBF	0.89	0.89	0.89	0.89

Table-1: Summary for classification performance by using SVMs with different kernels after gene selection by using statistical techniques

Table-1 show that, SVM radial basis function performs the best. Of the three, product kernels, dot-product and quadratic product have better ROC

score than cubic-product. These results indicate that over-fitting causes the misclassification for the cubic-product kernel. If more samples are obtained and they are not separable linearly, nonlinear classification may perform well [3].

Table-2 summarizes the performance for each k -nn classifier. The ROC scores recorded in the table are the average ROC score over all trials with the number of selected genes from 1 to 100.

k-nn	Fisher	Golub	Mann	t-test
1-nn	0.79	0.79	0.79	0.79
2-nn	0.79	0.79	0.79	0.79
3-nn	0.86	0.85	0.86	0.85
4-nn	0.86	0.86	0.85	0.85
5-nn	0.86	0.86	0.85	0.85

Table-2: Summary for classification performance by using different k -nn after gene selection using statistical techniques

Table-2 show that k -nn with k more than 2 outperform k which is equals to 1 and 2. One of the reasons for this to happen is that in the case of mislabeled training samples, it will have much greater effect on the classification result of 1-nn since one mislabel will result in misclassifying the test sample. 3-nn and 4-nn is less prone to bias in the data and more tolerable to noise since it makes use of several training samples to determine the class of a test sample.

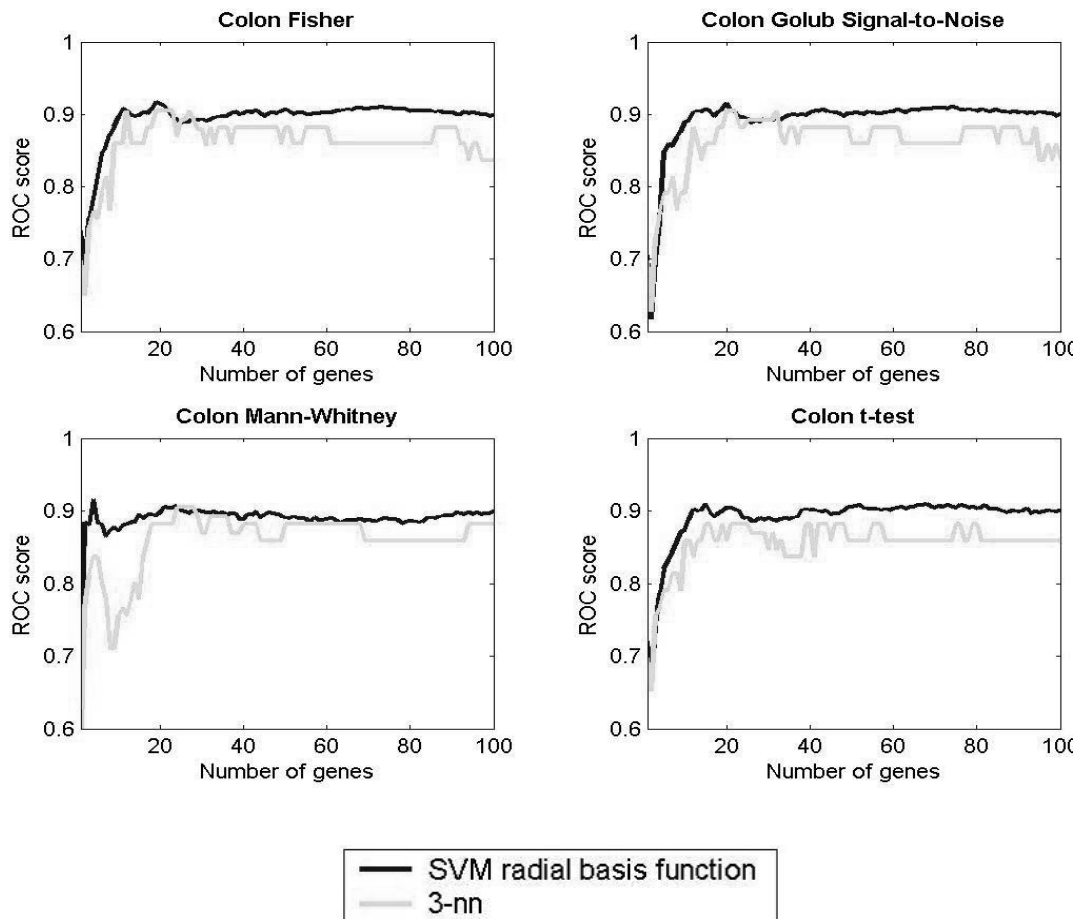


Figure-2: Classification performance between different classifiers after gene selection using statistical techniques (the best classifier is selected from SVM and k nn)

Figure-2 shows that SVM with radial basis function as the kernel function always produced higher ROC score than 3-nn. Generally, the results have lower ROC score with fewer genes for both classifiers. Lowest scores always drop between the numbers of genes from 1 to 15 except for Mann-

Whitney Rank Sum Statistic. One reason for the lower scores might be due to the characteristic of genes itself where genes do not act alone, but they interact with other genes for certain functions. For example, if Gene A and Gene B are in the same function it could be that they have similar regulation and therefore similar expression profiles. If Gene A has a good discriminative score it is highly likely that Gene B will, as well.

Hence the statistical techniques are likely to include both genes in a classifier, yet the pair of genes provides little additional information compared to either gene alone. If there are 5 functions in the dataset, 10 genes for each function, and if the genes in first function have the highest scores, so these 10 genes might be selected for the classification task. In this case, the genes being selected are highly redundant and thus provide little additional information. The peak performance for SVMs and k -nn always drop from the number of genes between 15 and 30. When the number of genes increase from 30 to 80 generally, the ROC score for SVMs and k -nn becomes more stable, because the possibility to select meaningful genes increase.

7. Summary

This paper reports the application of different statistical techniques to the colon dataset. These techniques include Fisher Criterion, Golub Signal-to-Noise, traditional t-test, and Mann-Whitney Rank Sum Statistic. By using these techniques, the data is rank based on the discriminative score and top T numbers of genes are selected. In conjunction with these gene selection techniques, several SVMs and k -nn classifiers are applied. Based on the genes selected by the gene selection techniques, ROC score of different combination of gene selection techniques and classifiers are obtained for analysis. The main objective of this experiment is to study the impact and importance of applying gene selection techniques prior to the classification task. Results show that a better classification performance is achieved by the classifiers if informative genes are first selected. However, finding a way to reduce redundant genes being selected in order to obtain a better classification performance is important.

8. Reference

- [1] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines And Other Kernel-Based Learning Methods*. New York: Cambridge University Press.
- [2] Devore, J.L. (1995). *Probability and Statistics for Engineering and the Sciences*. 4th edition, California: Duxbury Press.
- [3] Domura, D., Nakamura, H., Tsutsumi, S., Aburatani, H. and Ihara, S. (2002). Characteristics of Support Vector Machines in Gene Expression Analysis. *Genome Informatics*. (13):264 – 265.
- [4] Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison Of Discrimination Methods For The Classification Of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*. 97(576): 77 – 87.
- [5] Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002). Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. *Statistica Sinica*. (12):111 – 139.
- [6] Friedman, M. and Kandel, A. (1999). *Introduction to Pattern Recognition*. London: Imperial College Press.
- [7] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*.(286):531 – 537.
- [8] Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J.P. and Poggio, T. (1999). Support Vector Machine Classification of Microarray Data. *S. Technical Report 182*. AI Memo 1676, CBCL.
- [9] Smiatacz, M., Malina, W., Versatile Pattern Recognition System Based On Fisher Criterion. ul. G. Narutowicza 11/12, 80-952 Gdańsk

SESSION

BIOINFORMATICS DATABASE, DATAMINING, AND PATTERN DISCOVERY TECHNIQUES

Chair(s)

TBA

Clustering on Protein Sequence Motifs using SCAN and Positional Association Rule Algorithms

Bernard Chen¹, Ben Nordin¹, Sriram Bobba¹, Devendar Singireddy¹, Brad Taylor¹, Sinan Kockara,
and Mutlu Mete²

¹University of Central Arkansas, Department of Computer Science

²Texas A&M University-Commerce, Department of Computer Science

Abstract— The role of protein sequence motifs is in predicting functional or structural portion of other proteins including prosthetic attachment sites, enzyme-binding sites and DNA /RNA binding sites, and so on. A fixed window size is usually predefined to discover protein sequence motifs for many algorithms and techniques. However, the predefined window size may deliver a number of similar motifs simply shifted by some bases or including mismatches. In this paper, we use the positional association rules algorithm to form motifs network and adapt a Structural Clustering Algorithm for Networks named SCAN to recognize similar motifs. Although association rule based algorithms have been widely adapted in association analysis and classification, few of those are designed as clustering methods. With the SCAN analysis, the qualities of the clusters are further improved.

Index Terms— Positional Association Rules, SCAN, Protein Sequence Motifs

I. INTRODUCTION

Bioinformatics is the science of interpreting data from observations of biological process whose data is managed and mined [2]. Unlike data generated in various fields to support a hypothesis, the biological data is generated assuming that it contains vital information, and this information might answer several important questions. [3].

One of the most important applications of data mining is in the field of bioinformatics, because of its huge mass of data and hidden patterns particularly in proteomics data. The proteomic data consisting of sequence motifs in recurring patterns has the capability to predict a protein's structure and functionalities [8]. In order to identify sequence motifs, most algorithms need to specify a fixed size for the motif in advance. These algorithms deliver a similar number of motifs since they have a fixed size (1), include mismatches, or (2) are shifted by one base [5]. The problem of mismatches is addressed by showing that some groups of protein motifs occur in recurring patterns. The first problem implies that some group motifs may be similar to one another; the second problem probably can be more easily seen in this way: If there exists a biological sequence motif with length of 12 and we set the window size to 9, it is highly possible that we discovered two similar sequence motifs where one motif covers the front part of the biological sequence motif and the other one covers the rear part [8].

The Association Rule [1, 6, 7] is used to extract important information from large repositories of data. For example, association rules can discover the support and confidence of "if A occurs then B will occur." This can be expanded to any number of item sets whether it is three, four, or more. To put forth this kind of DNA/Protein bioinformatics data into Association Rules, each protein is regarded as a transaction and the sequence motifs as items in the transaction. Some of the papers that were referenced apply the Association Rule in this manner [1, 8]. Although Association Rule plays an important role in extracting recurring patterns from protein sequences, there is still one more criteria to be considered. The motifs in a protein occur in specific distance intervals, so it is vital to discover the distance between the occurrence of motifs A and B. Therefore, a new Positional Association Rule Algorithm is proposed in [8]. The Positional Association rule is simple extension of the Basic Association rule with a new parameter named "Distance Assurance".

It is proved that the fixed window size problem can be solved by generating clusters with the help of the Positional Association Rules Algorithm in [8]. In this paper, Structural Clustering Algorithm for Networks (SCAN) [10], a new clustering algorithm for networks, is applied to generate clusters from the Positional Association Rules. SCAN is a popular tool for analyzing graphs. SCAN's ultimate goal is to divide the nodes in the graph into three categories: clusters, hubs, and outliers. It creates clusters from structurally similar nodes [10]. For example, social networks may suggest a friend to you because you share similar friends with that person (i.e. you both belong to the same cluster). Nodes that belong to more than one cluster may bridge the two clusters together. SCAN identifies nodes of this pattern as hubs. Finally, SCAN marks structurally dissimilar nodes as outliers, which may be discarded as noise data [10].

In this paper, we propose that one can use SCAN to refine positional association rule results in order to increase the quality of the resulting clusters. We apply proposed approach to alleviate the first problem "include mismatches" caused by the fixed window size approach. The set of rules produced using the positional association rule are fed into SCAN, to generate clusters, outliers, and hubs. The outliers and hubs were discarded while the clusters were retained since the primary goal is to increase the quality of the clusters that SCAN revealed. Higher-quality SCAN clusters are verified with the quality of the positional association rule clusters.

The rest of the paper is organized into four more sections. Section II provides a detailed explanation of the algorithm.

Section III follows with details about the Experiment. Section IV shows the results of this work. Finally, the paper is concluded with Section V.

II. ALGORITHM

2.1 Positional Association Rules Algorithm

Algorithm: Positional Association Rule with the Apriori Concept
Input: Database, D, (Protein sequences as Transactions and Sequence Motifs as items), min_support, min_confidence, and min_distance_assurance
Output: P, positional association rules in D.
Method:

```

(1) L = find_frequent_itemsets(D, min_support)
(2) S = find_strong_association_rules(L, min_confidence)
(3) for (k=2; Sk ≠ ∅; k++)
(4)   for each strong association rule, r ∈ Sk
(5)     antecedent_motif = Apriori_Motif_Construct(r_ant)
(6)     consequent_motif = Apriori_Motif_Construct(r_con)
(7)     if antecedent_motif == NULL or consequent_motif == NULL:
(8)       goto Step (4)
(9)     for each protein sequence, ps ∈ D
(10)      for (ant_position=1; |ps|; ant_position++)
(11)        if antecedent_motif start appear on ps[ant_position]:
(12)          r_ant_count++
(13)          for (con_position=1; |ps|; con_position++)
(14)            if consequent_motif start appear on ps[con_position]:
(15)              distance = ant_position - con_position
(16)              r_distance++
(17)          Pk = { rant | rant > min_distance_assurance * r_ant_count }
```

Apriori_Motif_Construct(itemset)

```

(1) if |itemset| == 1:
(2)   return itemset
(3) else:
(4)   for each positional association rules in Pdistance
(5)     if all items in the itemset appear in the positional association rule:
(6)       return the new motif constructed by the positional association rule
(7)   return NULL
```

Figure 1 The Pseudocode of Positional Association Rule with the Apriori concept

The Association Rule in Data Mining generates item sets which occur frequently with certain rules occurring in a particular format, say $(X \Rightarrow Y)$ i.e. “if X occurs then Y occurs” with the condition that all of these item sets must pass a minimum support and confidence. A new Positional Association Rule, proposed in [8], has another parameter called “distance assurance.” The Positional Association Rule identifies a frequent item set with a certain frequent distance (d) and applies this distance once it obtains strong Association rules with a minimum confidence and minimum support. Where support and confidence is defined as:

$$Support(X \Rightarrow Y) = \frac{|X \cup Y|}{|T|}$$

$$Confidence(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|}$$

Where $|T|$ is the total number of transactions, $|X|$ is the number of transactions in T that contains at least one X , $|X \cup Y|$ is the number of the transactions in T that contain both X and Y . The newly proposed “distance assurance” is defined as:

$$Dis.Assurance(X \Rightarrow Y) = \frac{\|X \overset{d}{\cup} Y\|}{\|X\|}$$

Where $\|X\|$ is the total number of times that X appears in T , d indicates the distance, $-\infty < d < \infty$. Where $X \overset{d}{\Rightarrow} Y$ denotes “if X appears, then after the distance of d , Y appears,” $\|X \overset{d}{\cup} Y\|$ is the total number of times in T that when X occurs and after the distance of d , Y occurs. Figure 1 shows pseudo code for the Positional Association Rule Algorithm and a detailed description is available in [8].

2.2 SCAN Algorithm

SCAN is short for Structural Clustering Algorithm for Networks. While many algorithms find just the clusters in a network, SCAN finds the hubs and outliers. The identification of hubs is the real strength of SCAN, as hubs bridge clusters, and spread its influence from cluster to cluster. The usefulness on identifying outliers on the other hand, is simply in knowing that the outliers can be ignored. Outliers have little influence on their connected cluster, or on the cluster’s network.

SCAN works by looking at the neighborhood of vertices instead of only their direct connections. This allows the detection of hubs and outliers. Not only is the algorithm useful, but it is also efficient with a running time of $O(n)$.

When running SCAN, the algorithm labels a newly found vertex as unclassified. From here it checks to see if this vertex has a minimum amount of connections in a cluster. If so, it uses this new found core as a springboard to search for more vertices. Finally, once SCAN visits all vertices, it identifies the vertices that connect to two or more clusters as hubs, and vertices that connect to only one cluster as outliers. The more connections a vertex has to a cluster, the more influence that vertex has on the cluster.

2.3 The combination of Positional Association Rules algorithm and SCAN Algorithm

In this paper, in order to alleviate the first problem “include mismatches” caused by the fixed window size approach, we combine the positional association rules algorithm with SCAN to identify protein sequence motifs that similar to each other. First of all, positional association rule algorithm with distance equals to zero is implemented to identify protein sequence motifs that occur on the same position. The rationale behind this is that if two (or more) motifs occur on the same position frequently enough (pass the minimum distance assurance), they should be similar to one another. As the result, the network-like graph such as Figure 2 is generated. Next, the associations were converted into two columns of data for input into SCAN (as showed in *Results*). The data was used to run SCAN multiple times for each distance assurance with different values of μ and ϵ . Finally, clusters are generated by the proposed approach, which combines the positional association rules algorithm and SCAN. Secondary structure information is taken and analyzed the quality of each SCAN-generated cluster. Detail results with different parameters are available in *Results*.

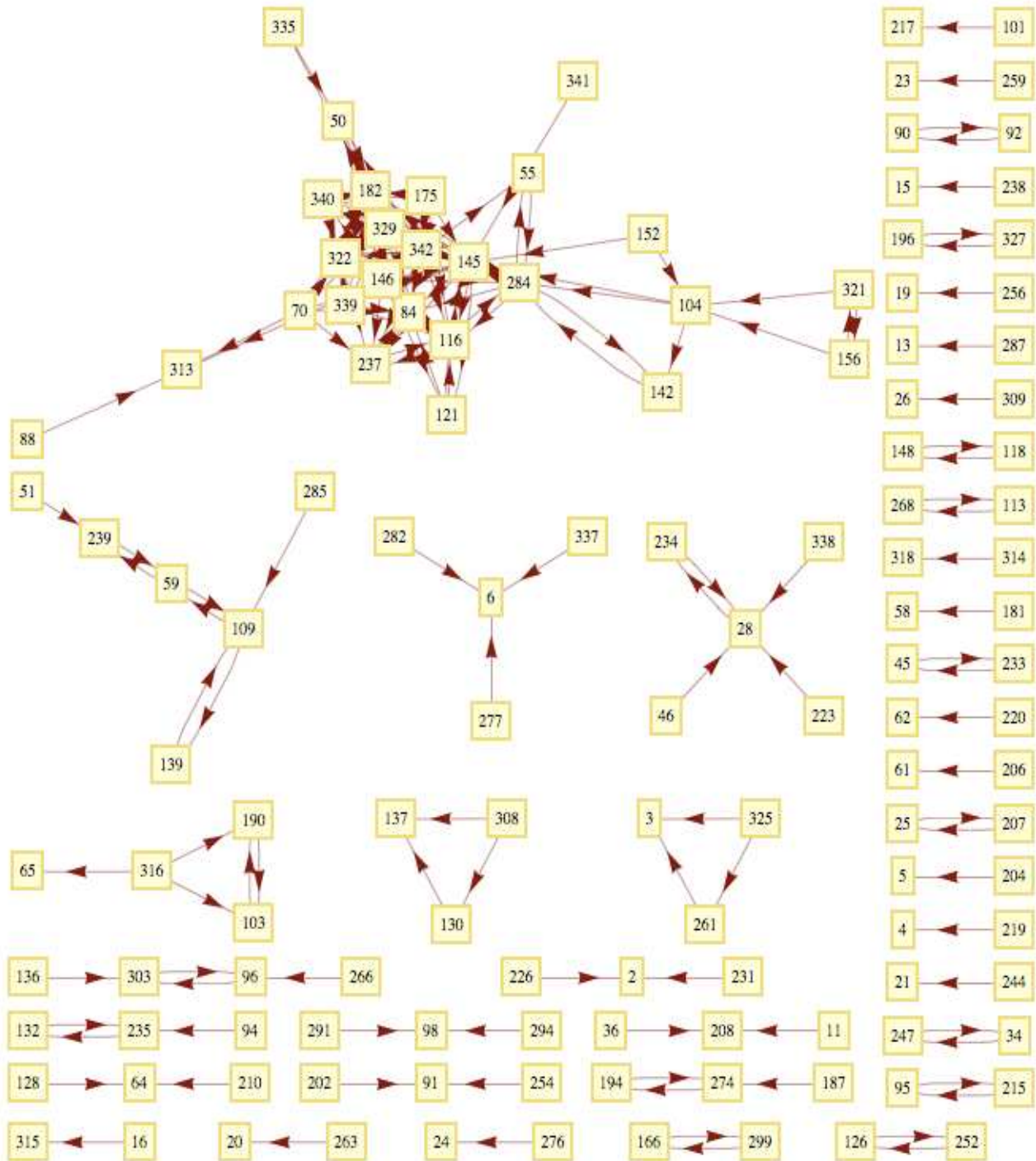


Figure 2: Directed graph generated from positional association rules based on minimum support, confidence, and distance assurance equal to 20%, 70% and 50% respectively.

III. EXPERIMENT AND PARAMETERS SETUP

3.1 Dataset

The original dataset used in this work includes 2710 protein sequences obtained from Protein Sequence Culling Server (PISCES) [11]. It is the dataset that was used in [8,12] to generate protein sequence motifs. No sequence in this database shares more than 25% sequence identity. The frequency profile from the HSP [13] is constructed based on

the alignment of each protein sequence from the protein data bank (PDB) where all the sequences are considered homologous in the sequence database. For the frequency profiles (HSP) representation for sequence segments, each position of the matrix represents the frequency for a specified amino acid residue in a sequence position for the multiple sequence alignment. Twenty rows represent 20 amino acids and 9 columns represent each position of the sliding window. Secondary structure was also obtained from DSSP [14], which

is a database of secondary structure assignments for all protein entries in the Protein Data Bank, for evaluation purposes. DSSP originally assigns the secondary structure to eight different classes. According to previous related research [12, 15], those eight classes were converted into three based on the following method: H, G and I to H (Helices); B and E to E (Sheets); all others to C (Coils). 343 different sequence motifs with window size of nine generated from previous work [12] are included in this paper. The dataset actually used in this work comes from [8] and contains more than 2000 protein sequence as transactions vary in amount of motifs (items). Each transaction sequence is sorted and organized by distance value, the items on the same line having a distance of zero from one another. The secondary structure data contained nine values for each 343 motifs, each value corresponding to its H, E, or C secondary structure percentage.

3.2 Positional Association Rule

The protein sequences are treated as transactions and the sequence motifs are treated as items of the transaction. Firstly, the association rules are generated from the data. As we mentioned in section 2.1, only tradition association rules are not sufficient due to the protein motifs occurring at positions. "Distance assurance" measure is incorporated. In this paper only a distance measure of zero is taken into account which means the protein sequence motifs which occur at same positions are considered.

3.3 Running SCAN for refining clusters

The SCAN proposed in [10] is used to generate clusters from the rules generated as described in section 3.2. When a member of the generated clusters is identical to a neighboring cluster, their combined structure will add up to a bigger cluster. So, the number of common neighbors is normalized by the geometric mean of the two neighborhood sizes.

$$\sigma(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{|\Gamma(v)| |\Gamma(w)|}}$$

where, $\Gamma(v)$ and $\Gamma(w)$ denotes the neighborhood of v and w respectively. When assigning a member to a cluster a threshold ε is applied to the computed structural similarity. Also μ number of neighbors with a structural similarity and exceeding the neighborhood threshold ε is required to decide whether a vertex is a core.

The values of ε and μ are varied to generate various clustered files. The ε is varied from 0 to 0.5 and μ is varied between 1 and 2 only although various values of μ has been used they, all proved to be ineffective.

3.4 Dissimilarity Measure

The following formula is used to calculate the dissimilarity between two sequence segments:

$$\text{Dissimilarity} = \sum_{i=1}^L \sum_{j=1}^N |F_k(i, j) - F_c(i, j)|$$

Where L is the window size and N is 20 which represent 20 different amino acids. $F_k(i, j)$ is the value of the matrix at row i and column j used to represent the sequence segment. $F_c(i, j)$ is

the value of the matrix at row i and column j used to represent the centroid of a give sequence cluster. The lower dissimilarity value is, the higher similarity two segments have.

3.5 Structural Similarity Measure

Cluster's average structure is calculated using the following formula:

$$\sum_{i=1}^{ws} \max(P_{i,H}, P_{i,E}, P_{i,C})$$


ws

Where ws is the window size and $P_{i,H}$ shows the frequency of occurrence of helix among the segments for the cluster in position i . $P_{i,E}$ and $P_{i,C}$ are defined in a similar way. If the structural homology for a cluster exceeds 70%, the cluster can be considered structurally identical [13]. If the structural homology for the cluster exceeds 60% and lower than 70%, the cluster can be considered weakly structurally homologous [15].

IV. RESULTS

The positional association rule runs six times with distance assurance values of 10%, 20%, 30%, 40%, 50%, and 60%; while the minimum support and confidence is set as 20% and 70% based on the optimal parameter setup of previous work [1]. Once complete, the file was translated into a two-column format representing the associations. For example, $A \rightarrow B$ would become line "A B." The two column files were then fed into SCAN. An example is given in Figure 3 with minimum distance assurance equals to zero.

A	B	Distance Assurance
226	2	--> 68.1172 (814)
2	226	--> 33.388 (814)
1	2	--> 21.2466 (392)
2	1	--> 16.0788 (392)
6	2	--> 20.6297 (249)
2	6	--> 10.2133 (249)
62	2	--> 23.615 (341)
2	62	--> 13.9869 (341)



226	2
2	226
1	2
2	1
2	6
6	2
2	62
62	2
2	62
62	2

Figure 3: Conversion of the Positional Association Rules output to SCAN input

However, besides the data, SCAN requires two other parameters: ε and μ . μ is varied between 0 and 3 with step-size of 1. ε is between 0 and 1 to generate various clustering files and optimum clustered data is chosen. In the first run of SCAN, some limitations on the parameters were determined. First, μ seems to only be effective at values 1 or 2. A value of zero results in all clusters and no outliers, a value higher than two results in all outliers and no clusters. SCAN produces hubs with values of ε greater than 0.5, so ε was restricted to lower values.

Hubs were determined to be an undesirable component in this research because they were not included with the clusters. This caused isolation of major cluster components. For example, Figure 4 shows four motifs that should belong to the same cluster. If ε was set too high, Motif #6 would be

classified as a hub, removing it from the cluster. Since 282, 337, and 277 are not associated with any other motifs, they are removed as outliers.

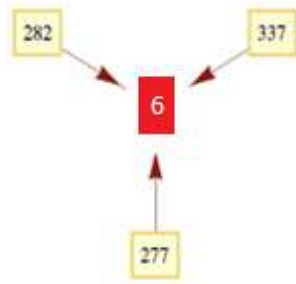


Figure 4: A Cluster with a Potential Hub

In the end, SCAN was run with distance assurance between 10% and 60%, M between 1 and 2, and E between 0.1 and 0.5. To ease the process of running SCAN on all of these parameter combinations, a script was created to run them in batch. The SCAN algorithm is a pre-packaged Java application. the algorithm was called with the appropriate combination of parameters and it gave the output files containing the clusters, hubs, and outliers obtained from the association rule data.

Next, a second script was ran, which fed each SCAN output file into the quality algorithm. The quality algorithm implements the Structural Similarity Measure discussed in section 3.5. The algorithm takes the SCAN output file and file containing motif structure information as parameters. Once complete, the algorithm produced an output file containing a percentage on each line representing a cluster's quality. Finally, a third, simple script was run to summarize the quality results and place them into range groups including >80, 70-80, 60-70, and <60. An example summary is shown in Figure 5.

DA	Mu	EPS	Cluster #	Quality	Q80	Q70	Q60	Qlow
50	1	0.1	1	0.742129	0	1	0	0
50	1	0.1	2	0.714095	0	1	0	0
50	1	0.1	3	0.615433	0	0	1	0
50	1	0.1	4	0.694561	0	0	1	0
50	1	0.1	5	0.739924	0	1	0	0
50	1	0.1	6	0.694304	0	0	1	0

Figure 5: Sample Quality Summary

Initially, all of the summary files were combined to determine which parameters gave the best results. The most favorable combination was a distance assurance of 50%, M of 1, and E of 0.3. Distance assurance had the most significant impact on cluster quality. ϵ , as shown in Figure 6, has little or no effect on quality.

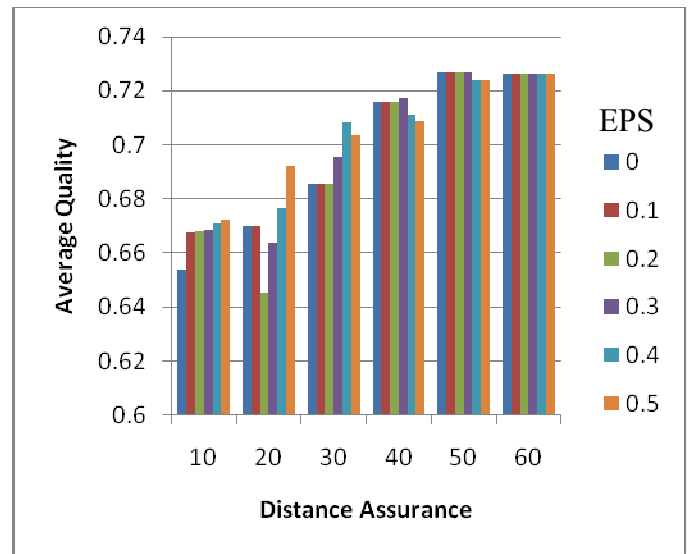


Figure 6: Dist. Assurance & ϵ Quality, M = 1

M has a slight effect on quality, but still does not compare to distance assurance. Figure 7 shows M's effect.

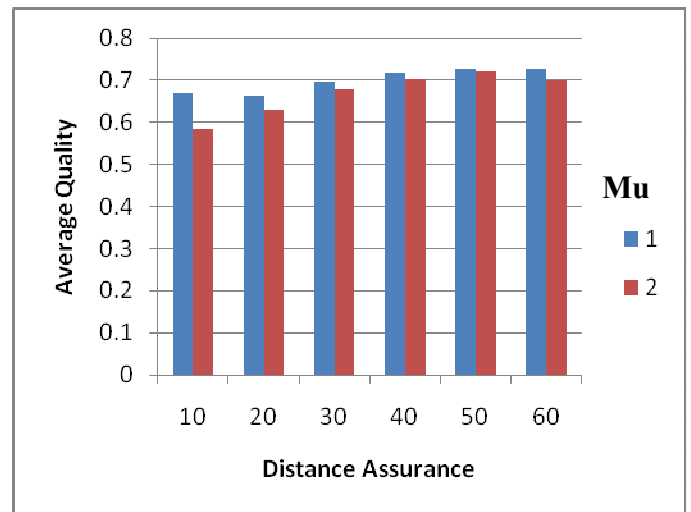


Figure 7: Dist. Assurance & μ 's Quality, EPS = 0.3

With these new findings, it can be concluded that the SCAN parameters ϵ and μ have little effect on cluster quality as long as they stay within the range tested above. Distance assurance's effect on the result demonstrates the impact of the positional association rule analysis method. Running SCAN on the data provided two important pieces of information: clusters and outliers. Each cluster provides a graph structure containing the original associations. This allows observations to be made on groups of associations rather than one at a time. The outliers remove noise, or associations of little significance. The positional association rule does a good job of eliminating outliers based on occurrence statistics, but SCAN takes it a step further and analyzes relationship structures.

V. CONCLUSION

For data mining in the field of bioinformatics, the ability to find recurring patterns in proteomics data enables the discovery of a protein's structure and functionality. Most enumerative algorithms require the size of the motif to be set in advance. This can cause errors such as mismatches and bases that are off by one. However, the *Positional Association Rule* can be used as a remedy to these problems through the use of a distance assurance.

It is known that Association Rules can already be well used in Classification techniques, and *Chen et al.* [1] proved that it can also be used for Clustering purposes. In this paper, we further combine the positional association rules algorithm with the SCAN algorithm. With the SCAN data sorted, concentration solely on the clusters further increased the cluster quality.

ACKNOWLEDGMENT

The work of Bernard Chen was supported in part by UCA's University Research Council (URC). The work of Sinan Cockara was supported in part by UCA's University Research Council (URC).

REFERENCES

- [1] Bernard Chen, Michael Miller, Timothy Montgomery, Terrance Griffin, "Clustering Using Positional Association Rules Algorithm on Protein Sequence Motifs", International Conference on Bioinformatics & Computational Biology (BIOCAMP2010), Las Vegas, USA, pp.75~80.
- [2] Lonardi Stefano, Chen Jake, "Biological Data Mining": Chapman and Hall/CRC 2010 Computational Biology and Bioinformatics, IEEE/ACM Transactions on April-June 2010.
- [3] Arno Siebes, V. Hlavac, K.G.Jeffery and J. wiedermann (Eds): SOFSEM 2000, LNCS 1963, PP.54-55, 2000@ springer –Verlag Berlin Heidelberg 2000.
- [4] Interestingness of Discovered Association Rules in terms of Neighborhood-Based Unexpectedness (1998), Guozhu Dong, Jinyan Li.
- [5] Ohler u. & Niemann, H. (2001), Identification and analysis of eukaryotic promoters: Recent Computational Approaches, Trends in Genetics. 17,56-60.
- [6] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", VLDB 1994.
- [7] R. Agrawal, T. Imielinski and A. Swami, "Mining Associations between Sets of Items in Large Databases", *ACM SIGMOD Int'l Conf. on Management of Data*, Washington D.C., May 1993.
- [8] Bernard Chen, and Sinan Kockara, "Mining Positional Association Super-Rules on Fixed-Size Protein Sequence motifs", *IEEE BIBE 2009, Taichung, Taiwan*, proceeding pp. 1-8
- [9] Haoudi, Abdelali; Bensmail, Halima "Bioinformatics and data mining in proteomics" Expert Review of Proteomics, Volume 3, Number 3, June 2006 , pp. 333-343(11)
- [10] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, Thomas A. J. Schweiger, "SCAN: A Structural Clustering Algorithm for Networks", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, August 12-15, 2007, San Jose, California, USA
- [11] Wang, G. & Dunbrack, R. L. (2003) PISCES: A Protein Sequence Culling Server in Bioinformatics pp. 1589-1591, Oxford Univ Press.
- [12] Chen, B., Tai, P. C., Harrison, R & Pan, Y.(2006) FGK Model : An Efficient Granular Computing Model for Protein Sequence Motifs Information Discovery, Iasted Proc. International Conference on Computational and Systems Biology (CASB), Dallas.
- [13] Sader, C. & Schneider, R. (1991) Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment, *Proteins: Structure, Function & Genetics.* 9, 56-68.
- [14] Kabsch, W. & Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers.* 22, 2577-2637.
- [15] Zhong W., Altun G., Harrison R., Tai P. C. and Pan YI, (2005) Improved Kmeans Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property, *IEEE Trans. On Nanobioscience*, Vol 4, pp. 255-265.

Constructing Super Rule Tree (SRT) for Protein Motif Clusters Using DBSCAN

Bernard Chen, Sait Suer, Muhyeddin Ercan, Rahul Tada, Recep Avci, and Sinan Kockara
University of Central Arkansas, Department of Computer Science, USA

Abstract—

Searching for protein sequence and structural motifs is one of the most important topics in Bioinformatics, because the motifs are able to determine the role of the proteins. A fixed window size is usually defined in advance for the most of motif searching algorithms. The fixed window size may result in generating a number of similar motifs shifted by one to several bases or including mismatches. In this study, to confront the mismatched motifs problem, we use the super-rule concept to construct a Super-Rule-Tree (SRT) which is generated by the DBSCAN clustering algorithm. This SRT recognizes the similar motifs. Analysis of the hierarchical DBSCAN generated Super-Rule-Tree shows a better quality in secondary structure similarity evaluation than the previous studies'. We believe that the combination of DBSCAN and SRT concept may provide a new point of view to similar researches which require predefined fixed window size.

Keywords: **Super-Rule-Tree (SRT), DBSCAN, protein sequence motif.**

1. INTRODUCTION

ALL living organisms require proteins to maintain chemical and physical activities. Proteins are made of 20 types of amino acids [1]. Each protein has its own unique structure and function depending on the sequence and the type of its amino acids. From the point of view of biology and bioinformatics, to reveal the functionality of a protein, it is necessary to obtain the structure of the protein. Hence, an understanding of the formation of amino acids that synthesize the protein is crucial. Analyzing the sequence of amino acids yields some sequence patterns called motifs which have biological significance and repeat frequently. One of the most important Bioinformatics research fields in sequence analysis is searching for motifs, since these recurring patterns have the potential to determine a protein's conformation, function and activities [2].

Proteins are usually grouped based on their structural similarities in order to determine their functional properties. Therefore, to group the proteins, clustering of motif sequences is important. Just like proteins, discovered protein sequence motifs are usually categorized into protein families; PROSITE [3], PRINTS [4], and BLOCKS [5] are three most popular motifs databases that follows this trend. Since sequence motifs from PROSITE, PRINTS, and BLOCKS are developed from multiple alignments, these sequence motifs only search for conserved elements of sequence alignment from the same protein family and carry little information about conserved sequence regions, which transcend protein families [6].

In order to obtain protein sequence motifs which transcend protein family boundaries, we applied our Super GSVM-FE model on all of our information granules so that we obtained

541 extracted high-quality protein sequence motifs in our previous work [7]. However, the most challenging factors of identifying the motifs by clustering them appropriately emerge from the ambiguity and the variability of their sizes. Therefore, a pre-determined size is mostly used in the motif researches. However, two major problems stem from this fixed size namely; mismatches and shifted by one base [8]. The first problem can be simply expressed as the probable similarity of two or more motif groups. The second problem 'shifted by one base' causes to identify one motif more than once as if they are two or more different motifs. For example, if a biological sequence is longer than the fixed size, it is possible to identify the front part and the rear part as two different motifs. In this paper, we try to solve 'grouping similar motifs including mismatches' problem by using super-rules concept [9]. This problem previously was dealt in [2]. In their study, they made an improvement of the HHK Clustering Algorithm [2] and by using the super-rules concept they clustered the motifs and found the similarities among them in the form of a Super-Rule-Tree (SRT).

In this paper; however, we worked out the first problem by using famous clustering algorithm so called Density Based Spatial Clustering of Applications with Noise (DBSCAN) [10] in order to acquire more accurate results. We worked on 541 high-quality protein sequence motifs extracted by Super GSVM-FE model [7]. Then we applied the DBSCAN algorithm on these motifs at different levels of hierarchy to obtain the ideal SRT. DBSCAN algorithm requires two parameters called 'Eps (epsilon)' [10] and 'MinPts (minimum points)' [10]. Eps is the maximum radius of the neighborhood which is to be examined to form a cluster and MinPts is the minimum number of elements required to form a cluster. We applied DBSCAN for all possible values of epsilon and minPts and plotted different graphs taking into consideration minPts, epsilon, number of outliers, number of clusters, and comparatively size of clusters to choose the best pair of parameters. A comprehensive quality comparison of our new Super-Rule-Tree (SRT) with the one in the previous study [2] is also presented.

The remainder of the paper is organized as follows. Section 2 describes the DBSCAN and Super Rule Tree (SRT). Section 3 discusses how we setup the experiment with the DBSCAN and an explanation for determination of parameters. The SRT with comparisons and conclusions are given in section 4 and section 5.

2. METHODOLOGY

2.1 DBSCAN

Density Based Spatial Clustering of Applications (DBSCAN) with Noise is a notable clustering algorithm. It requires two parameters namely Eps and MinPts. Important terms and their definitions are listed below.

- a) *Eps*: Maximum radius of the neighborhood to be considered while forming clusters.
- b) *MinPts*: Minimum number of points required to form a cluster.
- c) *Eps-neighborhood* [10]: A point q is said to be in the Eps-neighborhood of the point p , if the distance between p and q is less than or equal to Eps.
- d) *Core points and Border points* [10]: Points inside the cluster are called core points and points on the border of the cluster are called border points.
- e) *Directly density-reachable* [10]: A point q is directly density-reachable from a point p w.r.t Eps and MinPts, if q belongs to the Eps-neighborhood of p and the number of points in the Eps-neighborhood of p is greater than or equal to MinPts (see Figure 2.1). If p and q are core points, then directly density-reachable is symmetric i.e., p is directly density-reachable from q and vice versa. However, this condition fails if either p or q is a border point.
- f) *Density-reachable* [10]: A point p is density-reachable from a point q w.r.t Eps and MinPts, if there exists a set of points between q and p such that every point in this set is directly density-reachable from its precede.
- g) *Density-connected* [10]: If there exists a point x such that the points, p and q are both density-reachable from x , then p is said to be density-connected to q w.r.t Eps and MinPts.
- h) *Noise*: Noise is a set of points in a database that does not belong to any cluster. These points are also called as *outliers*.

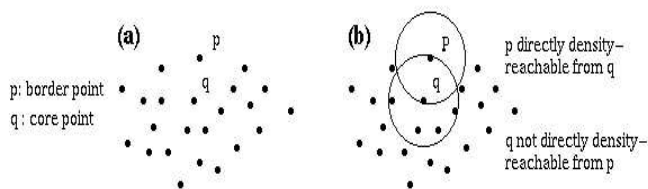


Figure 2.1: DBSCAN application on a 2D data set [10]

This clustering algorithm follows the procedure of finding all points density-reachable from an arbitrary starting point, depending on the Eps and MinPts. If the starting point is a core point then the procedure begins building a cluster. On the other hand, if it is a border point the algorithm cannot go further, i.e., it cannot find any point density-reachable from the starting point. This procedure is followed until all of the points in the Eps-neighborhood are touched or visited at least

once. After all of the points in a cluster are visited, the algorithm chooses a new arbitrary starting point to generate other clusters.

For the given example in Figure 2.1, it is not complicated to find the range of parameters and it is not difficult to visualize the data so that the parameters can be determined by starting from 0 to the extreme value, i.e. the distance between the farthest elements. However, in our case, the elements (points) have 180 dimensions or attributes; so, it is difficult to visualize a data in 180 dimensions and challenging to determine the ideal parameters as well as determining a range for parameters. Thus, for Eps, we started from 0 in which every element was found as an outlier. Then we use brute-force approach to reach a point where all the elements form just a single cluster. This approach helped us to find the extreme values for parameters. We further investigated to find the best parameters. Parameters are considered the best possible when the cluster to outlier ratio becomes maximum. This is explained in section 4 with details. ‘Manhattan Distance’ was used as a distance measure which is the sum of absolute differences between attributes of two elements.

2.2 Super Rule Tree (SRT):

The data set contains 541 motifs, in which each motif has some rules. DBSCAN was used to cluster these motifs based on similarity and then assemble the rules in each motif to generate super rules. Once the rules are generated, it is possible to form another layer of super rules (super-super rules). By this manner, a tree like structure (Super-Rules-Tree structure) is formed using these super rules. These super rules represent a harmonic rule pattern and the essential underlying relationship of classification [9]. Because the super-rules are generated from each of the motifs, it is easy to understand the general trend and ignore the noise and also interactively focus on the important aspects of the domain by using super-rules and selectively view the original detail rules in the corresponding motif [9].

3. EXPERIMENTAL SETUP

3.1 Data set:

The original data set including 2710 protein sequences had been obtained from Protein Sequence Culling Server (PISCES) by Wang and Dunbrack [11]. This data set was used in [2] and [7] to generate protein sequence motifs. No sequence in this database shares more than a 25 per cent sequence identity. We also obtained the secondary structure from DSSP [12] which is a database of secondary structure assignments for all protein entries in PDB. In this database there are 8 different classes of for secondary structures. Chen et al. replaced those 8 classes with 3 classes by assigning H, G and I to H (Helices); B and E to E (Sheets); and all others to C (Coils).

541 different sequence motifs were generated in [7] with a window size of nine from the original data set. Each window is represented by a 9x20 matrix plus additional nine corresponding representative secondary structure information

and it corresponds to a sequence segment. Twenty amino acids are represented by 20 columns and each position of the sliding window is represented by 9 rows. Chen et al. has obtained 541 high quality clusters extracted by super GVSM-SE model and each cluster is represented in 180 dimensions in the first data set. In this study, the 541 clusters obtained from [7] have been used as the data set. In addition to these clusters, the data set which includes the secondary structure of these clusters have also been used.

3.2 Dissimilarity Measure

In this paper Manhattan distance has been used as the dissimilarity measure. Manhattan distance indicates a grid-like path while traveling from one point to another. It is also known as the city block metric. According to Zhong et al. [6], this dissimilarity measure is more suitable for this field of study since all positions of the frequency profile are considered equal.

The Manhattan Distance for the data set is calculated by the following formula:

$$\text{Dissimilarity} = \sum_{i=1}^L \sum_{j=1}^N |F_k(i, j) - F_c(i, j)|$$

Where L is the window size and N is 20 representing 20 different amino acids. $F_k(i, j)$ is the value of the matrix at row i and column j and represents the sequence segment. $F_c(i, j)$ is the value of the matrix at row i and column j and represents the centroids of a give sequence cluster. The lower the dissimilarity value, the higher similarity the two segments have.

3.3 Structure Similarity Measure

In order to get the secondary structure and measure the quality of each cluster the following formula has been used.

$$\text{Secondary structural similarity} = \frac{\sum_{i=1}^{ws} \max(P_{i,H}, P_{i,E}, P_{i,C})}{ws}$$

Where ws is the window size, C_i , E_i and H_i correspond to the frequency of Coils, Sheets and Helices respectively and $P_{i,H}$ shows the frequency of occurrence of helix among the segments for the cluster in position i . $P_{i,E}$ and $P_{i,C}$ are defined in a similar way.

If the structural homology for a cluster exceeds 70%, the cluster can be considered structurally identical. If the structural homology for the cluster exceeds 60% and is lower than 70%, the cluster can be considered weakly structurally homologous [6].

3.4 Cluster-Outlier Ratio

A ratio has been used as a criterion to find the ideal parameters Eps and $MinPts$ for DBSCAN. The ratio is calculated by using the following formula:

$$\text{Cluster_Outlier_ratio} = \text{num_cluster} / \text{num_outliers}$$

As the ratio increases, the optimum parameters are obtained. However, this ratio is considered in an interval where number of outliers does not equal to zero or the number of elements.

4. EXPERIMENTAL RESULTS

4.1 Determination of Eps and $MinPts$ for Super-Rule-Tree (SRT) construction

Clusters are formed by applying the DBSCAN algorithm on the original data set. But, before that, the most important issue is to determine the values of Eps and $MinPts$. To determine a logical Eps and $MinPts$ value, the DBSCAN is applied on the original data with Eps ranging from 100 to 500 and $MinPts$ ranging from 2 to 7. These possible parameter pairs were chosen in this range because beyond these boundaries the algorithm accumulates all elements into one cluster or it determines all the elements as outliers. Graphs were plotted for all the values of Eps and $MinPts$ based on the number of clusters formed and the number of outliers. Since the logical Eps and $MinPts$ cannot be determined based on the mentioned criteria, the Cluster-Outlier ratio has been used. This ratio was compared for each Eps and $MinPts$ value within the range and determined its maximum values so that the number of clusters is higher and the number of outliers is less. After graphs were plotted based on different parameters it was determined that the appropriate $MinPts$ value is 2, otherwise the number of clusters declines significantly as shown in the figures below.

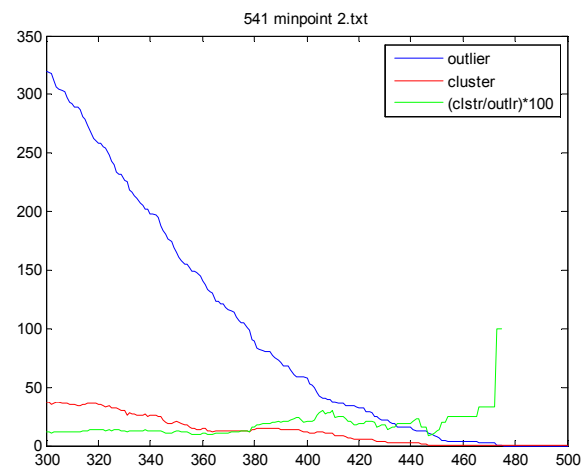


Figure 4.1: Graph for 541 clusters with $MinPts=2$

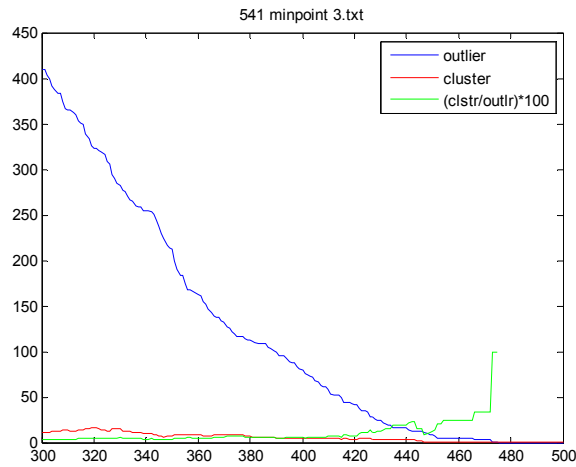


Figure 4.2: Graph for 541 elements with MinPts=3

The x-axis represents Eps. In Figure 4.1, it is revealed that at Eps =406, the clusters to outliers ratio is maximum and at the same time the number of clusters is reasonably high (greater than 1). In Figure 4.2, the ratio of cluster to outlier decreases significantly. A similar trend is observed for MinPts greater than 3, so the parameters are Eps =406 and MinPts =2 for 541 clusters.

4.2 Applying DBSCAN on the sub clusters

As the DBSCAN is applied with Eps=406 and MinPts=2, 12 sub clusters have been found, where the first sub clusters holds 463 elements i.e. 85 percent of the total elements accumulated in one sub cluster. Therefore, we believe it is necessary to cluster these 463 elements and form SRT structure.

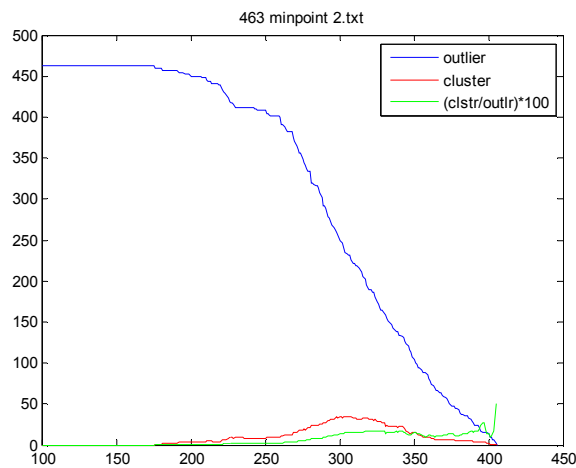


Figure 4.3: Graph for 463 clusters with MinPts=2

In order to apply DBSCAN on this sub cluster we followed the same procedure to determine the Eps and MinPts. From

Figure 4.3, the optimum Eps value was empirically found to be 396 and MinPts 2. DBSCAN was applied with these parameters and found 4 sub clusters, where the first sub cluster holds 438 elements, which is majority of the data. Needless to say, we cluster these elements via DBSCAN again.

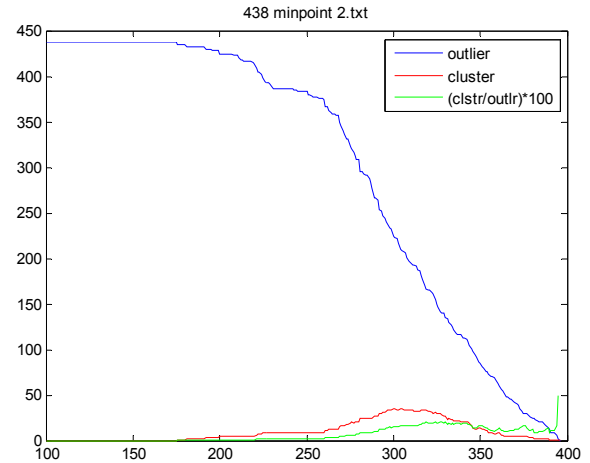


Figure 4.4: graph for 438 elements with MinPts = 2

After the determination procedure was followed for Eps and MinPts and their values are found to be 327 and 2 respectively (as shown in Figure 4.4). The DBSCAN was applied with these parameters and found 29 sub clusters with the first sub cluster holding 126 elements. We stopped further clustering after level 4 (the parameters are determined through figure 4.5 with Eps=319 and MinPts=2) because there is no sub-clusters with more than 100 elements after that. Figure 4.6 shows the multi-layered DBSCAN generated SRT structure, with all the super rules in each motif at each level of DBSCAN application.

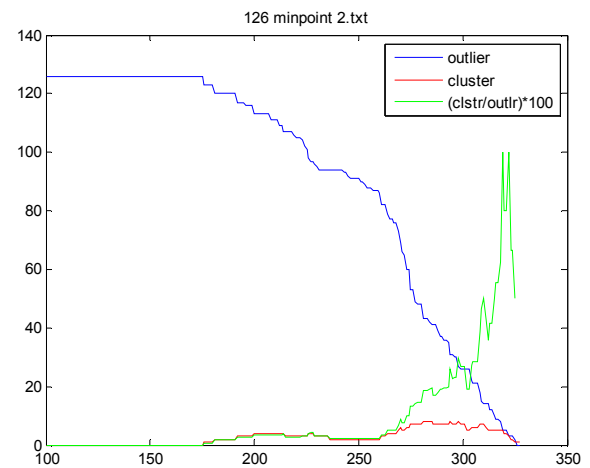


Figure 4.5: Graph for 126 elements with MinPts=2

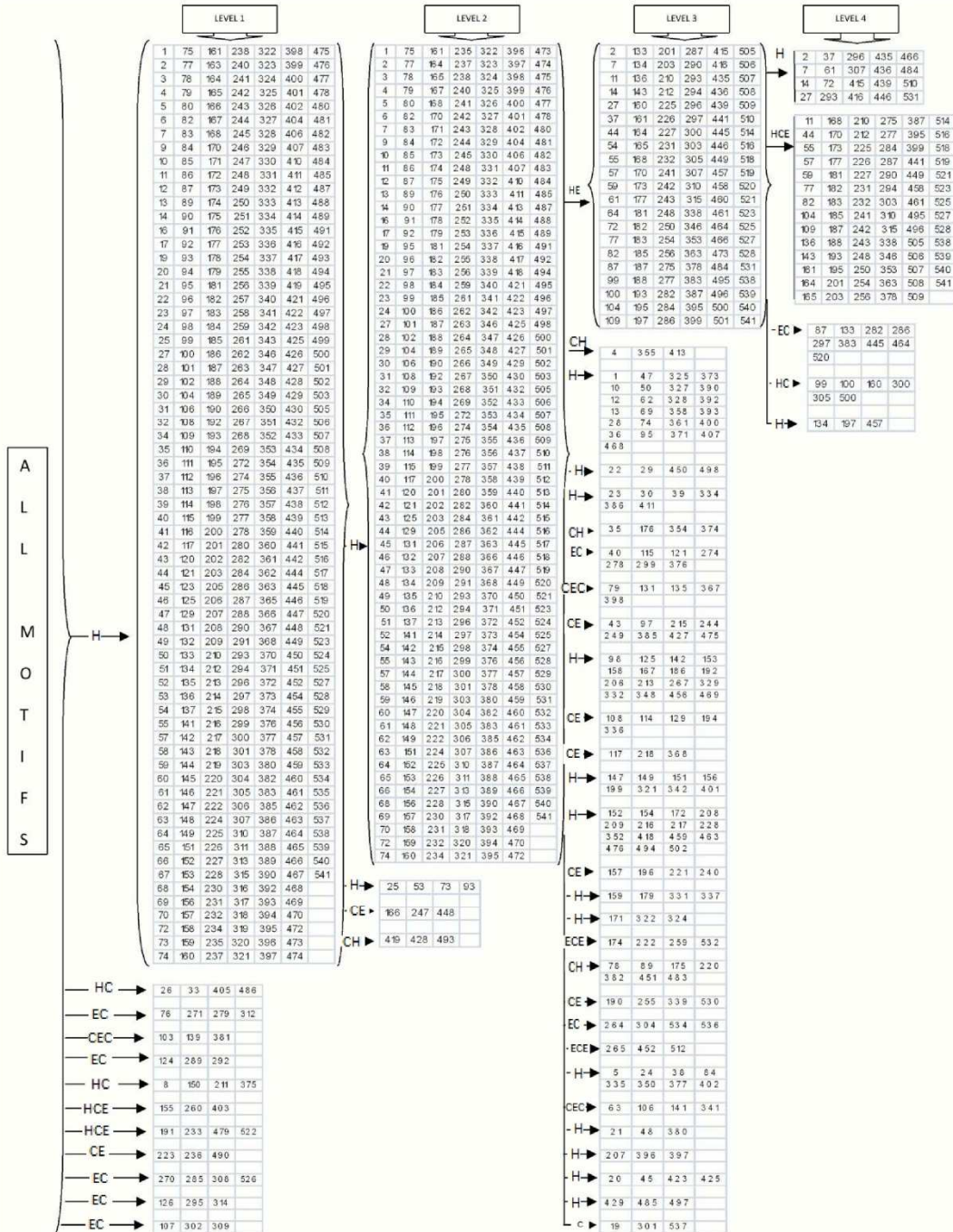


Figure 4.6: A Super Rule Tree with 4 levels of hierarchy

4.2 Super-Rule-Tree comparison

The Super-Rule-Tree generated in this paper is based on the top-down approach; while the Super-Rule-Tree made in [2] is based on the bottom-up method. The major reason causes the difference is according to the number of clusters generated from the clustering algorithms. In [2], the HHK clustering requires no parameters and generates high number of clusters. For example, the HHK clustering algorithm generates 108 clusters when it is applied on 541 protein sequence motifs. Due to the fact that the number of clusters is too large to handle, another level of clustering is applied; thus, a Super-Rule-Tree is formed to have a more generalized view. On the contrary, DBSCAN generates 12 clusters when it is applied on 541 protein sequence motifs with first cluster contain over 85% protein sequence motifs. Clearly, it is necessary to apply DBSCAN on the first cluster. Therefore, a Super-Rule-Tree is formed to have a more specialized view.

“Which SRT is better?” In order to answer this question, we evaluate the SRT level by level using secondary structural similarity. Table 4.1 demonstrates the average cluster quality for each level. Level 1 indicates the first clustering results applied on original 541 protein sequence patterns. Level 2 demonstrates the clustering results on the next level. Since the SRT in [2] contains only 2 levels, we can not compare both Super-Rule-Trees directly. However, it is clear to see that the SRT constructed in this paper is better than the previous works in secondary structure point of view. This mainly because the DBSCAN has the ability to filter out several outliers by setting up Eps and MinPts; while the HHK clustering algorithm can not sieve out outliers because it is a non-parameter approach.

Table 4.1 Secondary structure similarity evaluations on SRT level by level

Average Cluster Quality	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
SRT in this paper	75.98%	76.69%	75.64%	73.50%
SRT in [2]	69.02%	63.48%	NA	NA

5. CONCLUSION

In this paper, we propose that DBSCAN can be utilized to form the Super-Rule-Tree structure. We demonstrate a detailed process and a high quality Super-Rule-Tree, which gives a clear big picture of relations between protein sequence motifs. The improved secondary structure similarity on the SRT provides a better insight of the discovered protein sequence motifs that transcend protein family boundaries. We believe many further researches can be derived from this work.

REFERENCES

- [1] Fan, K. and Wang, W. (2003) ‘What is the minimum number of letters required to fold a protein?’, *J. Mol. Biol.*, 328, 921-926.
- [2] Bernard Chen, Jieyue He, Stephen Pellicer and Yi Pan. (2010) ‘Using Hybrid Hierarchical K-means Clustering Algorithm for Protein Sequence Motif Super-Rule-Tree (SRT) Structure Construction’, *International Journal of Data Mining and Bioinformatics (SCI indexed)*, Volume 4 - Issue 3, pp. 316-330.
- [3] N. Hulo, C. J. A. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch, (2004) ‘Recent improvements to the PROSITE database,’ *Nucleic Acids Res.*, vol. 32, Database issue: D134-137, 2004
- [4] T. K. Attwood, M. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N. Naudling, L. McGregor, A. Mitchell, G. Moulton, K. Paine, and P. Scordis, (2002) ‘PRINTS and PRINTS-S shed light on protein ancestry,’ *Nucleic Acid Res.* vol. 30, no. 1, pp. 239-241.
- [5] S. Henikoff, J. G. Henikoff and S. Pietrokovski, (1999) ‘Blocks+: a non redundant database of protein Alignment blocks derived from multiple compilation,’ *Bioinformatics*, vol. 15, no. 6, pp. 417-479
- [6] Zhong, W., Altun, G., Harrison, R., Tai, P.C. and Pan, Y. (2005) ‘Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property’, *NanoBioscience, IEEE Transactions on*, Vol. 4, pp.255–265.
- [7] Chen, B., Pellicer, S., Tai, P.C., Harrison, R. and Pan, Y. (2008) ‘Efficient super granular SVM feature elimination (Super GSVM-FE) model for protein sequence motif information extraction’, *Int. J. Functional Informatics and Personalized Medicine*, Vol. 1, pp.8–25.
- [8] Ohler, U. and Niemann, H. (2001) ‘Identification and analysis of eukaryotic promoters: recent computational approaches’, *Trends in Genetics*, Vol. 17, pp.56–60.
- [9] He, J., Chen, B., Hu, H.J., Harrison, R., Tai, P.C., Dong, Y. and Pan, Y. (2005) ‘Rule clustering and super-rule generation for trans membrane segments prediction’, *IEEE Computational Systems Bioinformatics Conference Workshops (CSBW'05)*, Stanford University, California, USA
- [10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). ‘A density-based algorithm for discovering clusters in large spatial databases with noise’. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*
- [11] Wang, G. and Dunbrack, R.L. (2003) *PISCES: a Protein Sequence Culling Server*, *Bioinformatics*, Vol. 19, No. 12, pp.1589–1591.
- [12] Kabsch, W. and Sander, C. (1983) ‘Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features’, *Biopolymers*, Vol. 22, pp.2577–2637.

Mapping Genes to Diseases With Translational Data Mining

M. von Korff¹, A. Klenk¹, and T. Sander¹

¹Research Information Management, Actelion Pharmaceuticals Ltd., Allschwil, Switzerland

Abstract - A new software tool (*Gene2diseaseMapper*) is presented that takes the HUGO symbol of a gene as input and delivers a ranked list of diseases, considering microarray experiments. Gene name synonym expansion was used to generate a query for MEDLINE. Retrieved article records were filtered by a disease stoplist which was created from Medical Subject Headings (MeSH). From the ArrayExpress microarray database, all experiments were retrieved in which the gene was differentially expressed. The experiments were searched for disease terms extracted from the MEDLINE articles. A similarity function was developed to compare the MeSH terms and the terms in ArrayExpress. A scoring function was implemented which ranked the disease MeSH terms according similarity and frequency. The method was explored with 12 genes for whose corresponding protein a drug was approved or is under development. *Gene2diseaseMapper* was able to find the diseases of the approved drugs in 11 out of 12 cases.

Keywords: Bioinformatics, Translational medicine, Data mining, Microarray, Medline, Medical subject headings

1 Introduction

Many diseases are related to a change in the expression of proteins. The largest changes in protein expression can be found in cancer related diseases. Inflammation-related diseases also cause large changes in the protein expression pattern. A prominent example is rheumatoid arthritis, which affects millions of people worldwide. Neurodegenerative diseases such as Alzheimer, Parkinson and Huntington also show different protein expression patterns when compared to healthy control groups. Changes in protein expression can also be used to detect endogenous biomarkers; i.e. molecules which indicate a disease state [1]. For several years, microarrays have enabled expression profile analysis of the whole genome. The expression profiles of tens of thousands of genes can be explored in one experiment. In combination with information about experimental conditions, control groups, test groups and treatment these profiles are a valuable source for data mining. Because many journals require submission of the microarray data to one of the public repositories such as Gene Expression Omnibus [2] (GEO) or ArrayExpress [3] together with a publication, an enormous

and quickly growing resource of medical information is available. The Gene Expression Atlas of the ArrayExpress database contains curated data from more than 5600 experiments. The differential expression of the genes is already calculated and the database is programmatically accessible. A p-value is given for every calculated differential expression value, indicating its reliability. The experiments are described and annotated with the Experimental Factor Ontology (EFO) [4].

A central point in a drug discovery program is the determination of the protein that will be targeted by the drug. This is often done at the beginning of a project, when a gene is chosen for cloning and expression to establish a biological screening assay. It is also possible to start a drug discovery program with a phenotype-based approach, but a drug will hardly be approved without a defined target protein. This target protein has to fulfill manifold requirements. It has to be related to a disease with a certain chance to cure it or, if not possible, to palliate symptoms. The gene encoding the target protein has to be known in order to develop biological screening assays.

Searching the medical literature for the relation between target and disease is the starting point in drug discovery. The huge expenditures for pharmaceutical research during the last decades resulted in a plethora of publications. Most of the medical information is not published in open access journals and is therefore not freely accessible. But almost all relevant biomedical literature is indexed in MEDLINE [5]. With more than 17 million bibliographic records MEDLINE is the largest repository for biomedical literature. Interfaces like PubMed enable human and programmatic access. What makes MEDLINE interesting for drug discovery is not only the specialization in life science-related subjects, but the hierarchical indexing system used to categorize the collected publications. The medical subject headings (MeSH) thesaurus consists of a controlled vocabulary, the MeSH descriptors, supplementary concepts and entry points [6]. The hierarchical structure of the MeSH thesaurus can be mapped onto a tree with general concepts close to the root and specific concepts in the leaves. Each node in the tree contains a unique node name, a MeSH descriptor, related concepts and entry points. There is no MeSH descriptor for the root node. Articles in MEDLINE are indexed with MeSH descriptor terms by searching the articles for entry points.

As a working hypothesis for this examination, it was assumed that the vocabulary from ArrayExpress and the standardized vocabulary from the medical subject headings can be used to detect overlapping information. Starting with the approved symbol for a gene it should be possible to detect information about related diseases in MEDLINE. Searching a microarray database with a gene symbol and related disease information should retrieve evidence on differential expression of that gene in a disease.

2 Methods

2.1 Stoplists

Stoplists are lists of node names that are used to activate or inactivate branches in the MeSH tree [7]. Only active MeSH terms are used for searching corresponding expressions in the microarray data. A simple stoplist for diseases activates the whole branch 'C' in the MeSH tree. Without sub-branch C22, which contains MeSH terms related to animal diseases, branch C contains 10782 nodes with 4466 unique MeSH headings (stoplist disease). Branch C04 in the MeSH tree was deactivated while searching microarray datasets for diseases which are not related to any form of cancer (neoplasms). Branch C04 contains only cancer-related MeSH terms. In addition, nodes in other branches were deactivated if their heading was equal to one of the headings in the C04 branch. After applying the disease stoplist omitting cancer, 8905 MeSH nodes with 3807 unique descriptor headings remained (stoplist disease, no cancer). Separating between neoplasm and other diseases is necessary, because neoplasia causes so many changes in gene expression that the relations between gene expression and other diseases would not be recognized. Of course this kind of restriction can also be applied to other diseases; e.g., inflammation-related diseases causes manifold changes in gene expression.

2.2 Programmatic access to PubMed

The MEDLINE databases can be accessed programmatically via the Entrez tools [8]. A query, containing a search term, submitted to MEDLINE via the PubMed interface returns a list of identifiers (PMID) which is used to obtain the publication records \mathbf{R} . These records contain bibliographic information, often an abstract and the MeSH term headings which were used to index these articles.

2.3 Gene names and synonyms

A table with Human Genome Organization (HUGO) ids, gene names, approved symbols and synonyms was retrieved from HGNC (HUGO Gene Nomenclature Committee) [9]. The HUGO Gene Nomenclature Committee is located at the European Bioinformatics Institute and works under supervision of the Human Genome Organization. From HGNC a table with gene names and their synonyms was retrieved. The

MEDLINE database Gene also delivered HUGO ids, gene names and synonyms. [10] There is not a complete overlap between the synonyms in the two databases.

2.4 Searching PubMed records with gene names

To generate the query for searching the PubMed database, the approved symbol from the HUGO Gene Nomenclature Committee (HGNC) was used to find the synonyms from PubMed Gene and genenames.org. The synonyms were combined in a string by using 'OR' and sent as a query to PubMed. Without any further specification all fields in the PubMed database were searched. Depending on the gene symbol, a few records to the extent of several ten thousand were retrieved. All records which did not contain at least one active MeSH descriptor of the disease branch were skipped (stoplist disease, no cancer). The result was a dataset \mathbf{R}_{Gene} for each gene. For the genes TNFSF11 and TPPP the branch C04, containing cancer-related diseases, was also activated (stoplist disease).

2.5 Searching ArrayExpress database

The Gene Expression Atlas of the ArrayExpress database contains curated and re-annotated microarray datasets [3]. This database was queried with the HUGO symbol for the gene under consideration. All experiments $\mathbf{MA}_{\text{Gene}}$ in which this gene was differentially expressed were retrieved. A gene experiment record contains the identifier of a microarray experiment in which the gene is up or down-regulated. Connected with the microarray experiment identifier is a record containing the experiment title, a description, the sample attribute values and the experimental factor values.

2.6 Searching microarray experiments with disease MeSH terms

The microarray experiments $\mathbf{MA}_{\text{Gene}}$ were searched for matching disease MeSH terms from \mathbf{R}_{Gene} . Each disease term was compared with the title, the description, the sample attribute values and the experimental factor values. Each of the resulting similarities S_{Title} , $S_{\text{Description}}$, S_{Sample} and S_{ExpFac} was multiplied by S_{Disease} , the frequency of occurrence of the disease term in the retrieved publication records. The highest scores from all microarray experiments $\mathbf{MA}_{\text{Gene},i}$ were summed up. Because the terms which are used to annotate the experiments in ArrayExpress differ from the medical subject headings, a similarity function was needed to find similar terms. Each term, medical subject header \mathbf{t}_{MeSH} or from a microarray experiment \mathbf{t}_{MA} , was decomposed into a list of unique words \mathbf{u}_{MeSH} and \mathbf{u}_{MA} . A complete similarity matrix between these two lists was calculated by single word comparison using the Levenshtein similarity function [11]. From this matrix the optimum list of similarity pairs was

derived and their median taken as total similarity score for $\text{sim}(\mathbf{u}_{\text{MeSH}}, \mathbf{u}_{\text{MA}})$. If at least one word from \mathbf{u}_{MeSH} did not fit with a similarity ≥ 0.8 to any word in \mathbf{u}_{MA} the similarity $\text{sim}(\mathbf{u}_{\text{MeSH}}, \mathbf{u}_{\text{MA}})$ was set to 0. The score for a disease term $s_{\text{MA,MeSH}}$ was computed as the sum of all products of the frequency of occurrence of this term in the publication records multiplied by the maximum similarity of this term with a corresponding term in the microarray annotation.

3 Experiments

3.1 Targets with approved drugs

Seven targets for which a recently approved drug was available were chosen from the literature ($\text{GeneSet}_{\text{Mature}}$) (Table 1) [12]. With the HUGO approved gene symbol and the found synonyms a query string was generated and the PubMed records were retrieved as described in the paragraph “Searching PubMed records with gene names“. The retrieved records were filtered with the disease filters and the remaining records underwent a first evaluation. From the MeSH headings a simple histogram was generated with the most frequent MeSH terms at the top. An example is given for gene HTR1A in Table 2. The rank of the indication equal to the indication of the approved drug was taken as a figure of merit for the applied algorithm. One rank score was obtained for the PubMed record derived MeSH term histogram and one for the sorted scores $s_{\text{MA,MeSH}}$ of the microarray experiment to MeSH term comparison.

3.2 Targets with drugs in development

Another set of five genes ($\text{GeneSet}_{\text{New}}$) was selected for which a drug was in development for the encoded protein [12]. The genes in $\text{GeneSet}_{\text{New}}$ are much less well explored than the genes in $\text{GeneSet}_{\text{Mature}}$, as can be seen from the number of retrieved PubMed records (Table 3). “Neuroinflammatory disease” is the indication of the corresponding drug for gene ALCAM. Because there is no MeSH term “Neuroinflammation” the indication was set to

Table 2. MeSH term histogram for HTR1A with expanded query. Applied stoplist: disease, no cancer. “Frequency” is the frequency of occurrence of the MeSH headings in the 17617 PubMed records.

Rank	Disease MeSH heading	Frequency
1	Depression	206
2	Schizophrenia	178
3	Pain	162
4	Body Weight	152
5	Inflammation	133
6	Genetic Predisposition to Disease	123
7	Hypertension	118
8	Hypothermia	112
9	Heart Failure	100
10	Catalepsy	95

inflammation.

4 Results and conclusions

For dataset $\text{GeneSet}_{\text{Mature}}$ all approved drug indications were found by the MeSH term histograms (Table 4, index 1-7) and all found indications had a histogram rank below five, except for TPPP which was at rank 42. In dataset $\text{GeneSet}_{\text{New}}$, containing less explored genes (Table 4, index 8-12) also all drug indications were found. Two outliers were observed with the gene ALCAM and SLC6A7 on rank 27 and 29 respectively. In 11 of 12 cases the indications were confirmed by microarray experiments. For F13A1 (Factor XIII deficiency) no matching sample or condition was found in Gene Expression Atlas. Table 5 shows that the number of gene name synonyms ranges from 7-22. Comparing the retrieved number of articles for the HUGO symbol only and the query containing the gene name synonyms demonstrates a huge increase in retrieved articles by using synonyms (Table 1 and 5).

Table 1. Seven drug targets with at least one approved drug on the market ($\text{GeneSet}_{\text{Mature}}$). “HUGO” is the approved symbol for the target protein encoding gene. “Articles” is the number of articles retrieved from PubMed for the expanded gene query. “Articles, no expansion” is the number of filtered articles querying PubMed with the HUGO symbol only.

Index	HUGO	Drug	Indication	Articles	Articles, no expansion
1	HTR1A	Vilazodone	Depression	17617	103
2	TNFSF11	Xgeva	Bone metastases	5956	1
3	TPPP	Eribulin	Breast neoplasm	8796	42
4	GHRH	Tesamorelin	Obesity HIV patients (Obesity)	9705	2891
5	GLP1R	Victoza	Diabetes mellitus	866	36
6	PDE4	Roflumilast	Chronic obstructive pulmonary disease	2835	158
7	F13A1	Corifact	Factor XIII deficiency	3179	68

Table 3. Five drug targets with a drug in development (GeneSetNew). For explanations see Table 1.

Index	HUGO	Drug	Indication	Articles
8	ALCAM	AT-002 (CD166)	Neuroinflammatory disease (Inflammation)	507
9	APOC3	(Isis pharmaceuticals)	Cardiovascular disease	1544
10	SLC6A7		Alzheimer	1896
11	MAPKAPK5	GLPG-0259	Rheumatoid arthritis	247
12	CXCL16		Inflammation	241

The ratio between unique MeSH terms and the number of retrieved article records shows a roughly tenfold reduction taking the median of all values. CXCL16 is an interesting outlier, because with 238 retrieved records 114 MeSH terms were found. Remember that these are MeSH terms from the diseases stoplist that excluded cancer-related terms. This indicates that this target is active in a multitude of disease processes. The microarray experiments gave additional information. The number of experiments in which the gene under consideration was differentially expressed ranged from 233 for APOC3 to 741 for ALCAM. Interestingly, ALCAM is the gene with the second lowest number of matches between MeSH terms and microarray experiments. This means that many sample values and conditions from the microarray experiments did not match any one of the 77 disease MeSH terms which were used to index the literature containing one of the ALCAM gene name synonyms.

In conclusion, the proposed method summarizes up to thousands of MEDLINE publication records and relates the indexing MeSH terms to hundreds of microarray experiments in the Gene Expression Atlas. After sorting disease related

MeSH term lists according to their score $s_{MA,MeSH}$, indications for approved drugs and drugs under development were at the top of the lists. Disease stoplists as filters for the indexing MeSH terms together with publicly available microarray data were successfully applied to targets of approved drugs and drugs under development. This demonstrates that Gene2diseaseMapper implements a new data mining method which prioritizes indications for targets in drug discovery programs.

Table 4. Result table for GeneSetMature (index 1-7) and GeneSetNew (index 8-12). "Indication" is the indication given for the drug targeting the protein encoded by "Gene". "Rank PubMed histogram" is the rank of the indication in the histogram of the MeSH terms which were derived from the PubMed query with the corresponding gene names. In "Rank PubMed-microarray" the rank of the indication according to the scored microarray experiments "MA score" is given. "MA score" is the resulting score from the evaluation of the microarray experiments with the disease MeSH terms. a No experiment with Factor XIII deficiency was found in the Atlas DB.

Index	Gene	Indication	Rank PubMed histogram	Frequency	Rank PubMed-microarray	MA score
1	HTR1A	Depression	1	189	1	2057
2	TNFSF11	Bone metastases	3	259	6	798
3	TPPP	Breast neoplasm	42	27	6	504
4	GHRH	Obesity HIV patients (Obesity)	4	190	1	784
5	GLP1R	Diabetes mellitus	1	195	1	14
6	PDE4	Chronic obstructive pulmonary disease	3	51	3	104
7	F13A1	Factor XIII deficiency	1	139	a	0
8	ALCAM	Neuroinflammatory disease (Inflammation)	27	3	13	48
9	APOC3	Cardiovascular disease	8	46	10	50
10	SLC6A7	Alzheimer	29	4	34	4
11	MAPKAPK5	Rheumatoid arthritis	6	2	10	4
12	CXCL16	Inflammation	1	14	1	240

Table 5. Details of the search for the genes in GeneSetMature (index 1-7) and GeneSetNew (index 8-12). "Synonyms" is the number of gene name synonyms that were used to query MEDLINE. "Articles" contains the number of unique MEDLINE article records that were retrieved by the query. "Unique MeSH terms" is the number of unique MeSH terms found in the indexing section of the articles. Column six gives the ratio between unique MeSH terms and the number of unique articles. Column seven contains the number of microarray experiments where the gene was found to be differentially expressed. Column eight indicates how many disease MeSH terms matched on at least one microarray experiment.

Index	Gene	Synonyms	Articles	Unique MeSH terms	Ratio MeSH/Articles	MA Experiments with differentially expressed genes	Matching MeSH on MA Experiments
1	HTR1A	16	17617	5044	0.29	258	279
2	TNFSF11	22	5956	5438	0.91	326	212
3	TPPP	16	8796	9070	1.03	313	315
4	GHRH	14	9705	4933	0.51	250	288
5	GLP1R	7	866	597	0.69	293	52
6	PDE4	11	2835	859	0.30	474	171
7	F13A1	16	3179	2068	0.65	429	282
8	ALCAM	12	507	298	0.59	741	75
9	APOC3	9	1544	1549	1.00	233	87
10	SLC6A7	7	1896	582	0.31	243	140
11	MAPKAPK5	9	247	69	0.28	451	33
12	CXCL16	13	247	69	0.28	404	77

5 Acknowledgement

We thank Susan Flores for editorial assistance.

6 References

- [1] Y. Bauer, P. Hess, C. Qiu, A. Klenk, B. Renault, D. Wanner, R. Studer, N. Killer, A. K. Stalder, M. Stritt, D. S. Strasser, H. Farine, K. Kauser, M. Clozel, W. Fischli, and O. Nayler. "Identification of Cathepsin L as a Potential Sex-Specific Biomarker for Renal Damage"; *Hypertension*, 57, 4, 795-801, Feb, 2011.
- [2] <http://www.ncbi.nlm.nih.gov/geo/>, accessed March 7 2011.
- [3] <http://www.ebi.ac.uk/arrayexpress/>, accessed Feb 13 2011.
- [4] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. "Modeling sample variables with an Experimental Factor Ontology"; *Bioinformatics*, 26, 8, 1112-1118, Apr, 2010.
- [5] <http://www.nlm.nih.gov/pubs/factsheets/medline.html>, accessed Feb 8 2011.
- [6] H. J. Lowe, and G. O. Barnett. "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches"; *JAMA*, 271, 14, 1103-1108, Apr, 1994.
- [7] Don R. Swanson, Neil R. Smalheiser, and Vetle I. Torvik. "Ranking indirect connections in literature-based discovery: The role of medical subject headings"; *Journal of the American Society for Information Science and Technology* 57, 11, 1427-1439, 2006.
- [8] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvermin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmsberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. "Database resources of the National Center for Biotechnology Information"; *Nucleic Acids Res*, 34, Database issue, D173-180, Jan, 2006.
- [9] <http://www.genenames.org>, accessed Mar 29 2011.
- [10] <http://www.ncbi.nlm.nih.gov/gene>, accessed Feb 10 2011.
- [11] Fred J. Damerau. "A technique for computer detection and correction of spelling errors"; *Communications of the ACM*, 7, 3, 171-176, 1964.
- [12] <http://www.pharmaprojects.com/>, accessed Mar 30 2011.

Protein Sequence Motif Extraction Using Decision Forest

Bernard Chen¹, Cody Hudson¹, Minwoo Kim¹, Aaron Crawford¹, John Wright¹, and Dongsheng Che²

¹University of Central Arkansas, Department of Computer Science, Conway, AR, 72034

²East Stroudsburg University, Department of Computer Science, East Stroudsburg, PA, 18301

Abstract— As one of the more prominent areas of bioinformatics research, protein sequence analysis has gathered considerable interest. The structure, function, and activities of the protein are strongly linked to structural motifs found in its sequence data. Building off of past research, we propose a new granule model that combines the strength of fuzzy logic and granule computing, with the speed and robustness of a decision tree for the purpose of identifying and extracting protein motif data that transcends protein families. We propose parameters for the model and test their effectiveness using several measures of accuracy and quality. The end result, a decision tree example, is explored for its usefulness in this endeavor.

Index Terms—FGK Model; Decision Forest; Entropy Threshold; Protein Sequence Motif;

I. INTRODUCTION

As one of the basic components of an organic body, proteins have been of prominent interest for many years now in various fields of study. As such, their shape, their functions, and the analysis thereof have become increasingly important. In the past, the process by which one would link both protein structure and shape to its function was through arduous and time consuming methods[1] that included well known processes such as crystallography[2], spectroscopy, and various others. However, in recent years, the promising field of bioinformatics and its accompanying data mining techniques has broken into novel ground by looking not directly at the shape of the protein, but rather at its base composition. Doing so allows the prediction of the three dimensional shape of the protein within an acceptable threshold of accuracy.

To understand this, one must understand that a protein can be described by three basic categories: primary, secondary, and tertiary structure. A protein's primary structure, or "base composition," is its amino acid sequence. These are the building blocks of proteins and the repeating patterns therein are known as motifs. Each of these amino acids can have non-covalent, intermolecular reactions with other amino acids within the protein, causing repeating patterns of folds and sheets within the protein's structure. This localized substructure describes the protein's secondary structure. Finally, the tertiary structure of the protein is the overall three dimensional shape. This is important because not only does the tertiary structure of a protein denote its function, but

biochemical research and data would suggest a protein's shape is heavily determined by its primary structure (assuming the absence of any denaturing agents, such as heat or acid)[3]. This supports the idea of using data mining and bioinformatics as a tool for analyzing the primary structure in order to predict the tertiary structure of a protein.

Naturally, in order for analysis of protein data to occur, the data has to be both available and numerous, which suggests that databases are good repository of protein information. Three of the most popular protein databases would include PROSITE[4], PRINTS[5], and BLOCKS[6]. Each describes, in some detail, the various structures of the protein, and, to some degree, also supports the idea that reoccurring primary and secondary structural patterns suggest common tertiary structure.

Various researchers have tried using such databases and numerous techniques[7] to glean some meaningful correlation between protein structure and its three dimensional shape. One such study by Han and Baker utilized their K-means clustering algorithm [8, 9]. Using said algorithm, the protein motifs discovered by it, and an additional algorithm, Hidden Markov Model [10], they were able to predict with some level of success the local tertiary structure of various proteins. In the previous works related to this paper, a Fuzzy C-means algorithm was used to initially break the data into ten subsets. A K-Means algorithm was then utilized to refine each subset. This combination (noted as the FGK Model), was used to not only analyze similarities among protein structures, but also to eliminate low quality data [11]. Support Vector Machines were then proposed to be used for the purpose of predicting the shape of the protein using the above analysis [12].

Granted such, the methodologies proposed within this paper suggest the use of decision trees in the stead of SVM. This method would be used to adequately analyze a protein's primary and secondary structure, as well as offer the ability to use such trees for the prediction of the protein's tertiary structure.

Decision tree algorithms offer output in an easy to understand format, showing precisely how the algorithm made its decisions [13, 14]. Unfortunately, the algorithm requires the calibration of several parameters, including entropy threshold, data classifiers, and labelers. This paper discusses how each parameter has been chosen for further research on the matter.

Therefore, the ID3 (Itemized Dichotomizer 3) decision tree[14] is being proposed to extend the before mentioned previous works related to this paper (the FGK Model)[11]. With its ability to define whether proteins belong to a given cluster, it will be instrumental in eliminating noisy or meaningless data. As the purpose is to discover small, sequential patterns within the amino acid sequence in order to relate to common tertiary structures, it is only natural that not all data will be important. Thus, in this paper, the use of the ID3 decision tree algorithm in order to relate patterns of primary and secondary protein structure to its tertiary structure will be discussed. Just as well, the processes by which its parameters are decided for this particular solution will be heavily discussed primarily through the use of statistical charts describing the output of the decision trees. The following sections of the paper will be arranged as such: methods (describing both present and past approaches used to solve this problem), experimental setup (describing the input in more detail, all utilized equations, etc.), results, future works, and conclusion.

II. METHODS

2.1 Data Set Challenges-Large and Random

As one might suspect, to adequately analyze protein primary sequences, one must overcome the challenges the data presents. The sheer size of the dataset can make even fairly robust data mining techniques seem rather inadequate. Coupled with the inherent random and noisy nature of pulling data from various, somewhat disparate databases [4, 5, 6], the task becomes even more difficult. This is particularly despairing in the case of using a decision tree as it is fairly susceptible to outliers and random data. However, previous works suggest that a preliminary analysis of the data with the "FGK Model"[11], tackles both of the before mentioned problems with a promising level of success. The data can then be further and efficiently processed by the proposed ID3 algorithm.

Granted such, our previous works refers to the experiments of Wei et al [15], which handles, specifically, the randomness aspect of the protein data set. Using the basic idea of the K-Mean clustering algorithm, one will note that all initial centroids are randomly chosen. This potentially renders the algorithm worthless in data that is fairly random in the first place. Instead, they proposed that one run the K-Means algorithm five times. In each round, the randomly generated initial points that had the potential to form clusters with high structural similarity were chosen for the improved K-Means clustering algorithm. These were checked against other potential points, and if its minimum distance fell within a given threshold, it was included as an initial centroid.

The method used in the "FGK Model" was similar, but used a method more akin to averaging the results of the five K-Means runs to produce centroids for a sixth iteration. The resulting clusters from this additional run of the "Greedy K-Means"[11] algorithm used these centroids to produce clusters of various qualities. These qualities are determined by analyzing secondary structural similarity of the proteins in each cluster (the equation for such is given in section 3). Each cluster and its respective centroids are ranked by these

structural similarity values, under the safe assumption that centroids that produce higher quality clusters are more desirable.

2.2 Fuzzy Greedy K-Means (FGK) Model

The problem of an overly large and complex dataset is still a prominent issue. Although the five iterations of the traditional K-Means algorithm and then a sixth application of the so-called "Greedy K-Means" algorithm sufficiently handle a great deal of the noise in the data, it is still undesirably inefficient when dealing with the entire data set at one time. However, the proposed FGK Model presents a solution via a simple concept of granular computing. The concept proposes that a divide-and-conquer idea be used to break the original problem into various subsets that can be more easily processed by any given algorithm. In other words, it breaks the original data set into "information granules." [16, 17] Although one might argue that this is simply spreading the running time across various subsets, this isn't true. This is especially important in the case of the K-Means algorithm, which has a running time that increases significantly with a larger dataset.

Therefore, the combination of the "Greedy K-Means" algorithm, and the concept of granular computing produces the FGK Model. The FGK model essentially breaks the protein data set into ten information granules using Fuzzy C-Means. Then performing the five iterations of the traditional K-Means algorithm and the sixth Greedy K-Means run solves both issues with data complexity and size. The resulting output groups the data into ten information granules containing any number of clusters containing any number of proteins.

2.3 Decision Tree Forest Model

Now, we know the FGK Model adequately processes the data, clustering it according to its primary structure into both granules and further into clusters. Even stating so, this model still needs further tools to produce any novel or interesting findings. Thus, the ID3 decision tree algorithm[14] is proposed to further the model. This produces a mechanism that, once trained in the typical fashion, can tell if any given random protein belongs to a cluster with decent prediction accuracy. This is to say that this paper suggests that a "forest" of decision trees is to be created for each cluster in each granule (producing, with the given dataset, a total of 799 decision trees). Each decision tree in the so called forest is trained on the individual clusters' proteins. This would imply that each decision tree will have a basic idea of the inherent sequential patterns (i.e. motifs) within each protein set, such that it can be used to then analyze a given protein's primary sequence. If the decision tree produces a "yes" (the meaning of which, in this particular context, will be explained in the experimental setup section) for that given protein, then this would suggest that the protein has similar characteristics to the homologous proteins within the cluster, including tertiary structural characteristics. A model of such can be seen in Figure 1, combining the elements of the FGK Model and the new Decision Tree Forest Model, to produce a novel approach that takes the analysis power of decision trees and combines it with the data sorting and cleaning power of the FGK-Model.

Thus, the basic concept of the ID3 algorithm will be followed heavily to produce each of the 799 decision trees.

The algorithm, while simple, is fairly robust with large data sets and adequately accurate for this particular task. Granted such, it seems obligatory to note that any future works related to this would make use of much more appropriate decision tree algorithms, as the ID3 algorithm is largely a proof of concept. This is not to say that any results produced by this algorithm are not applicable, but rather that this research team realizes there are more appropriate, albeit more complex, decision tree algorithms to apply.

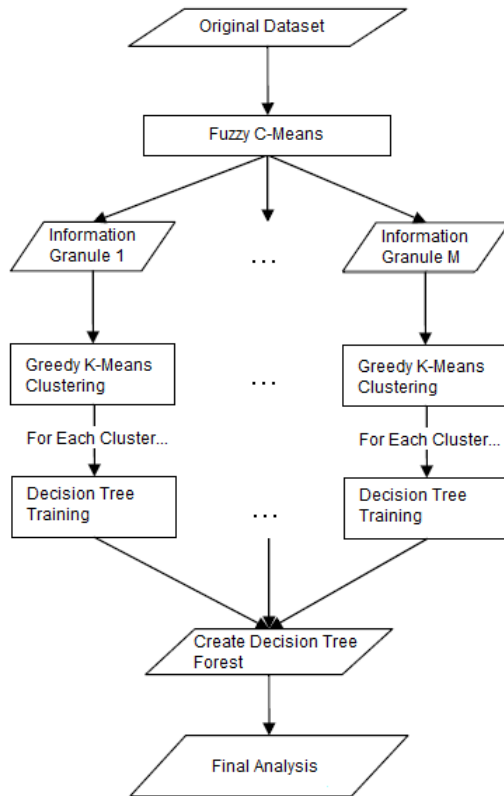


Figure 1. The FGK-Decision Tree Forest Model

III. EXPERIMENTAL SETUP

3.1 Dataset

The incoming dataset that is first analyzed by the overlying FGK-Model is composed of 2710 protein sequences obtained from the Protein Sequence Culling Server (PISCES)[18]. None of the protein sequences within this database share more than a 25% sequence identity. Sliding windows with nine successive residues are generated from each protein sequence, such that each window represents one sequence segment of nine continuous positions. Granted such, more than 560,000 segments are generated by this method. Also added to this dataset is the protein's frequency profile, generated from the HSSP[19]. This frequency is based on the alignment of each protein sequence from the Protein Data Bank (PDB), where all the protein sequences are considered homologous in the sequence database. The secondary structure of each protein is also generated from DSSP[20], which is simply a database containing secondary structural assignments for all protein entries in the Protein Data Bank.

The FGK-Model will take this dataset and produce 799 clusters divided among ten information granules. Each granule will have a varying number of clusters within it (this number is determined by a function explained in section 3.4). Each cluster, itself, will have a varying number of protein sequence information in it as well. They will also be of a varying secondary structural similarity (explained in section 3.7). Each of these clusters will then be used as the dataset for the induction of each individual decision tree for reasons described in the Methods section.

3.2 Representation of Sequence Segment

As mentioned, the sliding windows of nine successive residues are generated from all of the 2710 protein sequences. Each window corresponds to a sequence segment, which is represented by a nine by twenty matrix, plus an additional nine places corresponding to the secondary structure data obtained from DSSP. Twenty rows represent twenty amino acids and nine columns represent each position of the sliding window. For the frequency profile (HSSP) representations of the protein sequence segments, each position of the matrix represents the frequency for a specified amino acid residue in a sequence position for the multiple sequence alignment. The secondary structure generated from DSSP is simplified from its original eight different classes, down to three. In this paper, structures denoted by H, G, and I are converted to H (Helices), B and E are converted to E (Sheets), and all other structures are converted to C (Coils).

3.3 Distance Measure

As the FGK-Model contains K-Means at its core, a distance formula is imperative. According to various sources[9,15], the most appropriate distance formula to use is the city block metric, as each position in the generated frequency profile will be considered equally. Thus, the following formula is used to calculate the distance between two sequence segments when clustering [9]:

$$\text{Dissimilarity} = \sum_{i=1}^L \sum_{j=1}^N |F_k(i, j) - F_c(i, j)|$$

Where L is the window size (in this case nine) and N is twenty, representing the twenty different amino acids. $F_k(i, j)$ is the value of the matrix at row i and column j , which represents, in this case, the sequence segment. $F_c(i, j)$ is the value of the matrix at row i and column j , which represents the centroid of a given sequence cluster.

3.4 FGK-Model Parameter Setup

For the Fuzzy C-Means Clustering that is included in the FGK-Model, the fuzzification factor is set to 1.05 and the number of clusters is set to ten. These settings yielded the best results for this particular dataset. The reason for this being, if the fuzzification factor was to remain constant, but the number of clusters was set to twenty, the membership function would produce nearly equal membership to all clusters for each segment. If one was to decrease the fuzzification factor instead, overflow becomes probable.

In order to separate the information granules generated by the above Fuzzy C-Means results, the membership threshold is

set to twelve percent. Using this value, fifteen percent of the dataset is filtered out and the remaining eighty-five percent is assigned to one or more of the clusters. The formula that dictates how many clusters should be included in each information granule is given below:

$$C_k = \frac{n_k}{\sum_{i=1}^m n_i} \times \text{total number of cluster}$$

Where C_k denotes the number of clusters assigned to an information granule k . The number of members belong to information k is denoted as n_k . The number of clusters in FCM is denoted as m . Although using this methodology causes the total data size to increase from 413 MB to 529 MB, as well as an increase in total number of members from 562745 to 721390, it allows for one to deal with one information granule at a time. For example, the largest information granule generated contains 136112 members. From that granule, 151 clusters should be computed from those members, generating a data with the size of 99.9 MB. Compared with the original dataset, the largest granule is only twenty-five percent the size. Therefore, the computation time for all information granules (231720 seconds) is a mere twenty percent of the running time of other leading research [15] (1285928 seconds). These results support the idea that the FGK-Model is a viable one for reducing space and time complexities.

3.5 Decision Tree Induction

The ID3 decision tree algorithm [14], as most classifying algorithms, requires a period of training to produce any level of output. For each cluster generated by the overlying FGK-Model, a decision tree will be trained and generated by considering the frequency profile of each segment in a given cluster. This training produces a resulting decision tree that will now represent the sequential motifs in said cluster. This particular implementation of the ID3 algorithm uses the general formulas for producing both entropy and information gain, both given below:

$$\text{Entropy}(S) = - (S_Y/S_C) \log_2(S_Y/S_C) - (S_N/S_C) \log_2(S_N/S_C)$$

Where S is a collection of total size S_C , S_Y is all items belonging to a given cluster, and S_N is all items not belonging to a given cluster. How these items are labeled is described in section 3.6.

$$\text{Gain}(S, A) = \text{Entropy}(S) - (S_V/S_C) \text{Entropy}(S_V)$$

Where S is a set of each value v of all possible values of attribute A , S_V is the subset of S in which attribute A has the value v , and S_C denotes all items in set S .

3.6 Class Labeling

To generate each label that will determine whether or not a given protein is to be classified as a “yes” protein (that is, it belongs to its cluster generated by the FGK-Model) or a “no” protein, one must consider the secondary structure. For each cluster, a representative secondary structure is generated by determining the secondary structural motif (H, E, or C in this

paper) that is most characteristic (that is, the motif with the highest count in that particular column). This is done in each of the nine secondary structural positions for that particular cluster. Once the representative secondary structure for that cluster is generated, each of the proteins are then analyzed for their similarity to this representative structure by both position and the motif at that position and then given an appropriate score. For example, if the representative structure for the cluster (assuming only three structural positions) is HHH, and an individual protein sequence has a secondary structure HEH, then this protein would be given a score of two out of three. For this research, the scores range from 0 (that is, the protein has no similarity to the given representative structure of the cluster) to 9 (which denotes a protein that is fully representative of the cluster). Labeling can then be performed based on this score, such that any values over a certain number, what we will call our label pivot (a parameter discussed in section 3.10), are then considered a “yes” protein. All others would be considered a “no” protein.

3.7 Secondary Structural Similarity Measure

Used in the FGK-Model, the formula to calculate a cluster’s secondary structure similarity is given by the following formula:

$$\text{Secondary structural similarity} = \frac{\sum_{i=1}^{ws} \max(P_{i,H}, P_{i,E}, P_{i,C})}{ws}$$

Where ‘ws’ is the window size and $P_{i,H}$ shows the percentage of helix (H) occurrences among the segments for the cluster in position ‘i.’ $P_{i,E}$ and $P_{i,C}$ are defined in a similar way in respect to sheets and coils.

Granted such, if the generated structural homology for a given cluster is seventy percent or greater, the cluster can be considered structurally identical [19]. If it falls between sixty percent and seventy percent, it can be said to be weakly structurally homologous [15].

3.8 Average Node Secondary Structural Similarity Measure

Decision trees are defined, primarily, by their nodes, not by clusters. Given such, it is necessary to also include an average node secondary structural similarity measure, given by the following formula:

$$\frac{\sum_{i=1}^n |\text{Secondary_Structural_Similarity}|}{n}$$

Where the “Secondary_Structural_Similarity” is the equation defined in section 3.7 and number of decision nodes is denoted as ‘n.’

3.9 Ideal Prediction Accuracy Measure

To aid in choosing appropriate parameters, another measure that is made for each decision tree is its ideal prediction accuracy. A twenty-fold cross validation, or similar measure, isn’t used in this particular case due to the sheer size of the data as well as the fact that each decision tree is tested on twenty-one different entropy threshold values (described in section 3.10). Instead, the ideal prediction accuracy measure is generated by simply running the training data (that is, the frequency profile of each protein in a given cluster) through

the same tree it produced. This is done by comparing the labels given to the test data by methods explained in section 3.6, against the decisions made by the decision tree for each protein. This summation of all correctly made decisions (regardless of whether or not it is a “yes” decision or a “no” decision) is divided by the number of decisions made. This gives a percentage that shows directly how changing entropy thresholds affects the predicting power of a given decision tree.

3.10 Decision Forest Parameter Setup

The ID3 decision tree, in this particular implementation and application, has three primary parameters, some of which have already been defined: label pivot, attribute range set, and entropy threshold[14]. The label pivot determines what range of labels, as described in section 3.6, are considered “yes” labels, and, alternatively, the range that denotes “no” labels. Naturally, the magnitude of this number has a large effect on the outcome of the decision trees. The attribute range set is composed of a short list of amino acid frequency ranges that serve as the classifying attributes. The length of this list and the distance between each of the bounds of the ranges also has a prominent effect on the decision tree, and its respective measures. The most sensitive parameter, however, is the entropy threshold, or, rather, the allowed level of randomness before the decision tree can make a decision. As one might expect, the closer the threshold is to 1.0, which is the maximum entropy a dataset can have, the shorter and less effective the decision tree becomes. Yet, an entropy threshold that is too restrictive (i.e. close to 0.0) would be detrimental to the purposes of this research for reasons explained more in depth in the Experimental Results section.

The parameters tested in this experiment include two label pivots (six and seven), two attribute range sets ($\{0-4, 5-7, 8-14, 15-29, 30-100\}$, $\{0-7, 8-14, 15-29, 30-100\}$), and twenty-one different entropy thresholds, ranging from 0.0 to a maximum of 1.0 while incrementing by 0.05 units. All of these parameters were tested on all 799 protein clusters, such that 268,464 unique tuples were generated, giving various measures described in each of the sections above. The results of these tests are described in the Experimental Results section.

IV. EXPERIMENTAL RESULTS

4.1 Parametric Criteria

For each of the 799 protein clusters generated by the FGK-Model, and for each of the parameter choices as described in section 3.10, an array of measures were recorded. This data was used for the purpose of deciding upon the most appropriate values for the three parameters for the decision tree implementation. These measures included ideal prediction accuracy, average node secondary structure similarity, average *yes node* secondary structural similarity, decision node count, yes decision node count, and number of proteins classified within those yes nodes. Also included was a range of values that counted the percentage of decision nodes that had a secondary structure similarity measure of over 90%, 90-80%,

80-70%, 70-60%, and less than 60% structural homology. These values were used to determine what combination of entropy threshold, attribute range set, and label pivot would produce the optimal output for this research, based on various criteria. Obviously, one vies for high ideal prediction accuracy, because it implies high actual prediction accuracy such that parametric combinations that yielded these were kept. Likewise, a secondary structural similarity measure that is greater is more desirable than one that is not, with more emphasis placed on those combinations that yielded high average *yes node* secondary structural similarity measures. This is because the nodes that belong to the cluster (i.e. “yes” nodes) are statistically more important.

Inverse to the other measures, it was decided that a *lower* node count (that is, the count of decisions made) would be more favorable. This is due to the fact that this research aims to find protein sequence motifs that transcend protein families. If the node count is too high, and approaches the number of proteins, this implies that each node represents approximately one protein. As each decision node, ideally, should represent a given motif among the proteins it represents, it makes no sense to have a system in which each node only represents one protein. This, in itself, implies higher entropy and fewer items in the attribute range list.

Finally, it was decided that those parameters that gave higher percentages of nodes that have 70% structural homology or above (see section 3.7), were ideal.

4.2 Parametric Results

Given the parameters, four distinct data sets were created from analyzing and averaging the appropriately weighted values from the 268,464 generated tuples. The graphs denoting these four data sets can be seen in the following figures. Ideal prediction accuracy, given by a red line refers to the measure described in section 3.9. Its value refers to the right y-axis. “Yes” node secondary structural similarity, given by a purple line, is exactly that, again referring to the right y-axis. Total secondary structural similarity, given by a green line, is the measure of all nodes’ secondary structure. It, too, refers to the right y-axis. Total node count, a light blue line, is simply the number of all decision nodes, and it refers to the left y-axis. Since high quality nodes are important, we also show the percentage of nodes with greater than 90% structural similarity, given by a gray-blue line. This, again, is given by the right y-axis. Finally, the “yes” node count, denoted by an orange line, just refers to the number of yes decision nodes.

As one can see in each of the four figures, an entropy threshold of 0.75 is marked by a vertical red on the graph, noting the various measures at that entropy. One might note that the percentage of nodes with a 90% structural similarity line falls sharply on all four graphs *after* an entropy of 0.75. One might also note that ideal prediction accuracy follows a similar trend, but to a much less severe degree, just as average *yes node* secondary structure similarity measure. An entropy threshold of 0.75 also falls in the mid-range of the average node count, implying that it would not yield data too far dichotomized, nor would it yield completely random output. Keeping in mind all criteria spelled out in section 4.1, it would appear that an entropy threshold of 0.75 is, indeed, the most appropriate for this research.

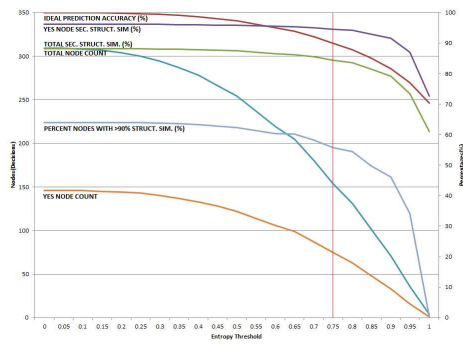


Figure 2 Seven Label Pivot, Large Attribute Range set

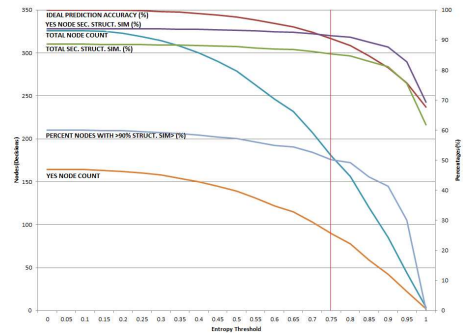


Figure 3 Six Label Pivot, Large Attribute Range Set

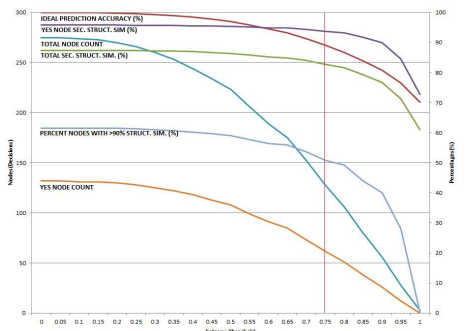


Figure 4 Seven Label Pivot, Reduced Attribute Range Set

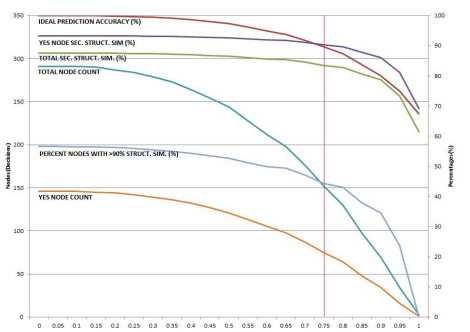


Figure 5 Six Label Pivot, Reduced Attribute Range Set

Parameters	>90%	90-80%	80-70%	70-60%	<60%
P7-R5	55.83%	15.56%	7.19%	9.18%	12.24%
P6-R5	50.27%	15.92%	13.05%	8.97%	11.80%
P7-R4	50.80%	17.38%	7.92%	10.00%	13.90%
P6-R4	44.39%	17.45%	14.57%	9.94%	13.65%

Table 1. Comparison of Decision Node Protein Secondary Structural Similarity Percentages.

To determine which label pivot and attribute range set is optimal, one can refer to the measures of nodal structural similarity percentages given in Table 1, which assumes our given entropy threshold of 0.75. In this table, P7 refers to a label pivot of seven, P6 refers to a label pivot of six, R5 refers to the large ($\{0-4, 5-7, 8-14, 15-29, 30-100\}$) attribute set, and alternatively, R4 refers to the small attribute range set. As one can see, taking only those percentages that refer to greater than 70% structural similarity (as, again, they can be considered structurally identical [15]), P6-R5 produces the best results, with P7-R4 producing the worst results. Note that while P6-R5 doesn't produce the optimal percentage of nodes with greater than 90% structural similarity, it does produce both the most over 70% and has the least percentage of nodes with less than 60% structural similarity. Taking in consideration other measures, such as node count and average yes node secondary structural similarity, P6-R5 consistently produces the most optimal output.

4.3 Example Decision Tree Result

Thus, given the parameters of a 0.75 entropy threshold, and the parametric combination denoted as P6-R5 (refer to section 4.2) one can produce a relatively simple decision tree to examine the effectiveness of the FGK-Decision Forest Model. The following figure examines a random file whose number of decisions was in the lower range, such that it could be easily displayed on paper. Note that this tree is not typical in that the average range for the node count with the given parameters is 150:

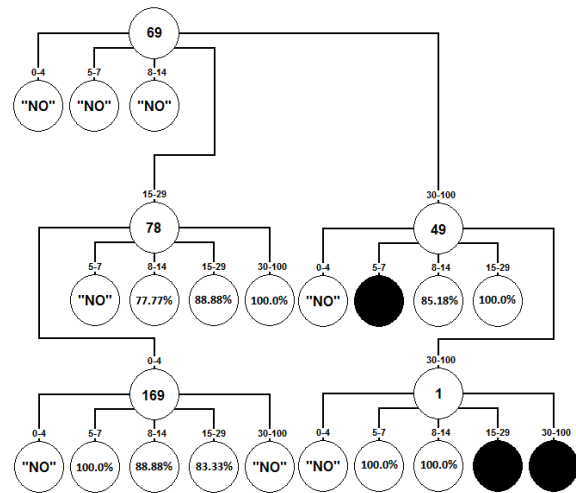


Figure 6 Granule 6-Cluster 93 Decision Tree

The method by which one would read Figure 6 is very simple. One starts at the top node, denoted here as '69,' and would work their way down to a given decision. The decision states whether or not a protein belongs to their FGK-Model generated clusters. The '69,' '169,' '78,' '1,' etc. are all dimensions for each protein generated by the sliding window technique. Each branch from each node denotes the attribute range that is used to further classify the data set. For instance, starting from the root node, '69,' if the frequency value of this dimension is between zero and four, then a "no" decision is made. In most cases, however, the decision tree must refer to

other dimensions and check their respective value before a decision can be made. The yes decision nodes are denoted as percentages, which detail the structural similarity of the proteins it describes. The black decision nodes denote a case in which no proteins in the training data could be represented by that particular path. These are interpreted as “no” decision nodes.

V. CONCLUSION

A newly proposed model, the FGK-Decision Forest Model, utilizes the data organizing prowess and robustness of granule computing granted by Fuzzy C-Means and Greedy K-Means clustering, and the clear and easily comprehensible analysis of the ID3 decision tree. Using this model, one splits the original protein data, generated by a sliding window technique, into various information granules of protein clusters via various iterations of Fuzzy C-Means and Greedy K-Means. Granted these clusters, a decision tree is generated for each. These decision trees each contain decisions that denote whether or not certain proteins belong to a given cluster. They also denote structural motifs, presented in the “yes” nodes of each tree. All of these decision trees come together to produce a “decision forest” in which one could potentially use to predict local tertiary structure by finding the decision tree and the motif contained therein that best fits the unknown protein, assuming paired tertiary structure data.

This paper focuses heavily on the parametric setup and analysis of the results of each. The three primary parameters tested were entropy, label pivot, and attribute range set. The entropy described the allowed randomness of the tree. It was set to 0.75, as it had the greatest tradeoff between all parametric criteria. The label was based on secondary structural similarity and the idea that 70% and greater secondary structural similarity was roughly identical. Two label pivots were tested, and a value of 6 was decided based on the quality analysis. The attribute range set was based on the frequency values produced by the sliding window technique. Two sets were tested, and the larger range set was used for its increased quality in regards to the parametric criteria.

A decision tree example is also shown, in which its usefulness for portraying clear and easily comprehensible analysis is examined. As each “yes” node denotes a structural motif, each “no” node denotes a set of proteins that need to be removed from the training data, and each black (that is, each node in which a decision was not generated) node denotes sections in which there are no structural motifs, it is clear that the decision tree is a promising method, at least graphically, for portraying protein data. Also, each decision tree can be used, without modification, to decide whether or not a protein belongs to the cluster represented by the protein, and with an associated prediction accuracy. This implies that the decision forest, as stated previously, can be used to generate local tertiary structural predictions with measurably accurate decisions.

While further development and research is needed to expand the flexibility and applicability of this model, it should be clear that it has potential to be adapted due to its promising robustness and efficiency, as well as the relative ease of

comprehending its output, such that its analysis is not constrained to one field. With our proposed expansions on the original implementation, we believe this model will be used widely for the above mentioned reasons.

ACKNOWLEDGMENT

The work of Bernard Chen was supported in part by UCA's University Research Council (URC) and Summer Stipend. The work of D. Che was partially supported by President Research Fund at East Stroudsburg University of Pennsylvania.

REFERENCES

- [1] J.M. Chandonia, S.E. Brenner, “The impact of structural geonomics: expectations and outcomes,” *Science*, vol. 311, pp. 347-351, 2006.
- [2] A. L. Spek, “Structure validation in chemical crystallography,” *Acta Crystallographica*, Section D, vol. 60, no. 4, pp. 148-155, 2004.
- [3] G. Karp, *Cell and Molecular Biology: Concepts and Experiments*, 6th ed., New York: John Wiley & Sons Inc, 2009, pp. 52-66.
- [4] N. Hulo, C.J.A.V. Sigrist, L. Saux, P.S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. de Castro, P. Bucher, A. Bairoch, “Recent improvements to the PROSITE database,” *Nucleic Acids Res.*, vol. 32, 2004.
- [5] T.K. Attwood, M. Blythe, D.R. Flower, A. Gaulton, J.E. Mabey, N. Naudling, L. McGregor, A. Mitchell, G.M. Paine, R. Scordis, “PRINTS and PRINTS-S shed light on protein ancestry,” *Nucleic Acid Res.*, vol. 30, no. 1, pp. 239-241, 2002.
- [6] S. Henikoff, J.G. Henikoff, S. Pietrokovski, “Blocks+: a non redundant database of protein alignment blocks derived from multiple compilation,” *Bioinformatics*, vol. 15, no. 6, pp. 417-479, 1999.
- [7] O. Carugo, “Rapid Methods for Comparing Protein Structures and Scanning Structure Database,” *Current Bioinformatics*, vol. 1, pp. 75-83, 2006.
- [8] K.F. Han, D. Baker, “Global properties of the mapping between local amino acid sequence and local structure in proteins,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 12, pp. 5814-5818, 1996.
- [9] K.F. Han, D. Baker, “Recurring local sequence motifs in proteins,” *Journal of Molecular Biology*, vol. 251, no. 1, pp. 176-187, 1995.
- [10] C. Bystroff, V. Thorsson, D. Baker, “HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins,” *Journal of Molecular Biology*, vol. 301, pp. 173-190, 2000.
- [11] B. Chen, P.C. Tai, R. Harrison, Y. Pan, “FGK model: An Efficient Granular Computing Model for Protein Sequence Motifs Information Discovery,” *Proceedings of IASTED CASB Dallas*, pp. 56-61, 2006.
- [12] B. Chen, M. Johnson, “Protein Local 3D Structure Prediction by Super Granule Support Vector Machines (Super GSVM),” *Proceedings of BMC Bioinformatics*, vol. 10, 2009.
- [13] S. R. Safavian, D. Landgrebe, “A Survey of Decision Tree Classifier Methodology,” *IEEE Trans. Systems, Man and Cybernetics*, vol. 21, no. 3, pp. 660-674, 1991.
- [14] J.R. Quinlan, “Induction of Decision Trees,” *Machine Learning*, vol. 1, pp. 81-106, 1986.
- [15] W. Zhong, G. Altun, R. Harrison, P. C. Tai, Yi. Pan, “Improve K-Means Clustering algorithm for Exploring Local Protein Sequence motifs Representing Common Structural Property,” *IEEE transactions on Nanobioscience*, vol. 14, no. 3, pp. 255-265, 2005.
- [16] T.Y. Lin, “Data Mining and Machine Oriented Modeling: A Granular Computing Approach,” *Journal of Applied Intelligence*, vol. 13, no. 2, pp. 113-124, 2002.
- [17] Y.Y. Yao, “On Modeling data mining with granular computing,” *Proceedings of COMPSAC 2001*, pp. 638-643, 2001.
- [18] G. Wang, R. L. Dunbrack, Jr., “PISCES: a protein sequence culling server,” *Bioinformatics*, vol. 19, no. 12, pp. 1589-1591, 2003.
- [19] C. Sander, R. Schneider, “Database of similarity derived protein structures and the structure meaning of sequence alignment,” *Proteins: Struct. Funct. Genet.*, vol. 9, no. 1, pp. 56-68, 1991.
- [20] W. Kabsch, C. Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, pp. 2577-2637, 1983.

A MATLAB TOOLBOX FOR DATA REDUCTION, VISUALIZATION, CLASSIFICATION AND KNOWLEDGE EXTRACTION OF COMPLEX BIOLOGICAL DATA

A. Mohammad-Djafari*, G. Khodabandelou† and J. Lapuyade-Lahorgue ‡

Laboratoire des signaux et systèmes (L2S)
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
plateau de Moulon, 3 rue Joliot-Curie, 91192 GIF-SUR-YVETTE Cedex, France

ABSTRACT

In this paper, first we present A Matlab toolbox which gives the possibility to simulate the data for testing the algorithms such as: Principal Component Analysis (PCA), Factor Analysis (FA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA) and many other classification methods which can be used in Data Reduction (DR), Data Visualization (DV), supervised and unsupervised classification of multivariate great dimensional biological data. Then, we describe some biological experiments related to studying the circadian cell cycles and cancer treatment where the biologists observe different kind of data such as the variations of temperature, activity, hormones, genes and proteins expressions. These data are often complex: multivariate, great dimensionality, heterogeneous, with missing data, and observed at different sampling rates. The classical methods of PCA, FA, ICA and LDA can not directly handle these data. In this paper, we show how this toolbox can help them to visualize, to analyse and to do classifications on these data and finally to extract some knowledge from them.

Keywords: Data visualization, Dimensionality reduction, Principal Component Analysis, Factor Analysis, Independent Component Analysis, Linear Discriminant Analysis, Bayesian inference, Sources separation, Inverse problems.

1. INTRODUCTION

In many biological experiments, we are always face to data sets which are heterogeneous, of great dimensionality with missing and outliers data. To understand these data, first we need to visualize them, but the great dimensionality of these data needs a Data Reduction (DR) step. Principal Component Analysis (PCA), Factor Analysis (FA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA) methods are the main classical methods for analyzing high dimensional data [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. PCA, FA and ICA are mainly used for dimensionality reduction and LDA for supervised classification. Even if these methods are well defined, still there exist different algorithms for their practical usage: PCA and FA are the most stable ones because they use quadratic criteria and L2 norms (second order statistics in statistical interpretation and Gaussian hypothesis in probabilistic interpretation) and so they are very simple to implement. The characteristics of the results

obtained by PCA and FA are well known. For example, we know that the factors are obtained up to a rotation indetermination. ICA is more complex because the criteria used to be optimized are often non quadratic (Kullback-Leibler divergence) and use higher order statistics (HOS) and non Gaussian probability laws. The corresponding algorithms are then more sophisticated. However the common properties of independent components are that they are obtained up to a permutation and scale factor indetermination. LDA can be considered as a particular supervised classification method where we know the number of classes.

In this paper, in a first step, we present, very shortly, but in a unifying way of forward and inverse problem, different multivariate data analysis tools. Then, we present a Matlab toolbox: to generate different factors with different properties; to generate different data sets with linear or non linear dependencies; to add different kind of errors; to apply different algorithms of PCA, FA, ICA, LDA, ... and to compare the obtained results. In a second step, we show some preliminary results for real data set obtained by biologists working on circadian and cell cycle influence on cancer. This work is done in collaboration within the European project EraSysBio.

2. A UNIFYING PRESENTATION OF MULTIVARIATE DATA ANALYSIS METHODS THROUGH FORWARD AND INVERSE MODELING

PCA, FA, ICA and LDA are classical methods of dimensionality reduction and data analysis. Due to the origin of these methods, there have been many different presentations and interpretations. Here, we present them in an unifying context of forward modeling and inversion. To do this, we start by defining the factors $\mathbf{f}(t) = [f_1(t), \dots, f_N(t)]$ which is an N -dimensional vector of time series. Here, we choosed time series due to our final application. However, the time index can be anything else, for example, just the index of experiments of a position on a line, in a plane or in space.

In a first step, we assume that the observed data $\mathbf{g}(t) = [g_1(t), \dots, g_M(t)]$ are obtained via a mixing (or loading) matrix \mathbf{A} of dimensions $[M \times N]$ through the forward model

$$\mathbf{f}(t) \longrightarrow \boxed{\text{Forward model } \mathbf{A}} \longrightarrow \oplus \xrightarrow{\downarrow \epsilon} \mathbf{g}(t) = \mathbf{A} \mathbf{f}(t) + \epsilon(t), t = 1, \dots, T \quad (1)$$

where ϵ represents the errors of modeling and T is the total number of observed samples.

Using this forward model, the objective of many data analysis methods such as PCA, FA, ICA and LDA is to obtain the factor \mathbf{f} and the loading matrix \mathbf{A} . Described as such, we see that this estimation problem is very ill-posed in the sense that we can find many

Senior Researcher at Laboratoire des signaux et systmes (L2S)

PhD candidate at Laboratoire des signaux et systmes (L2S)

Post-doc at Laboratoire des signaux et systmes (L2S)

This work is a part of ERASYSBIO-C5Sys European project "Circadian and cell cycle clock systems in cancer": <http://www.erasysbio.net/index.php?index=272>

combinations of factors and loading matrix which can satisfy this model. In the following, we use this model to explain the differences between PCA, FA, ICA and LDA.

PCA and FA methods try to find uncorrelated factors \hat{f} . Because correlation describes a linear dependence, the main assumption is then that \hat{f} has to be obtained through a linear combination of the data: $\hat{f}(t) = \hat{B}g(t)$, where the matrix B is called separating (or demixing or deloading) matrix.

$$g(t) \rightarrow \begin{cases} \text{Inference} \\ \text{PCA, FA, ICA} \\ \text{LDA, Bayes} \end{cases} \begin{matrix} \rightarrow \hat{A} \text{ or } \hat{B} \\ \rightarrow \hat{f}(t) = \hat{B}g(t) \end{matrix} \quad (2)$$

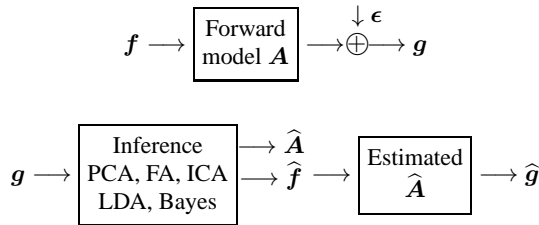
Here, we are not going to describe these algorithms which are described elsewhere in details [11, 12, 13, 14, 15], but we present a Matlab toolbox in which we implemented all these methods.

3. PRESENTATION OF THE MATLAB SIMULATION TOOLBOX

We have developed a menu driven simulation tool, which has, as the main menu, the following steps:

- Generation of different sources (factors) with different properties (Uniform, Gaussian, Mixture of Gaussian, ... ,
- Generation of different data sets with linear or nonlinear dependencies,
- Addition of different kind of errors,
- Application of different algorithms of PCA, FA, ICA, LDA, ... and
- Visualization and evaluation tools which give possibility to evaluate the performances of a given method or to compare the results obtained by two different methods.

As tools to measure the performances of these methods, we propose the following scheme:



and then compare \hat{g} with g , \hat{f} with f , \hat{A} with A , ...

As an example of using this simulation tool, we show here a complete set of figures detailing the different steps of simulation and inversion. Figure 1 shows an example of two sources f (generated via a mixture of two Gaussian model) and five data set g obtained via a mixing matrix A and addition of some noise ϵ using the forward model $g = Af + \epsilon$ and then the results obtained by FA and ICA.

As a second example, we show in Figure 2 two sources generated via a mixture of two Gaussian model. We then again used these sources to generate the data and applied different methods of PCA, FA, ICA (without using the class information) and LDA with using the class information.

As a third example, we show in Figure 3 two sources generated via a mixture of two uniforms model. We then again used these sources to generate the data and applied different methods of PCA, FA, ICA (without using the class information) and LDA with using the class information.

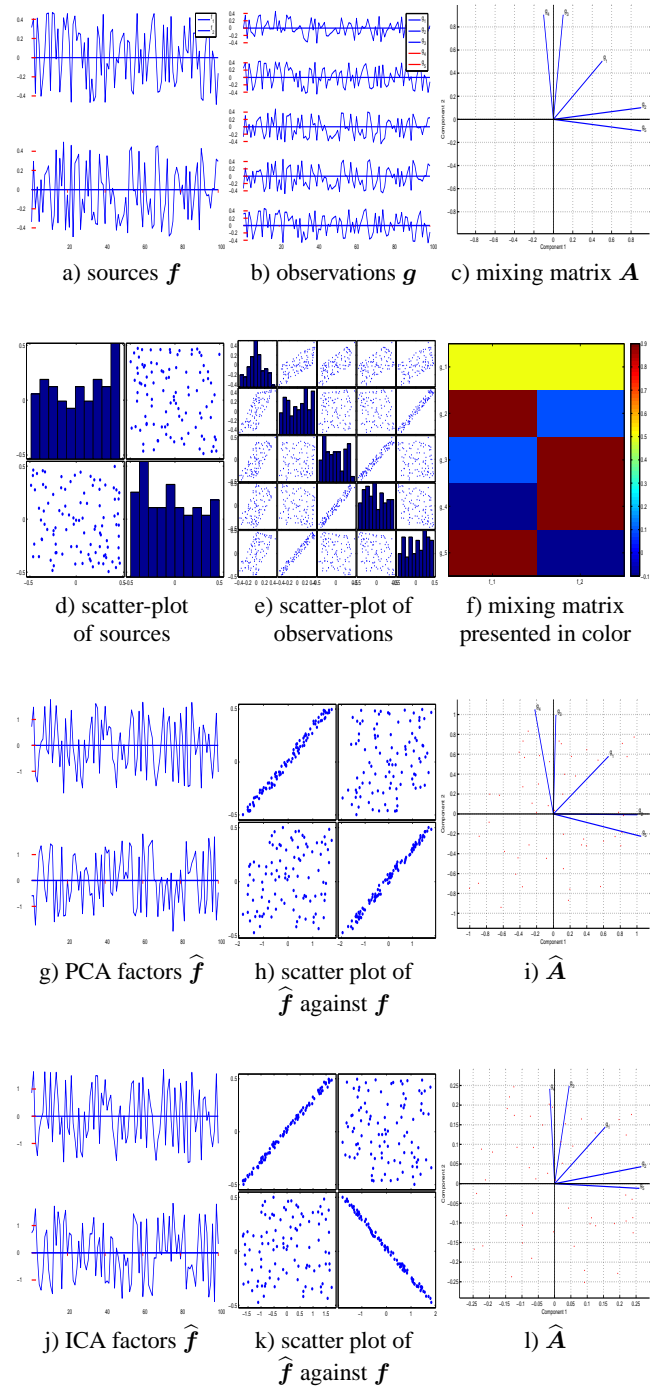


Fig. 1. Simulation of 2 sources f and 5 observations g with $T = 100$ samples: a) sources f , b) observations g , c) representation of the mixing matrix A , d) scatter-plots of the sources, e) scatter-plots of the observations, f) color presentation of the mixing matrix, g) PCA factors \hat{f} , h) scatter-plot of \hat{f} against f , i) representation of the estimated mixing matrix \hat{A} , j) ICA factors \hat{f} , k) scatter-plot of \hat{f} against f , l) representation of the estimated mixing matrix \hat{A} .

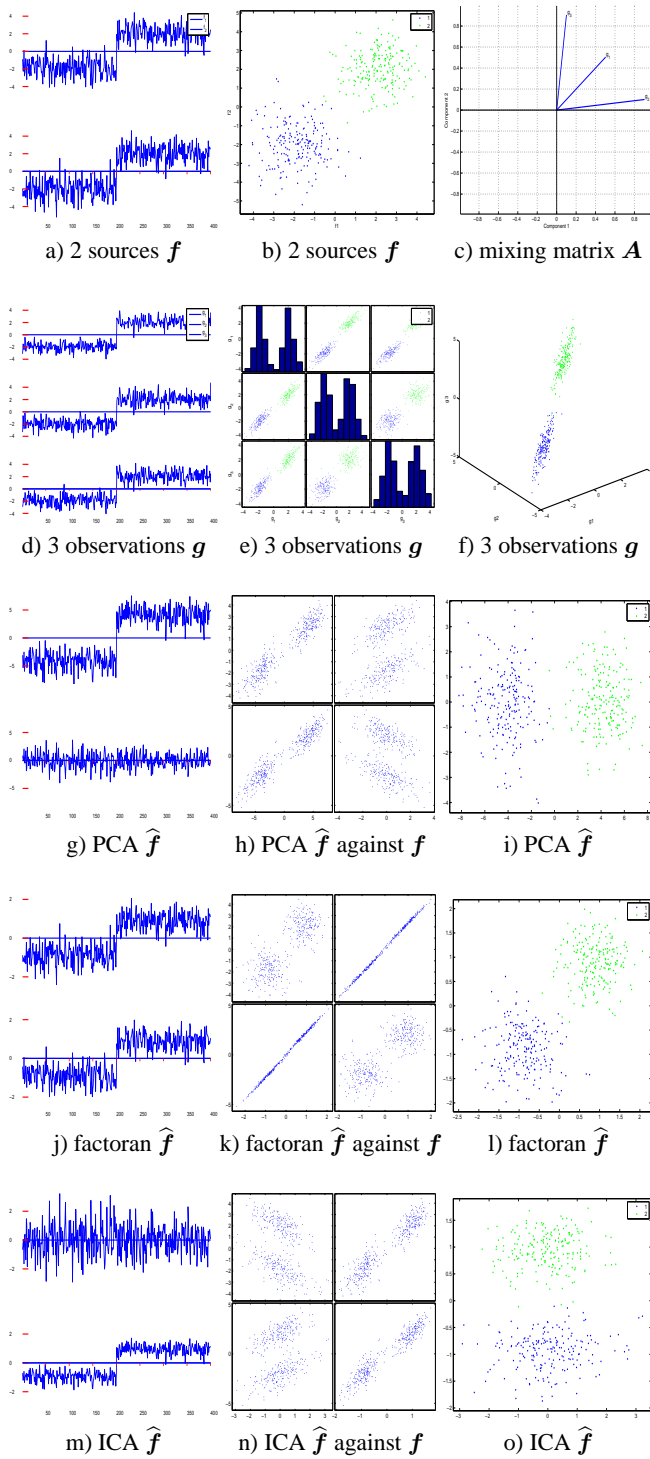


Fig. 2. Simulation of 2 sources (mixture of two Gaussian distributions) f and 3 observations g with $T = 400$ samples: a) sources f , b) observations g , c) representation of the mixing matrix A , d) scatter-plots of the sources, e) scatter-plots of the observations, f) spatial structure of the 3 sources, g) PCA factors \hat{f} , h) scatter-plot of \hat{f} against f and i) spatial structure of the PCA factors \hat{f} , j,k,l) the same with FA, m,n,o) the same with ICA.

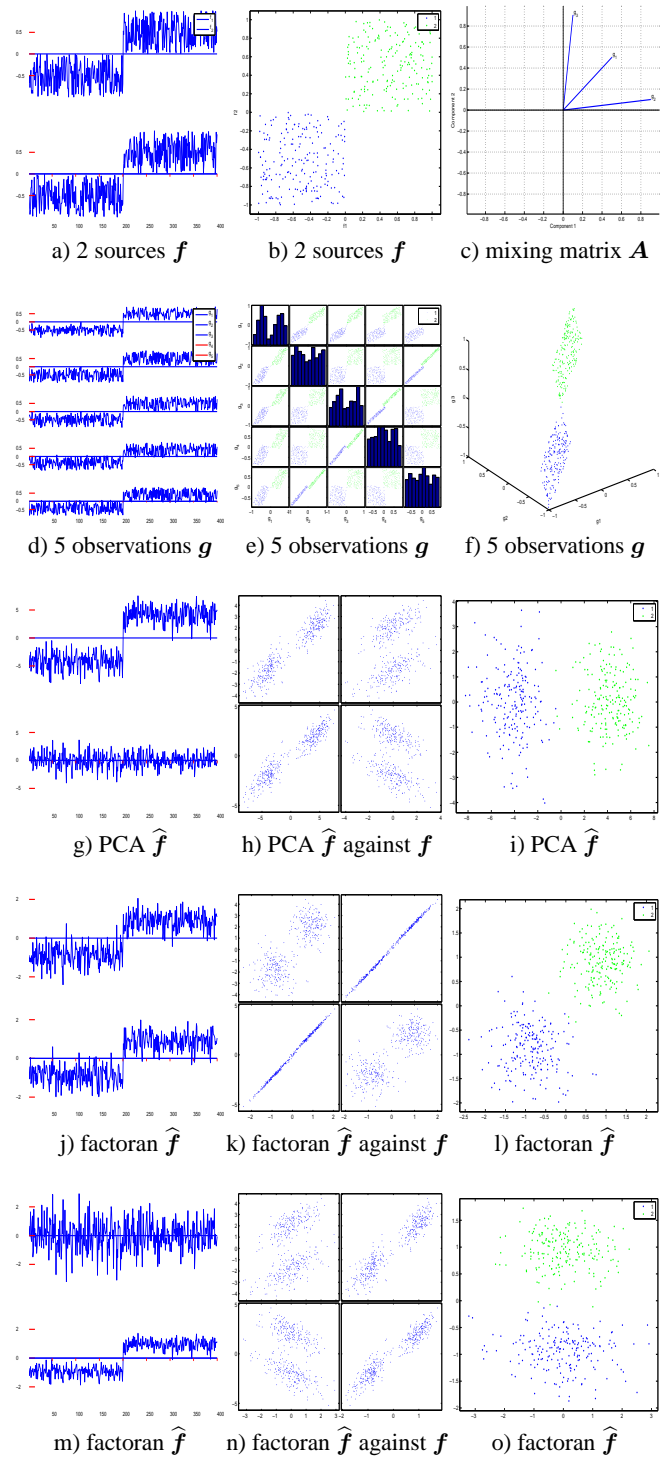


Fig. 3. Simulation of 2 sources (mixture of two uniform distributions) f and 3 observations g with $T = 400$ samples: a) sources f , b) observations g , c) representation of the mixing matrix A , d) scatter-plots of the sources, e) scatter-plots of the observations, f) spatial structure of the 3 sources, g) PCA factors \hat{f} , h) scatter-plot of \hat{f} against f and i) spatial structure of the PCA factors \hat{f} , j,k,l) the same with FA, m,n,o) the same with ICA.

4. APPLICATION ON REAL DATA

As we mentioned, we developed these tools for analyzing some biological data in relation with circadian cell cycle and evolution of cancer tumors in the context of the European project ERASYSBIO. A great number of experimentations have been done on mice. As an example, different quantities such as Temperature, Activity, different Hormones, different Genes expressions and different Proteins are measured during one or a few days and one of the problems addressed is finding the principal components or factors of some of these data.

In Figure 4, we show an example of such analysis on Gene expressions time series.

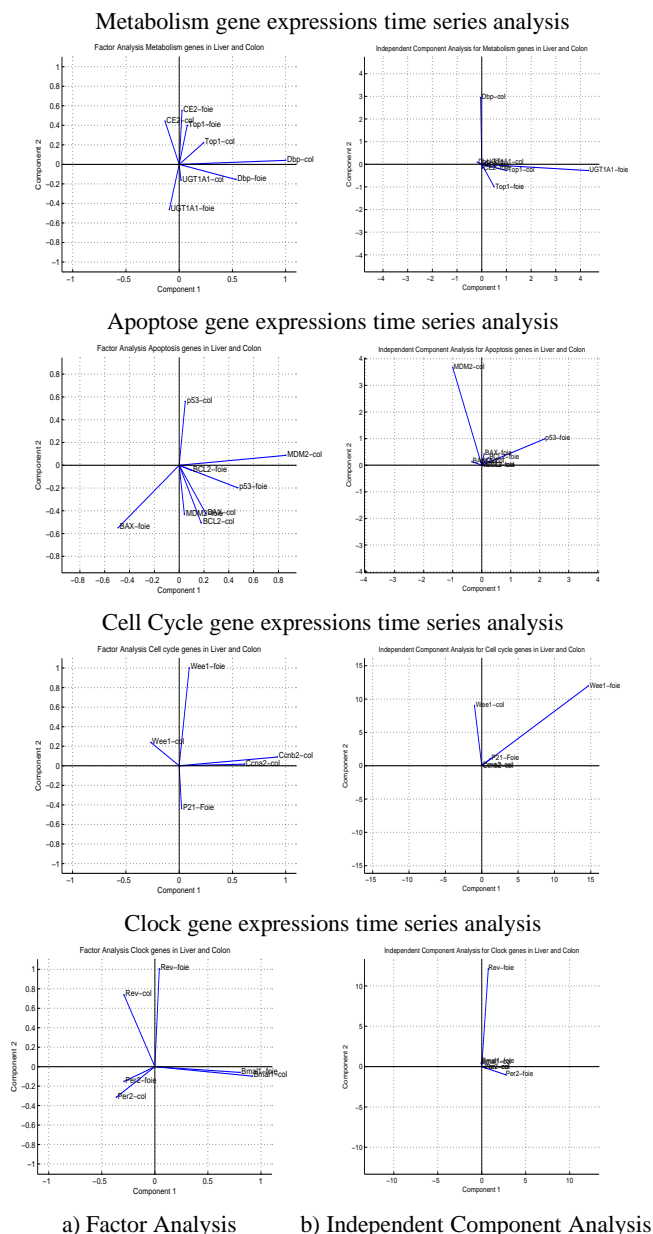
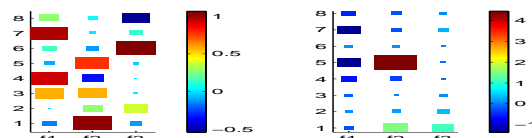


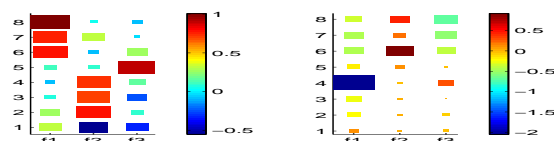
Fig. 4. A comparison of FA and ICA on three sets of gene expression data. These results are obtained with two factors.

For now, we just applied these methods directly on the time series data without accounting for time structure which is very important. However, the results obtained seem to have some significant importance for biologists. Here, we assumed only two factors. As we can see it seems that there is a need to increase the number of factors.

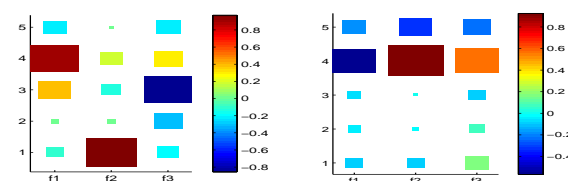
Metabolism gene expressions time series analysis



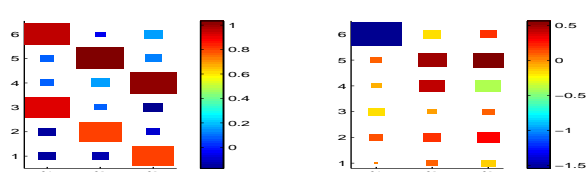
Apoptose gene expressions time series analysis



Cell Cycle gene expressions time series analysis



Clock gene expressions time series analysis



a) Factor Analysis b) Independent Component Analysis

Fig. 5. A comparison of FA and ICA on three sets of gene expression data. These results are obtained with three factors. Here, we used a different presentation of the loading matrix which is more appropriate for the cases where the number of factors is greater than two. This presentation is called Hinton where the values of the matrix are coded by color and by size of the patches.

In Figure 5, we show the same results with three factors. However, when the number of factors is greater than two, it is no more easy to represent them as bi-plot graphs of Figure 4. Here, we use a different presentation of the loading matrix which is more appropriate for the cases where the number of factors are greater than two. This presentation is called Hinton [16, 17] where the values of the

matrix are coded by color and by size of the patches.

In Figure 6, we show two results of Linear Discriminant Analysis on 14 genes expressions in Colon and 13 genes expressions in liver. As we can see, here two factors are enough to discriminate the three classes of mice. On this figure, at left, we see this discrimination and at right the weights of these genes in these two factors.

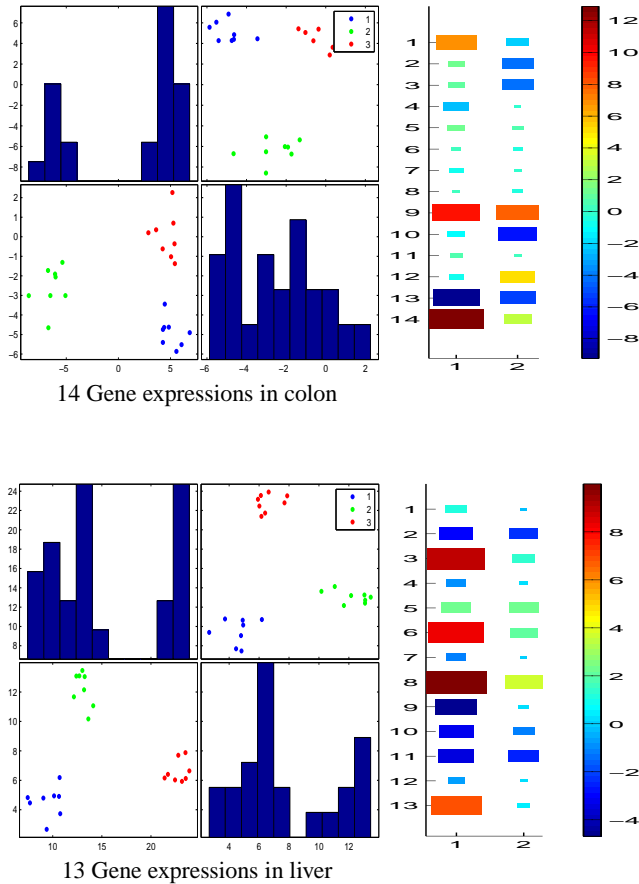


Fig. 6. Discriminant Analysis on real mice data: 13 genes expressions in liver have been used. Two factors have been enough to discriminate the three classes of mice (left). The weights of these genes in these two factors are shown on the right.

The main difficulties in these data are: great dimensionality (more than fifty), non-homogeneity (Temperature, Activity, Hormones, Genes, Proteins), presence of outliers data, missing data and lack of synchronization (for example, temperature is measured every 15 minutes but Genes expressions every 3 hours). We need to adapt these methods to account for all these difficulties. We are working on these difficulties and will report soon in details on them.

5. CONCLUSIONS

In this paper, first we introduced a unifying presentation of many classical data analysis methods such as PCA, FA and ICA based on forward modeling and inversion. This unifying presentation facilitates the comprehension of these different methods. We then presented a simulation Matlab toolbox which has the possibilities of generating sources and observations, doing FA, PCA and ICA and evaluating the performances of the proposed methods. Finally, we

used these tools for analyzing some biological data which seems giving important information, or at least confirm their intuition on the role of different quantities. We are still exploring these tools for the real application of biological data where we have to adapt more particularly these tools for the situations where:

- we have fewer number of data compared to the number of variables;
- the estimated covariance matrix of the data is not positive definite;
- the data are inhomogeneous;
- the data have different sampling rate;
- there are some non-observed values (missing data);
- there are outliers in the observed data (for example, measured temperature greater than 44 or less then 35, etc.).

6. PERSPECTIVES

When analyzing these biological data, the main questions we need to answer can be summarized as follows:

Variable section: One of the main questions asked very often is: If we had to redo other experiences, which ones of these quantities are the most importances to observe again. This is a very difficult question. The answer depends on the type of information we need to extract. Very often the quantities we have observed are linked (correlated or dependent). So, any selection of subset of variables causes, in some sense, loss of information. So, this question, very often, cannot be answered directly. We need modeling, the link between variables directly or in a transformed space, dimension reduction, clustering and classification, etc. Here are a few references concerning this subject [18, 19, 20, 21, 22]

Dimension reduction and Factor analysis: The second question is: Can we express the information content of all these data in a fewer set of factors or components? The main classical tools here are PCA, FA and ICA. One of the difficulties in these tools is the determination of the number of factors which is still an open problem [23, 9, 7]. When the number of factors is fixed, then these tools can be used easily. However, one of the drawbacks of these tools is the interpretation of the factors or components. Modeling the problem as an inverse problem of sources separation and using the the Bayesian approach are the promising tools to push farther these limitations [24, 25, 23, 26, 27, 28, 9, 29].

Discriminant Analysis:

Very often the observed data comes from different classes of subjects (male/female, healthy/Tumor,...) and we know the classes. In these cases, another question which arises is: Which of these variables or factors are the most discriminant between classes? Here are a few references concerning this subject [18, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39].

Clustering and classification:

Some times, in opposite of the previous case, we have only the data and we are asked to group or cluster them. This is also called *totally unsupervised classification*. In some other cases, we may know the number of classes and even the characteristics of each one of classes. The question is then to classify a given new observation. This is called *totally supervised classification*. When the number of classes is known, but the characteristics of each classe has to be *learned* from a *training set* of observations, then the problem is called *semi-supervised classification*. The estimation of number of classes is

related to *model selection*. Here are a few references concerning this subject [40, 33, 41, 8]

Graph of links and dependencies between variables:

One of the main steps of Knowledge extraction in studying biological data is producing a graph of dependencies between variables. To obtain such a graph we need to decide if two variables are dependent or not. We need then measures of dependencies to discover these dependencies [42, 43, 44, 45, 46, 47, 48]. One of the classical and most used is the Pearson's correlation ρ . When $|\rho|$ is near to one, we say that the two variables are dependent. However, when $|\rho|$ is near to zero or even zero, this does not mean that the two variables are independent. Indeed, $|\rho|$ measures only the linear dependence between those two variables. There are many other measures of dependencies that we can use which are more appropriate. For example, we use the Spearman's ρ_s and the Kendall τ jointly with Pearson's correlation ρ .

Graph of oriented dependencies between variables and causality: One of the last steps of Knowledge extraction in studying biological data is studying the oriented graph or causality [49, 50, 51]

7. REFERENCES

- [1] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, no. 1, pp. 113–127, 1994.
- [2] Pierre Comon, "Independent Component Analysis, a new concept ?," *Signal processing, Special issue on Higher-Order Statistics, Elsevier*, vol. 36 (3), pp. 287–314, Apr. 1994.
- [3] D. J. C. MacKay, "Maximum likelihood and covariant algorithms for independent component analysis," Tech. Rep., University of Cambridge, Cavindish Laboratory, Cambridge, UK, 1996.
- [4] K. Knuth, "Bayesian source separation and localization," in *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems, San Diego, CA*, A. Mohammad-Djafari, Ed., July 1998, pp. 147–158.
- [5] S. J. Roberts, "Independent component analysis: Source assessment, and separation, a Bayesian approach," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 145, no. 3, 1998.
- [6] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.
- [7] Ma Yi, P. Niyogi, G. Sapiro, and R. Vidal, "Dimensionality reduction via subspace and submanifold learning [from the guest editors]," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 14–126, march 2011.
- [8] R. Vidal, "Subspace clustering," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 52–68, march 2011.
- [9] K.M. Carter, R. Raich, W.G. Finn, and A.O. Hero, "Information-geometric dimensionality reduction," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 89–99, march 2011.
- [10] L. Carin, R.G. Baraniuk, V. Cevher, D. Dunson, M.I. Jordan, G. Sapiro, and M.B. Wakin, "Learning low-dimensional signal models," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 39–51, march 2011.
- [11] Y. Koren and L. Carmel, "Robust linear dimensionality reduction," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 10, no. 4, pp. 459–470, 2004.
- [12] A. Sharma and K.K. Paliwal, "Rotational linear discriminant analysis technique for dimensionality reduction," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 10, pp. 1336–1347, 2008.
- [13] Jing Peng, Peng Zhang, and N. Riedel, "Discriminant learning analysis," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 38, no. 6, pp. 1614–1625, 2008.
- [14] P. Chaudhuri, A.K. Ghosh, and H. Oja, "Classification based on hybridization of parametric and nonparametric classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1153–1164, 2009.
- [15] Taiping Zhang, Bin Fang, Yuan Yan Tang, Zhaowei Shang, and Bin Xu, "Generalized discriminant analysis: A matrix exponential approach," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 1, pp. 186–197, 2010.
- [16] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.
- [17] G.E. Hinton and R.R. Salakhutdinov, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, pp. 428–434, 2007.
- [18] J. Brezmes, P. Cabre, S. Rojo, E. Llobet, X. Vilanova, and X. Correig, "Discrimination between different samples of olive oil using variable selection techniques and modified fuzzy artmap neural networks," *Sensors Journal, IEEE*, vol. 5, no. 3, pp. 463–470, june 2005.
- [19] C. Fevotte and S.J. Godsill, "Sparse linear regression in unions of bases via bayesian variable selection," *Signal Processing Letters, IEEE*, vol. 13, no. 7, pp. 441–444, july 2006.
- [20] T. Trappenberg, J. Ouyang, and A. Back, "Input variable selection: mutual information and linear mixing measures," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 1, pp. 37–46, jan. 2006.
- [21] J.-J. Fuchs and S. Maria, "A new approach to variable selection using the tls approach," *Signal Processing, IEEE Transactions on*, vol. 55, no. 1, pp. 10–19, jan. 2007.
- [22] Lu Chuan, A. Devos, J.A.K. Suykens, C. Arus, and S. Van Huffel, "Bagging linear sparse bayesian learning models for variable selection in cancer diagnosis," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 11, no. 3, pp. 338–347, may 2007.
- [23] Farahmand A., M., Szepesvári C., and Audibert J.-Y., "Manifold-Adaptive Dimension Estimation," Proceedings of the 24th International Conference on Machine Learning, 2007.
- [24] A. Mohammad-Djafari, "Séparation de sources," in *Approche bayésienne en séparation de sources*, A. Mohammad-Djafari, Ed., Paris, 2006, Traité IC2, Série traitement du signal et de l'image, Hermès, (P. Common et Ch. Jutten ed.).
- [25] D.P. Wipf and B.D. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3704–3716, july 2007.

- [26] Erik G. Larsson and Yngve Selen, "Linear regression with a sparse parameter vector," *Signal Processing, IEEE Transactions on*, vol. 55, no. 2, pp. 451–460, feb. 2007.
- [27] E. Diederichs, A. Juditsky, V. Spokoiny, and C. Schutte, "Sparse non-gaussian component analysis," *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 3033–3047, june 2010.
- [28] Mohammad-Djafari Ali and Knuth K.H., "Bayesian approaches," in *Handbook of Blind Source Separation*, Pierre Comon and Christian Jutten, Eds., Elsevier Ltd, 2010, Academic Press.
- [29] J. Lapuyade and A. Mohammad-Djafari, "Nearest neighbors and correlation dimension for dimensionality estimation. application to factor analysis of real biological time series data," in *19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Michel Verleysen, Ed. 2011, ESANN 2011 Proceedings.
- [30] Yijuan Lu, Qi Tian, M. Sanchez, J. Neary, Feng Liu, and Yufeng Wang, "Learning microarray gene expression data by hybrid discriminant analysis," *Multimedia, IEEE*, vol. 14, no. 4, pp. 22–31, oct.-dec. 2007.
- [31] Jian Yang, A.F. Frangi, Jing-Yu Yang, David Zhang, and Zhong Jin, "Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 2, pp. 230–244, feb. 2005.
- [32] Taiping Zhang, Bin Fang, Yuan Yan Tang, Zhaowei Shang, and Bin Xu, "Generalized discriminant analysis: A matrix exponential approach," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 1, pp. 186–197, feb. 2010.
- [33] Pi-Fuei Hsieh, Deng-Shiang Wang, and Chia-Wei Hsu, "A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 223–235, feb. 2006.
- [34] Xuelian Yu, Xuegang Wang, and Benyong Liu, "A direct kernel uncorrelated discriminant analysis algorithm," *Signal Processing Letters, IEEE*, vol. 14, no. 10, pp. 742–745, oct. 2007.
- [35] Juwei Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 117–126, jan 2003.
- [36] T. Kurita, K. Watanabe, and N. Otsu, "Logistic discriminant analysis," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, oct. 2009, pp. 2167–2172.
- [37] Jian Yang and Chengjun Liu, "Horizontal and vertical 2dpcba-based discriminant analysis for face verification on a large-scale database," *Information Forensics and Security, IEEE Transactions on*, vol. 2, no. 4, pp. 781–792, dec. 2007.
- [38] Chein-I Chang and Baohong Ji, "Fisher's linear spectral mixture analysis," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 8, pp. 2292–2304, aug. 2006.
- [39] Changyou Chen, Junping Zhang, and R. Fleischer, "Distance approximating dimension reduction of riemannian manifolds," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 1, pp. 208–217, feb. 2010.
- [40] Nadler B., Lafon S., Coifman R.R., and Kevrekidis I.G., "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators," *Advances in Neural Information Processing Systems*, vol. 18, pp. 955–962, 2005.
- [41] Tian Lan and D. Erdogmus, "Local linear ica for mutual information estimation in feature selection," in *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, sept. 2005, pp. 3–8.
- [42] G. Qu, S. Hariri, and M. Yousif, "A new dependency and correlation analysis for features," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 9, pp. 1199–1207, sept. 2005.
- [43] F. Chin and H.C. Leung, "Dna motif representation with nucleotide dependency," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 5, no. 1, pp. 110–119, jan.-march 2008.
- [44] Deng Cai, Xiaofei He, and Jiawei Han, "Srda: An efficient algorithm for large-scale discriminant analysis," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 1, pp. 1–12, jan. 2008.
- [45] P.C.H. Ma and K.C.C. Chan, "Inferring gene regulatory networks from expression data by discovering fuzzy dependency relationships," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 2, pp. 455–465, april 2008.
- [46] S. Zafeiriou and I. Pitas, "Discriminant graph structures for facial expression recognition," *Multimedia, IEEE Transactions on*, vol. 10, no. 8, pp. 1528–1540, dec. 2008.
- [47] L. Yu, A. Mishra, and S. Ramaswamy, "Component co-evolution and component dependency: speculations and verifications," *Software, IET*, vol. 4, no. 4, pp. 252–267, august 2010.
- [48] Fan Wenfei, F. Geerts, Jianzhong Li, and Ming Xiong, "Discovering conditional functional dependencies," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 5, pp. 683–698, may 2011.
- [49] D.A. Bell, "From data properties to evidence," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 5, no. 6, pp. 965–969, dec 1993.
- [50] M.L. Raymer, T.E. Doom, L.A. Kuhn, and W.F. Punch, "Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 33, no. 5, pp. 802–813, oct. 2003.
- [51] Ghim-Eng Yap, Ah-Hwee Tan, and Hwee-Hwa Pang, "Discovering and exploiting causal dependencies for robust mobile context-aware recommenders," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 7, pp. 977–992, july 2007.

Portable measurement system of the spasticity based on the K-means clustering algorithm of the tonic stretch reflex threshold

C.G. Song¹, K.S. Kim¹, M.H. Kim¹ and S.H. Ryu¹

¹Department of Electronic Engineering, Chonbuk National University, Korea

Abstract - Conventional clinical scales used for the evaluation of the spasticity have some limitations, and their reliability remains controversial. The aim of this study is to develop a portable spasticity measurement system for quantifying the grade of spasticity based on the K-means clustering algorithm of the tonic stretch reflex threshold (TSRT). Fifteen stroke patients (age: 63.5 ± 15.6) participated in the study. As a result, there was a strong negative correlation ($r = -0.95$, $r^2 = 0.90$, $p < 0.05$) between the spasticity level and TSRTs. This result showed that our system could be made clinically available for the more reliable discrimination of the spasticity level, compared to conventional scales..

Keywords: spasticity, tonic stretch reflex threshold

1 Introduction

Spasticity is a major source of disability caused by central nerve injuries such as stroke. It is most commonly defined as a motor disorder characterized by a velocity-dependent increase in the muscle tone with exaggerated tendon jerks, resulting from hyper-excitability of the stretch reflex (SR), as one component of the upper motor neuron syndrome [1,2]. Spasticity is the manifestation of a lesion of the supraspinal motor pathways and is caused by adaptive changes in transmission in the spinal networks distal to a lesion of the descending motor pathways. Clinically, this implies increased muscle tone, enhanced tendon reflexes, involuntary reflex zones and clonus. The measurement of the spasticity is a difficult and unresolved problem, partly due to its complex and multi-factional nature. The previous methods of quantifying or qualifying the spasticity are based on clinical scales or the biomechanical and neurophysiological analysis of the limb resistance to passive or voluntary movements.

The clinical scales employed are the Ashworth scale (AS), modified Ashworth scale (MAS), Tardieu scale, composite spasticity index (CSI) and spasm frequency scale [3-7]. Among these scales, the MAS has been widely used in the clinical field since they are simple, easy to use and require no instrumentation. It rates the subjective impression of the evaluator of the amount of resistance felt during stretch of the

relaxed muscle. However, the amount of resistance felt results from the net EMG activity in the muscle without consideration of the velocity-dependence of the response, thus, the MAS measurement is a disagreement with Lance's definition [1]. Therefore, their inter- and intra-rater reliability remains controversial, because the scores are obtained based on the subjective feeling of the rater, such as the observation of the catch and spasm of the muscle, and largely rely on the experience of the examiner. Several researchers had reported the poor inter-rater reliability of AS [8,9], MAS [10,11] and Penn spasm frequency scale [8].

Neurophysiologic analysis is the measurement of the electrical activity, such as EMG signals in order to evaluate the spasticity. Several studies have accordingly used EMG to measure the responses evoked by either the stretching of the muscle (M-reflex), tendon tap (T-reflex) or electrical stimulation of the peripheral nerve supplying the muscle (H-reflex), in order to evaluate whether these responses are exaggerated in spastic individuals and related to the degree of spasticity [12]. However, several researchers reported that the ratio of the H-reflex to SR of the muscle (H/M ratio) was not correlated with the MAS, although it was increased in patients with spinal cord injury (SCI) [13,14] or stroke [15,16] compared to that in healthy subjects. The pendulum test, introduced by Wartenberg in 1951 [17], is a biomechanical method of evaluating muscle tone by using gravity to provoke muscle SRs during the passive swinging of the lower limb. Some researchers have reported that the ratio of the amplitude of the first swing to that of the final position is significantly correlated with clinical scales such as the AS and MAS scores in spastic patients [18-20]. However, this correlation depends decisively on the sitting posture and the ability of the person to fully relax. Also, it could only be applied to evaluate the spasticity in the knee flexor and extensor muscles and is limited to separate the increased resistance of the spastic muscles, due to the changes of the viscoelastic resistance from the velocity-dependent resistance [12].

The isokinetic dynamometer has been widely used for the quantitative assessment and evaluation of the spasticity. It allows the velocity and amplitude applied to evoke muscle stretches to be standardized; consequently it is able to quantify the velocity-dependent resistance according to the passive movement of the muscle. Firozabakhsh et al. [21]

found a significantly greater sum torque and slope of the torque-velocity regression lines in the spastic group compared to the normal group. Pisano et al. [22] demonstrated that the total stiffness indices (TSI), stretch reflex threshold speed (SRTS) and SR area were highly correlated with the AS. Pandyan et al. [23,24] showed that there was a high correlation coefficient between the MAS and resistance to passive movement (RTPM) and that the RTPM in the impaired arm was relatively larger than that in the non-impaired arm. Chen et al. [25] and Lee et al. [26] showed that there was a decrease of the biomechanical viscosity and reflexive EMG threshold (RET) of the biceps brachii after the injection of Botulinum toxin type-A.

The theoretical concept for tonic stretch reflex threshold (TSRT) measurement, based on motor control theory, was first published by Levin and Feldman [27]. The TSRT is based on the evaluation of the excitability of the motor neurons caused by both descending and segmental effects, and the measurement of these effects is the SR threshold, the integral part of the λ model of motor control. The SR threshold depends on the stretch velocity. In the λ model, the dynamic stretch reflex threshold (DSRT) is expressed in velocity and angular coordinates, i.e. the velocity and joint angle at which the muscle activity first appears. When calculated in such coordinates, the DSRTs and TSRT are expressed in relation to the actual configuration of the joint within the body frame of reference. In particular, when the threshold lies within the biomechanical range of the joint and the patient has no ability to shift this threshold angle, it separates the joint configurations in which the muscles are spastic from those in which they are not, thus quantifying an important, spatial aspect of the motor control [28]. Some researchers have reported the validation of the TSRT. Levin et al. [27,29] showed a negative correlation between the CSI and TSRT and positive correlation with the Fugl-Meyer scale in elbow flexors and extensors of the spastic patients with stroke. Jobin et al. [30] showed the good test-retest reliability of TSRT measurement for the children with cerebral palsy. Recently, Calota et al. [31,32] described a portable device for TSRT measurement and demonstrated the moderately high reliability of TSRT measurement for patients with moderate to high spasticity. These results indicate the TSRT could be a more representative measure since it satisfies Lance's criteria for the velocity-dependent increase of the spasticity.

The objectives of this study are to develop a hand-driven portable system for quantifying the grade of spasticity, which can calculate the bio-mechanical as well as neurophysiologic parameters, and to determine the relationship between the TSRT measured by the developed device and the level of the spasticity. The TSRT of each spastic patient was measured during both the extension and flexion of the forearm in order to take into account threshold of both agonist and antagonist muscles, and TSRTs obtained from all of the patients were grouped by means of K-means clustering method for the objective discrimination of the severity of the spasticity. We hypothesized that there would be a negative correlation

between them (i.e., the larger the severity of the spasticity, the smaller the TSRT) through the literature reviews of the previous papers [29-32]. We implemented this approach in a portable device and applied it to the evaluation of the spasticity

2 Portable spasticity measurement system

The developed spasticity measurement system is designed to measure the angle by means of a twin-axis flexible electro-goniometer (SG150, Biometrics Ltd., U.K.) and EMG signals by means of surface electrodes (Meditrace 200, Kendall, U.S.). Also, the angular velocity is calculated by the differentiation of the angle signals. This device is composed of a sensor module for signal conditioning and control module to monitor the measured data and the physiological parameters. All signals are pre-processed by the signal conditioning circuit in the sensor module. Fig. 1 shows a block diagram of the developed system.

In order to store and analyze the data obtained during the

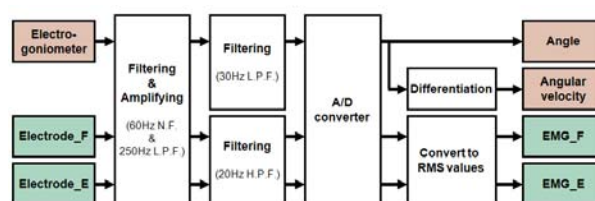


Fig. 1. Block diagram of the sensor module

study, data acquisition and analysis software were developed using the LabVIEW language (ver. 8.6, National Instruments™, U.S.). It can show the trace of the angle, angular velocity and two EMG signals continuously on a monitor using figures and numbers and simultaneously store the data on a hard-disk drive. Also, the system is equipped with a beep sound generator like a metronome, in order to announce the velocity of the flexion and extension of the upper limbs (stretch velocity) in a simple manner to both the subject and rater during the movements. The period between beep sounds can be selected manually in the range of 30 and

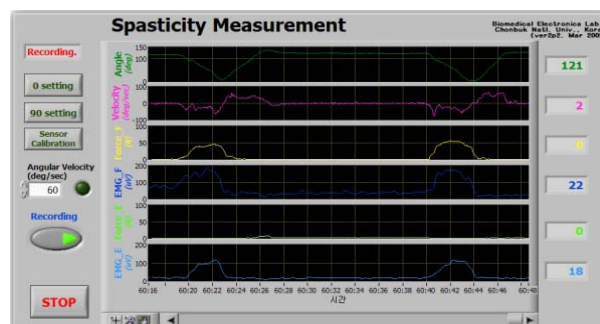


Fig. 2. Graphic user interface for data acquisition and parameter analysis

360 °/sec by the rater. Fig. 2 shows the graphic user interface for data acquisition and parameter analysis.

3 Materials and methods

3.1 Participants

Fifteen patients (7 males and 8 females) with stroke participated in the study (mean age 63.5 ± 15.6 , range 38-84 years) after giving their informed consent. Eleven patients had a cerebral infarction and four had a cerebral hemorrhage. The affected sides were 9 in the right arm and 6 in the left. The mean period since stroke was 6.7 months (range 1-36 months). The clinical grading by MAS was '1' in 6 patients, '1+' in 5 and '2' in 4. The MAS scores were considered as an auxiliary tool for the assessment of the trend of the TSRTs according to the level of the spasticity. All tests were performed on the more affected upper limb. Subjects were included if they had (a) sustained a stroke; (b) spasticity in the elbow flexors or extensors and (c) at least a 120° passive range of motion in the elbow joint. To calculate the clinical parameters, the presence or absence of spasticity in the elbow flexors and extensors was confirmed by manually stretching the elbow from full flexion to full extension at an arbitrary stretch velocity. Subjects were excluded if they (a) could not understand simple commands due to the decrease of their cognitive functions; (b) had subluxation or sprain of the shoulder or (c) elbow contracture.

3.2 Experimental protocol

Number Two clinicians (two males) evaluated each subject. The raters had different amounts of clinical experience (3.5 and 5.5 years). To ensure a standardized level of training, both evaluators received written documentation and participated in two one-hour training sessions with the developed device.

For the measurement of the angular displacement of the upper limb, a flexible electro-goniometer was placed on the lateral aspect of the elbow with the axis of rotation at the joint line and its two wings were fixed on the forearm and upper arm, respectively, by an elastic band. To monitor the activity of the elbow flexors and extensors, five surface electrodes (Meditrace 200, Kendall, U.S.) with a diameter of 10 mm were attached to the upper arm. The electrode sites were lightly shaved and cleaned with a 95 % ethanol mixture to reduce the skin impedance. For the elbow flexors, the active and reference electrodes were located on the biceps, while those used for the extensors were located on the triceps. A ground electrode was placed on the medial side of the elbow.

A motion from full flexion to full extension is defined as 'elbow extension', whereas 'elbow flexion' is defined as a motion from full extension to full flexion. 'Full extension' means an angle of 0° between the forearm and upper arm, while 'full flexion' means an angle of about 120°. One cycle

consisted of one elbow extension and one flexion over an approximate angle range of 120°→0°→120° in an approximate period of one cycle. The velocity of the extensors was determined as its total angular displacement per a period of one extension, while the velocity of the flexors was as the total displacement per a period of one flexion. If the periods of one extension and one flexion were 1 and 2 seconds, the stretch velocity were 120 °/sec in the extensors and 60 °/sec in the flexors, respectively.

The subjects lay on a bed in a relaxed position. The starting position involved the slight abduction of the shoulder, neutral position of the wrist and full extension of the elbow. In order to reduce the muscle tension, the subjects maintained this position for at least 2 minutes. The tests were performed after the rater checked whether all flexors and extensors were stabilized by monitoring their EMG signals. The subject's forearm was passively flexed and extended by the rater at a randomly selected stretch velocity among 60, 90, 120, 150 and 180°/sec, in order to avoid adaptation of the stretch response [33]. Measurements were performed repeatedly ten times at the selected stretch velocity with at least 10 seconds rest between sessions, because the motor unit recruitment threshold was 6 seconds during repeated contractions [34]. The total number of measurement sessions per subject was about 50. None of the sensors, including the electrogoniometer and surface electrodes, were displaced or

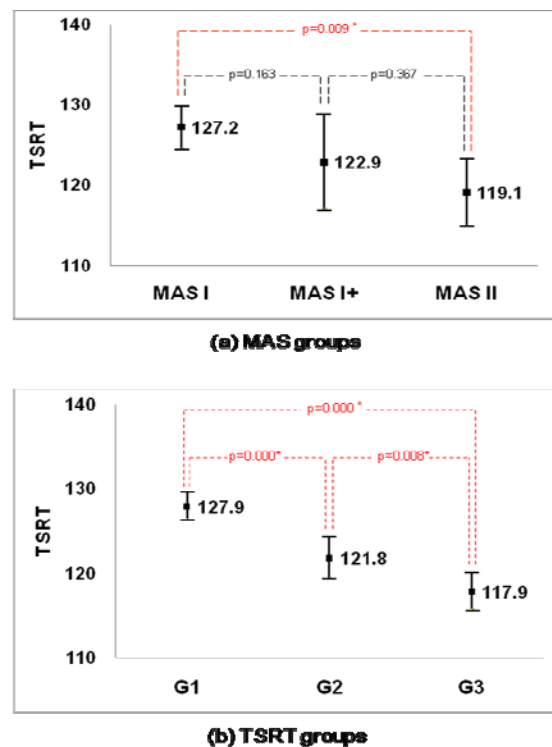


Fig. 3. Comparison of the TSRTs among the groups (a) classified by the MAS and (b) by the K-mean clustering of the TSRT (*: $p < 0.05$)

reattached during the test.

The DSRTs is defined as the joint angle and angular velocity value corresponding to the point at which the EMG signal increased by 2 times of standard deviations above the mean baseline EMG [31]. The baseline EMG was the EMG activity while the subject was at rest before beginning the evaluation session. At the end of each evaluation, the joint angle and angular velocity values at the time, when the first incident of the EMG activity of either the extensor or flexor was detected, were automatically obtained in the analysis software. Next, using these values the DSRT was calculated. Finally, the TSRT was computed by drawing the linear regression line through the DSRTs to zero velocity. The TSRT value was taken as the intercept of the regression line with the angle axis (dependent variable: angle, independent variable: angular velocity).

4 Results

Fig. 3 shows the comparison of the TSRTs among the groups classified by the MAS and K-mean clustering of the TSRTs. through the K-means clustering algorithm, the patients were classified into three groups (G1, G2 and G3) according to the criteria of the TSRTs. The centroids of each group were 127.9 in group G1, 121.8 in group G2 and 117.9 in group G3 and the Euclidean distances were 6.099 between groups G1 and G2, 10.052 between groups G1 and G3 and 3.952 between groups G2 and G3.

When grouping the patients according to the level of the MAS, the mean and standard deviation (S.D.) values of the TSRTs were 127.2 ± 2.5 in the MAS1, 122.9 ± 4.7 in the MAS1+ and 119.1 ± 2.6 in the MAS2 groups, respectively. In order to compare the differences of the TSRTs between the MAS groups, the one-way ANOVA test was performed. Consequently, the average TSRT in the MAS1 group were the largest, while that in the MAS2 group were the smallest ($p < 0.05$). Also, there was a negative correlation ($r = -0.74$, $r^2 = 0.54$, $p < 0.05$) between the TSRT and MAS. However, through the post hoc analysis, the differences between the average TSRT of the MAS1 group and that of the MAS1+ group ($p = 0.16$) and between the average TSRT of the MAS1+ group and that of the MAS2 group ($p = 0.37$) were not significant, as shown in Fig. 3(a).

On the other hand, when grouping by means of the K-means clustering of the TSRTs, the mean and S.D. values of the TSRTs of groups G1, G2 and G3 were 127.9 ± 1.6 , 121.8 ± 1.5 and 117.9 ± 1.3 , respectively. There was a strong negative correlation between the TSRTs and groups ($r = -0.95$, $r^2 = 0.90$, $p < 0.05$). Also, there were significant differences between the TSRTs of each group ($p < 0.05$), as shown in Fig. 3(b).

5 Conclusions

We developed a portable spasticity measurement system and classification algorithm for the objective and reliable discrimination of the level of spasticity based on the K-means clustering of the TSRTs. Our results showed the existence of a strong negative relationship between the TSRTs and classified groups ($r = -0.95$, $r^2 = 0.90$, $p < 0.05$). This demonstrates that our method could be made clinically available for the more objective and reliable discrimination of the spasticity, instead of the conventional MAS grade. In a future work, we will apply our system to a larger number of spastic patients with various upper motor-neuron disorders and verify its feasibility.

6 Acknowledgment

This work was supported by the second stage of Brain Korea 21 Project in 2011 and by the Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea Government Ministry of Knowledge Economy (No.20104010100660).

7 References

- [1] Lance JW, "Spasticity: Control of muscle tone, reflexes and movement: Robert Wartenberg lecture," *Neurology* 1980, 30(12):1303-1313.
- [2] Bell KR, Vandenborne K. *Contracture and limb deformities*, McGraw-Hill, New York, 1998.
- [3] Ashworth B. Preliminary trial of carisoprodol in multiple sclerosis. *The Practitioner* 1964;192(1):540-542.
- [4] Bohannon RW, Smith MB. Interrater reliability of a modified Ashworth Scale of muscle spasticity. *Physical Therapy* 1987;67(2):206-207.
- [5] Boyd R, Graham HK. Objective measurement of clinical findings in the use of botulinum toxin type A for the management of children with CP. *European Journal of Neurology* 1999;6(S4):S23-S35.
- [6] Levin MF, Hui-Chan CW. Relief of hemiparetic spasticity by TENS is associated with improvement in reflex and voluntary motor functions. *Electroencephalography and Clinical Neurophysiology* 1992;85(2):131-142.
- [7] Burridge JH, Wood DE, Hermens HJ, Voerman GE, Johnson GR, van Wijck F, Platz T, Gregoric M, Hitchcock R, Pandyan AD. Theoretical and methodological considerations in the measurement of spasticity. *Disability and Rehabilitation* 2005;27(1):69-80.

- [8] Priebe MM, Sherwood AM, Thornby JI, Kharas NF, Markowski J. Clinical assessment of spasticity in spinal cord injury: a multidimensional problem. *Archives of Physical Medicine and Rehabilitation* 1996;77(7):713-716.
- [9] Pandyan AD, Johnson GR, Price CIM, Curless RH, Barnes MP, Rodgers H. A review of the properties and limitations of the Ashworth and modified Ashworth scales as measures of spasticity. *Clinical Rehabilitation* 1999;13(5):373-383.
- [10] Allison SC, Abraham LD, Petersen CL. Reliability of the modified Ashworth scale in the assessment of plantarflexor muscle spasticity in patients with traumatic brain injury. *International Journal of Rehabilitation Research* 1996;19(1):67-78.
- [11] Blackburn M, van Vliet P, Mockett SP. Reliability of measurements obtained with the modified Ashworth scale in the lower extremities of people with stroke. *Physical Therapy* 2002;82(1):25-34.
- [12] Biering-Sorensen F, Nielsen JB, Klinge K. Spasticity assessment: a review. *Spinal Cord* 2006;44(12):708-722.
- [13] Little JW, Halar EM. H-reflex changes following spinal cord injury. *Archives of Physical Medicine and Rehabilitation* 1985;66(1):19-22
- [14] Crone C, Johnsen LL, Biering-Sorensen F, Nielsen JB. Appearance of reciprocal facilitation of ankle extensors from ankle flexors in patients with stroke or spinal cord injury. *Brain* 2003;126(2):495-507.
- [15] Levin MF, Hui-Chan C. Are H and stretch reflexes in hemiparesis reproducible and correlated with spasticity?. *Journal of Neurology* 1993;240(2):63-71.
- [16] Bakheit AM, Maynard VA, Curnow J, Hudson N, Kodapala S. The relation between Ashworth scale scores and the excitability of the alpha motor neurones in patients with post-stroke muscle spasticity. *Journal of Neurology, Neurosurgery and Psychiatry* 2003;74(5):646-648.
- [17] Wartenberg R. Pendulousness of the legs as a diagnostic test. *Neurology* 1951;1(1):18-24.
- [18] Leslie GC, Muir C, Part NJ, Roberts RC. A comparison of the assessment of spasticity by the Wartenberg pendulum test and the Ashworth grading scale in patients with multiple sclerosis. *Clinical Rehabilitation* 1992;6(1):41-48.
- [19] Fowler EG, Nwigwe AI, Ho TW. Sensitivity of the pendulum test for assessing spasticity in persons with cerebral palsy. *Developmental Medicine and Child Neurology* 2000;42(3):182-189.
- [20] Nordmark E, Andersson G. Wartenberg pendulum test: objective quantification of muscle tone in children with spastic diplegia undergoing selective dorsal rhizotomy. *Developmental Medicine and Child Neurology* 2002;44(1):26-33.
- [21] Firoozbakhsh KK, Kunkel CF, Scremin AME, Moneim MS. Isokinetic dynamometric technique for spasticity assessment. *American Journal of Physical Medicine and Rehabilitation* 1993;72(6):379-385.
- [22] Pisano F, Miscio G, Conte CD, Pianca D, Candeloro E, Colombo R. Quantitative measures of spasticity in post-stroke patients. *Clinical Neurophysiology* 2000;111(6):1015-1022.
- [23] Pandyan AD, Price CIM, Rodgers H, Barnes MP, Johnson GR. Biomechanical examination of a commonly used measure of spasticity. *Clinical Biomechanics* 2001;16(10):859-965.
- [24] Pandyan AD, Price CIM, Barnes MP, Johnson GR. A biomechanical investigation into the validity of the modified Ashworth scale as a measure of elbow spasticity. *Clinical Rehabilitation* 2003;17(3):290-294.
- [25] Chen JJ, Wu YN, Huang SC, Lee HM, Wang YL. The use of a portable muscle tone measurement device to measure the effects of Botulinum toxin type A on elbow flexor spasticity. *Archives of Physical Medicine and Rehabilitation* 2005;86(8):1655-1660.
- [26] Lee HM, Chen JJJ, Wu YN, Wang TL, Huang SC, Piotrkiewicz M. Time course analysis of the effects of Botulinum toxin type A on elbow spasticity based on biomechanic and electromyographic parameters. *Archives of Physical Medicine and Rehabilitation* 2003;89(4):692-699.
- [27] Levin MF, Feldman AG. The role of stretch reflex threshold regulation in normal and impaired motor control. *Brain Research* 1994;657(1):23-30.
- [28] Musampa NK, Mathieu PA, Levin MF. Relationship between stretch reflex thresholds and voluntary arm muscle activation in patients with spasticity. *Experimental Brain Research* 2007;181(4):579-593.
- [29] Levin MF, Selles RW, Verheul MHG, Meijer OG. Deficits in the coordination of agonist and antagonist muscles in stroke patients: implications for normal motor control. *Brain Research* 2000;853(2):352-369.
- [30] Jobin A, Levin MF. Regulation of stretch reflex threshold in elbow flexors in children with cerebral palsy: a new measure of spasticity. *Developmental Medicine and Child Neurology* 2000;42(8):531-540.

[31] Calota A, Feldman AG, Levin MF. Spasticity measurement based on tonic stretch reflex threshold in stroke using a portable device. *Clinical Neurophysiology* 2008;119(10):2329-2337.

[32] [32] Calota A, Levin MF. Tonic stretch reflex threshold as a measure of spasticity: implication for clinical practice. *Topics in Stroke Rehabilitation* 2009;16(3):177-188.

[33] Schmit BD, Dewald JPA, Rymer WZ. Stretch reflex adaptation in elbow flexors during repeated passive movements in unilateral brain-injured patients. *Archives of Physical Medicine and Rehabilitation* 2000;18(3):269-278.

[34] [Gorassini M, Yang JF, Siu M, Bennett DJ. Intrinsic activation of human motoneurons: Reduction of motor unit recruitment thresholds by repeated contractions. *Journal of Neurophysiology* 2002;87(4):1859-1866.

[35] Nuyens GE, Weerdt WJ, Spaepen AJ, Kiekens C, Feys HM. Reduction of spastic hypertonia during repeated passive movements in stroke patients. *Archives of Physical Medicine and Rehabilitation* 2002;83(7):930-935.

[36] Chung SG, Rey E, Bai Z, Rymer WZ, Roth EJ, Zhang LQ. Separate quantification of reflex and nonreflex components of spastic hypertonia in chronic hemiparesis. *Archives of Physical Medicine and Rehabilitation* 2008;89(4):700-710.

[37] Hornby TG, Kahn JH, Wu M, Schmit BD. Temporal facilitation of spastic stretch reflexes following human spinal cord injury. *The Journal of Physiology* 2006;571(3):593-604.

[38] Hagbarth KE, Hagglund JV, Nordin M, Wallin EU. Thixotropic behavior of human finger flexor muscles with accompanying changes in spindle and reflex responses to stretch. *The Journal of Physiology* 1985;368:323-342.

GPU Accelerated PK-means Algorithm for Gene Clustering

Wuchao Situ, Yau-King Lam, Yi Xiao, P.W.M. Tsang, and Chi-Sing Leung

Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China

Abstract - In this paper, a novel GPU accelerated scheme for the PK-means gene clustering algorithm is proposed. According to the native particle-pair structure of the PK-means algorithm, a fragment shader program is tailor-made to process a pair of particles in one pass for the computation-intensive portion. As the output channel of a fragment consisting of 4 floating-point values is fully utilized, overhead for each data points in searching for its nearest centroid throughout the particle-pair is reduced. Experimental evaluations on three popular gene expression datasets show that the proposed GPU accelerated scheme can attain an order of magnitude speedup as compared with the original PK-means algorithm.

Keywords: Gene clustering, K-Means, PK-means, GPU

1 Introduction

Nowadays, gene clustering has attracted more and more attentions as the advancement of the technologies both in microarray [1] and computing. Microarray technology allows producing gene expression data in lower cost and monitoring large amount of data simultaneously for whole genome though a single chip only [2]. There is a huge of gene expression data produced in laboratory. To study the interactions among thousands of genes in the massive data sets, cluster analysis is the first and important step. An efficient cluster analysis is required to rapidly extract useful information from the raw data. Later advanced processing can be done further, such as protein structure prediction, biological network modeling, by using some natural computing [3].

There are numerous methods developed for clustering in the past [4-7], and among them K-means, with its simplicity and effectiveness, is perhaps the most popular one. However, due to its sensitivity to the initial condition, it is easy to get trapped in local optimal. To overcome this problem, recently a new clustering algorithm, known as PK-means, is proposed [8], which merges K-Means with the particle swarm optimization (PSO) [9, 10] algorithm. Experimental evaluation shows that it can reach better clustering results. However, its computation time is rather long, especially for large dimensional dataset. The bottleneck lies in the operation of K-means, which is a basic part of PK-means.

To overcome this problem, we propose to introduce the graphics process units (GPU), with its powerful stream processing units, to perform the tedious K-means operation. As the PK-means is working on particle-pairs, each particle-pair is packed together and fit to the programmable graphics pipeline [11] in GPU, where the two particles are together evaluated within a single fragment program. With the output channel of each fragment fully utilized, overhead for each data point in searching for its nearest centroid throughout the particle-pair is reduced. Organization of this paper is listed as follows. Section 2 gives a brief review of the PK-means algorithm. In section 3, we describe GPU accelerated scheme for PK-means. This is followed by the experimental evaluation on the proposed method in Section 4. Finally, a conclusion summarizing the essential findings is drawn in Section 5.

2 The PK-means clustering algorithm

The PK-means clustering method is the integration of the particle-pair optimizer (PPO) [12] and the well-known K-means. The former is a variation of the traditional particle swarm optimization (PSO) algorithm, while introducing a smaller swarm size based on particle pairs. For the clustering problem, a particle's position is a set of K cluster centroids, each of which is a D -dimensional vector, i.e.,

$$X_{i,n} = (x_{i,n,1}, x_{i,n,2}, \dots, x_{i,n,K}), \quad (1)$$

where n is the iteration number. The velocity vector of this particle towards its next position is denoted by

$$V_{i,n+1} = wV_n + C_1r_1(p_{i,n} - X_{i,n}) + C_2r_2(Gbest - X_{i,n}), \quad (2)$$

where w is the inertia weight; $p_{i,n}$ is the best position for the particle ' i ' recorded so far; $Gbest$ represents the globally best position for the whole swarm throughout history; C_1 and C_2 are called acceleration factors; r_1 and r_2 are two random numbers within [0,1]. With the velocity vector available, the particle updates its position by

$$X_{i,n+1} = X_{i,n} + V_{i,n+1}. \quad (3)$$

To begin with, an initial swarm of four randomly generated particles is created and partitioned into two particle-pairs: $\{P_1, P_2\}$ and $\{P_3, P_4\}$, as shown in Fig 1. Each particle pair evolves independently. Particles in each pair update their positions and velocity according to Eqs. (2) and (3), and perform K-means to update and evaluate its fitness. After a certain number of iterations, two particles (denoted as EP1 and EP2) with the better fitness values in their respective particle-pair are selected and combined together to form an elitist particle-pair $\{EP_1, EP_2\}$. The latter will continue to evolve and finally the particle EP3 with a better fitness value as the winner of $\{EP_1, EP_2\}$ will represent the final solution.

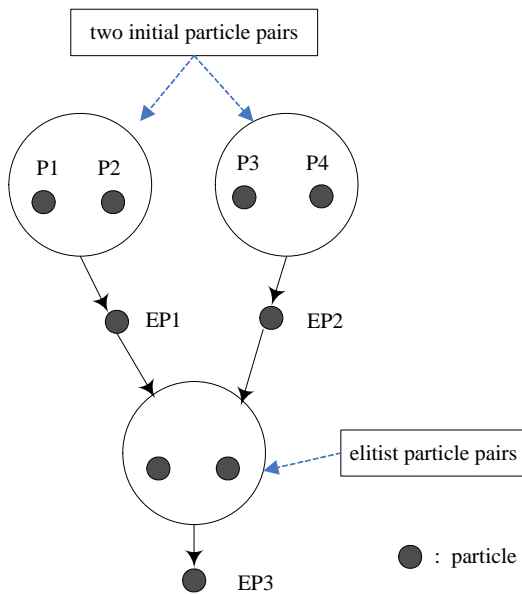


Fig 1. The evolution model of the Particle-Pair Optimizer.

3 The proposed GPU scheme for PK-means clustering

3.1 Overview of the GPU accelerated PK-means

The majority of computational cost in the PK-means algorithm lies in the operations of K-means for each particle-pair, and hence becomes the bottleneck of the algorithm. In view of this, we propose to convert this tedious step into a programmable graphics pipeline, where the Cg fragment shader program is tailor-made to perform the K-means operations for each particle-pair. An overview of the integration of GPU and PK-means algorithm is depicted in Fig. 2, where the building block “GPU accelerated K-means for particle-pair”, performing K-means for two particles, will be explained in detail in next subsection.

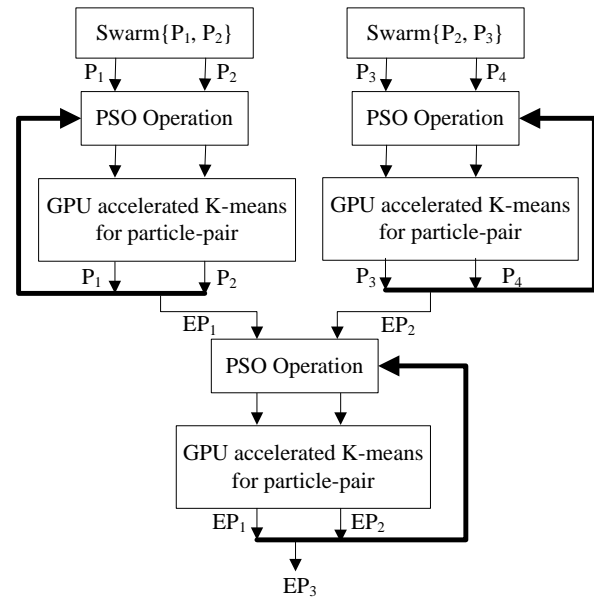


Fig 2. Overview of GPU accelerated PK-means

3.2 Fragment shader program based GPU Implementation of K-means for particle-pair

In the proposed GPU accelerated scheme, membership assignment for the particle-pair is conducted in the GPU while particle-pair updates (centroids updates) and the Mean Squared Error (MSE) calculation are carried out in the CPU. Fig. 3 gives the architecture of the GPU accelerated scheme.

Due to the large dimensionality of datasets we are dealing with (e.g. 77 for the Yeast cell-cycle data), both the data vectors and the particle-pair (each particle containing a set of K centroids) are stored in GPU textures, serving as look-up tables in the fragment shader program. Since a texture can store four single precision floating-point values (RGBA) per texel, a D -dimensional vector occupies $\lceil D/4 \rceil$ texels.

In the fragment shader program, each data vector is addressed by each fragment, and a pair of its nearest centroids is found in the two particles, respectively. Consequently, a pair of clustering memberships (each consisting of the ID of the nearest centroid and the nearest distance) is formed and rendered to the render texture. The latter is then downloaded to the CPU side where the particle-pair is updated and the MSEs are computed. Next round of iteration will be triggered from the CPU and the shader program repeats until a certain number of iterations has elapsed. Table 1 gives the Cg codes and Table 2 lists the notations for the fragment shader program.

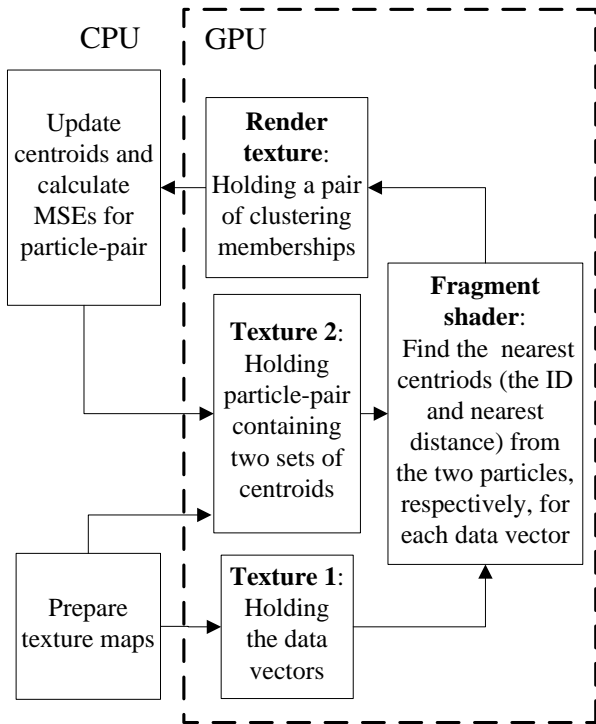


Fig 3. Overview of GPU accelerated K-means for particle-pair

Table 2 Notations for the fragment shader program in Table 1

Variable	Description
w	Equal to $\lceil D/4 \rceil$.
srctex	Texture holding the data vectors.
booktex	Texture holding the centroids.
index.y	The index of a vector in the data set.
codebook_size	Number of clusters/centroids.
minIndex1, minIndex2	The two IDs of the nearest centroid in the particle-pair, respectively.
mindist1, mindist2	Distances of a vector to its nearest centroids in the particle-pair, respectively.
memberships	A pair of clustering memberships, each consisting of the ID of the nearest centroid and the nearest distance.

Table 1 Fragment shader program - forming a pair of clustering memberships of the particle-pair for each data vector

```

#define num_of_centroids 256
#define FLT_MAX 3.402823466e+38F

void cgKMeans(
    float2 index: TEXCOORD0,
    uniform samplerRECT srctex,
    uniform samplerRECT booktex,
    out float4 memberships: COLOR0
){
    float mindist1, distance1, mindist2, distance2;
    float4 dn, dt;
    int i, k, minIndex1, minIndex2;
    minIndex1=minIndex2=-1;
    mindist2=mindist1=FLT_MAX;

    for ( k=0; k<num_of_centroids; k++)
    {
        distance2=distance1=0;
        for ( i=0; i<w; i++) // for all dimensions in a vector
        {
            // position of the data vector
            dn=texRECT(srctex,
                float2(i+(index.x-0.5)*w+0.5, index.y ) );

            // distance to the k-th centroid in particle 1
            dt= dn - texRECT( booktex, float2(i+0.5, k+0.5) );
            distance1 += dot(dt, dt);

            // distance to the k-th centroid in particle 2
            dt= dn - texRECT( booktex,
                float2(i+0.5, k+num_of_centroids+0.5) );
            distance2 +=dot(dt, dt);
        }

        if ( distance1<mindist1 ) // for particle 1
        {
            mindist1=distance1; //minimum distance
            minIndex1=k; // ID of the nearest centroid
        }
        if ( distance2<mindist2 ) // for particle 2
        {
            mindist2=distance2;
            minIndex2=k;
        }
    }

    // output the pair of clustering memberships
    memberships=float4 (minIndex1, mindist1,
        minIndex2, mindist2);
}
    
```

4 Experimental evaluation

The proposed scheme is evaluated with three popular gene expression datasets: Yeast cell-cycle [13] with 77 dimensions, Lymphoma [14] with 96 dimensions and Sporulation [15] with 7 dimensions. They have over 5 thousands, 4 thousands and 6 thousands of genes, respectively.

Performance of the GPU accelerated PK-means method is compared against the same PK-means algorithm implemented without GPU (referred to as the parent scheme). Both methods are applied to cluster each of the dataset into 256 clusters. To obtain reliable statistics, a total of 10 repeated trials for the three datasets are conducted. All the evaluations are based on the CPU (Intel Core2 Duo E6550 2.33GHz) and GPU (NVIDIA GTX260). The results of average computation time (in second) taken to reach convergence for both methods, are listed in Table 3. It can be seen that the GPU accelerated PK-means is at least 11 times faster than the parent scheme, and for the lower-dimensional dataset (i.e. Sporulation), over 20 times' speedup can be noted.

Table 3. Average computation time for the parent scheme and proposed scheme (Speed-up ratio: time of parent scheme / time of proposed GPU scheme)

Gene dataset	Scheme	Time (Sec)	Speed-up ratio
Yeast cell-cycle	Parent scheme	78.6	11.2
	GPU scheme	7.0	
Lymphoma	Parent scheme	68.4	11.4
	GPU scheme	6.0	
Sporulation	Parent scheme	16.7	20.5
	GPU scheme	0.8	

5 Conclusions

Gene cluster analysis plays an important role in discovering the function of gene. K-means is one of the well-known clustering methods for its simplicity and effectiveness. However, due to its sensitivity to the initial clustering, it is prone to be trapped in a local minimum. Recently, an enhanced clustering method, known as PK-means, which incorporates K-means with the particle swarm optimization is developed. Despite its success in finding better clustering results, the process is usually too time-consuming. The bottleneck lies in the K-means operation which is a basic portion of PK-means. To address the shortcoming, this paper proposes a novel GPU accelerated scheme for the PK-means algorithm. Based on the particle-pair structure of PK-means,

each particle-pair is packed together and fit to a tailor-made fragment shader program, where a pair of clustering membership is formed for the particle-pair and then sent to the entire output channel of each fragment. As the latter is fully utilized, overhead is reduced. Experimental evaluation on three gene expression datasets reveals that the proposed GPU accelerated scheme can attain an order of magnitude speedup as compared with the parent scheme.

6 References

- [1] P. Brown, D. Botstein, "Exploring the New World of the Genome with DNA Microarrays"; *Nature Genetics*, Vol. 21, 33-37, 1999
- [2] A. Brazma, A. Robinson, G. Cameron, M. Ashburner. "One-stop Shop for Microarray Data"; *Nature*, Vol. 403, 699-700, 2000
- [3] F. Masulli, S. Mitra. "Natural computing methods in bioinformatics: A survey"; *Information Fusion*, Vol. 10, issue 3, 211-216, 2009.
- [4] R. Shamir, R. Sharan. "Algorithmic approaches to clustering gene expression data". *Current Topics in Computational Biology*, MIT Press, Cambridge, MA, pp. 269-299, 2002
- [5] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein. "Cluster analysis and display of genome-wide expression patterns"; *PNAS*, Vol. 95, 14863-14868, 1998.
- [6] T. Kohonen; "The self-organizing map"; *Proc. IEEE* 78, 1464-1480, 1990.
- [7] J.C. Bezdek, R. Ehrlich, W. Full. "FCM: the Fuzzy c-means clustering algorithm"; *Comput. Geosci.* Vol. 10 issue 2-3, 191-203, 1984.
- [8] Z. Du, Y. Wang, Z. Ji. "PK-Means: A new algorithm for gene clustering"; *Comput. Biol. Chem.*, Vol. 32, issue 4, 243-247, 2008.
- [9] R. Eberhart, J. Kennedy. "A new optimizer using particle swarm theory"; In: *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*. IEEE Service Center, Nagoya, Japan, pp. 39-43. 1995.
- [10] J. Kennedy, R. Eberhart. "Particle swarm optimization"; In: *Proceedings of IEEE International Conference on Neural Networks*. IEEE Service Center, Piscataway, NJ, pp. 1942-1948, 1995.

- [11] R. Fernando, M. J. Kilgard. "The Cg tutorial: the definitive guide to programmable real-time graphics". Addison-Wesley, 2003.
- [12] Z. Ji, H., Liao, W. Xu, L. Jiang. "A strategy of particle-pair for vector quantization in image coding"; *Acta Electron. Sin.*, Vol. 35, issue 7, 86-89, 2007.
- [13] P.T. Spellman, et al. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization"; *Mol. Biol. Cell*, Vol. 9, 3273–3297, 1998.
- [14] A.A. Alizadeh, et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling"; *Nature*, Vol. 403, 503–511, 2000.
- [15] S. Chu, et al. "The transcriptional program of sporulation in budding yeast"; *Science*, Vol. 282, 699–705, 1998.

Computation of Coronary Sinus Pressure Using Pattern Recognition Techniques

Loay Alzubaidi¹, Ammar El Hassan², Jaafar Al Ghazo³

¹Department of Computer Science, Prince Muhammad bin Fahd University
AL-Khobar, Saudi Arabia
lalzubaidi@pmu.edu.sa

²Department of Information Technology, Prince Muhammad bin Fahd University
AL-Khobar, Saudi Arabia
aelhassan@pmu.edu.sa

³Department of Computer Engineering, Prince Muhammad bin Fahd University
AL-Khobar, Saudi Arabia
jghazo@pmu.edu.sa

Abstract - Pressure controlled intermittent coronary sinus occlusion (PICSO) has been found to substantially salvage ischemic myocardium. To indentify optimum occlusion and release points within PICSO cycles, two mechanisms are involved, however neither method is ideal. In this paper, a third method utilizing pattern recognition technology combined with ECG is introduced. This results in more efficient calculation of CSP parameters. Results of the new technique are shown from studying 3 groups of animals, namely sheep, pigs and dogs. The group size was 5, 5 and 3 respectively. All animals were drugged and anesthetized for the duration of the study.

Keywords: coronary sinus pressure (CSP), pressure controlled intermittent coronary sinus occlusion (PICSO).

1. Introduction

Pressure Controlled Intermittent Coronary Sinus Occlusion (PICSO) is implemented by means of a block which is applied via a catheter that intermittently obstructs the outflow from the cardiac veins in the right atrium, Figure 1 shows a single PICSO cycle of approx 16-seconds. The technique leads to an increase of Coronary Sinus Pressure CSP (systolic as well as diastolic) in the course of a few heart beats; controlled pressure increase can result in better distribution of the blood flow through the ischemic area. In order to maximize the effect of the PICSO procedure, it is

imperative that accurate Occlusion (Inflation) and Release (Deflation) points are identified within the PICSO cycle.

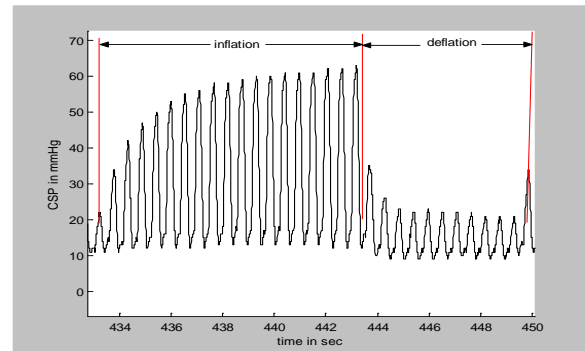


Figure 1 One PICSO with Inflation and Deflation time of 10/6 seconds respectively

Two techniques have been devised by the researchers to compute the Coronary Sinus Pressure (CSP) parameters (Systolic plateau, Diastolic plateau & the rise/release time) using a mathematical model which describes the increment and decrement of CSP in one PICSO cycle. The model consists of two parts that use 3-parameter double-exponential functions. This was fitted using the non-linear least square algorithms, as shown in Eq. 1 below:

$$P_{csp}(t) = \begin{cases} A * \exp(B * [1 - \exp(-C * t)]) - 1 & \text{when } 0 < t < T1 \\ D * \exp[E * [1 - \exp(-\frac{F}{t})]] - 1 & \text{when } T1 \leq t < T2 \end{cases} \quad (1)$$

Where $P_{csp}(t)$ = Coronary sinus pressure, and A, D, B, E, C, F are fitting parameters.

The first part of the equation (1a) describes the rise of the CSP during the inflation (occlusion) period.

$$P_{csp}(t) = A * \exp\{B * [1 - \exp(-C * t)] - 1\} \tag{1a}$$

The second part (1b) describes the release of the CSP during the deflation (release) period.

$$P_{csp}(t) = D * \exp\{E * [1 - \exp(-\frac{F}{t})] - 1\} \tag{1b}$$

The systolic peaks increase with the time during the inflation period. These peaks were fitted with the nonlinear least-square algorithms.

a) T90 Method

The *T90* method, developed by Schriener and Alzubaidi [2], is the first technique for calculating the CSP parameters during a PISCO cycle; this method yielded an approximate calculation by taking 90% of the predicted height of the systolic plateau. $P_{CSP}(t)$ reaches the maximum value when $t \rightarrow \infty$ in (Eq. 1a) as shown below

$$P_{csp}(t \rightarrow \infty) = A * \exp(B - 1) \tag{2}$$

Because, in mathematical terms, a plateau is never actually reached, it is meaningful to consider the time taken to reach 90% of the predicted height of the plateau. Figure 2 shows the systolic plateau and its rise time (RT).

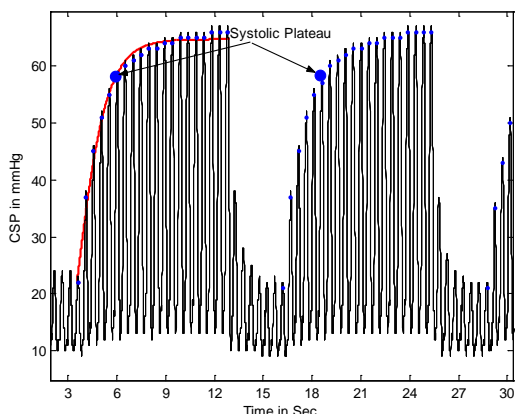


Figure 2 Systolic Plateau of CSP during the inflation period using T90 method

b) Time-Derivative Method

A new technique to describe the change of CSP parameters using the time-derivative method (dp/dt) was introduced by Alzubaidi L [1]. The new method is a more accurate means of calculating the systolic and diastolic plateau and the rise time of the PISCO cycle by determining the slope of CSP. The derived quantities serve as diagnostic parameters for a quantitative assessment of physiological condition and as predictors for an optimal adjustment of coronary sinus cycles.

The results of this technique were shown to bear a close resemblance to the clinical effect of coronary sinus occlusion. Fig. 3 illustrates comparison of the systolic plateaus of T90 and time-derivative methods

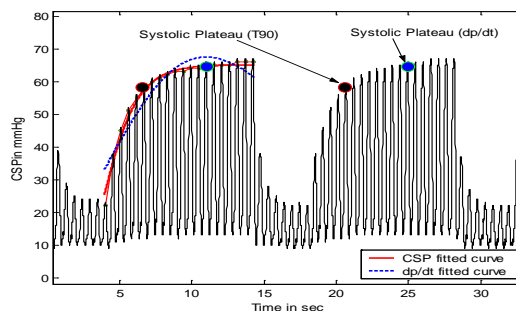


Figure 3 Systolic plateaus of CSP using the T90 and dp/dt methods

c) Weaknesses

There is room for improvement in both techniques above, as follows:

The T90 technique is fairly straight-forward, but the results are ultimately not as accurate as can be, this is due to the fact that the calculations are approximate.

Although the second (dp/dt) method is more accurate, it is also more complex and time-consuming. This is due to the extra overhead required for computing the slope of the CSP parameters using this approach.

What is needed, therefore, is an improvement in both accuracy and efficiency of the two algorithms above.

ECG Determines CSP Parameters

In this paper, we introduce a new technique to compute the rise and release of the CSP during the PISCO cycle. This technique can potentially yield more accurate figures using an ECG based calculation algorithm. The

new technique is a pattern recognition technique that recognizes the heart beat with lowest differential between the QRST interval and PQ interval during PICSO cycle.

Electrocardiogram (ECG)

The Electrocardiogram (ECG) is a biological signal. It is a quasi-periodical, rhythmically repeating signal, synchronized by the function of the heart, which acts as the generator of bioelectrical events. ECG is recorded by attaching a set of electrodes on the body surface such as chest, neck, arms, and legs.

ECG is an accurate, electrical manifestation of the contractile activity of the heart. By graphically tracing the direction and magnitude of the electrical activity that is generated by depolarization and repolarization of the atria and ventricles, the ECG chart provides information about the heart rate, rhythm, and morphology.

Each heartbeat can be observed as a series of deflections away from the baseline on the ECG. These deflections represent the time evolution of electrical activity in the heart which initiates muscle contraction. A single sinus (normal) cycle of the ECG, corresponding to one heart beat, is labelled with the letters P, Q, R, S and T on each of its switching/turning points as in Figure 4.

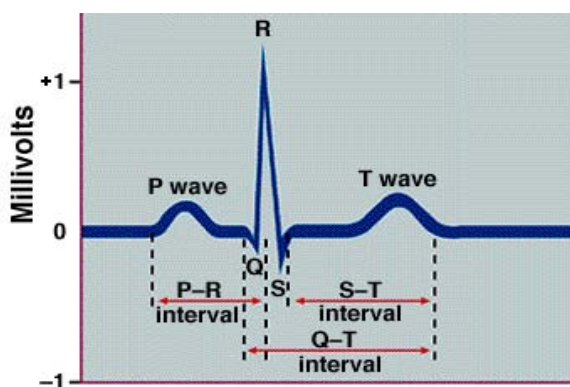


Figure 4 Morphology of PQRST for a single heartbeat

The ECG may be divided into the following sections:

- P-wave: a small low-voltage deflection away from the baseline caused by the depolarization of the atria prior to atrial contraction as the activation (depolarization) wave-front propagates through the atria.

- PQ-interval: the time between the start of atria depolarization and the start of ventricular depolarization.

- QRS-complex: the largest-amplitude portion of the ECG, caused by currents generated when the ventricles depolarize prior to their contraction. Although atria repolarization occurs before ventricular depolarization, the latter waveform (i.e. the QRS-complex) is of much greater amplitude and atria re-polarization is therefore not seen on the ECG.

- QRST-interval: the time between the onset of ventricular depolarization and the end of ventricular repolarization.

- ST-interval: the time between the end of S-wave and the beginning of T-wave. Significantly elevated or depressed amplitudes away from the baseline are often associated with cardiac illness.

- T-wave: ventricular repolarization, whereby the cardiac muscle is prepared for the next cycle of the ECG.

2. Methodology

The technique, which is based on pattern recognition concepts, was used to calculate the systolic plateau and the rise time (RT) of CSP by identifying a significant heartbeat (the heartbeat with lowest QRST & PQ interval variation). The rise time (RT) is the time between the start of the PICSO cycle and our significant heartbeat; now that RT is identified, it can be used in Eq1 (above) as a parameter to calculate the systolic plateau.

The physiological implication of this relationship is illustrated in Figure 5 and Table 1 below. All values, apart from the RT time, are in milliseconds.

$$PQ + QRST = \text{Heart Beat Interval}$$

The minimum difference between QRST & PQ is 76.66 ms

$$RT = 10.48 - 4.04 = 6.44 \text{ sec}$$

Hence, the systolic plateau can be calculated by substituting this RT value for the (t) parameter in Eq1 above.

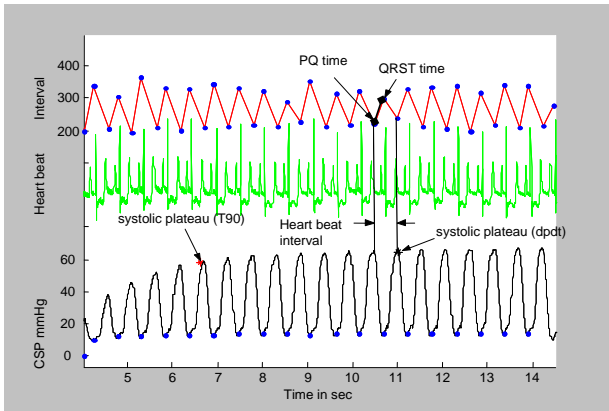


Figure 5 Relationship between CSP, ECG and the PQ & QRST intervals

Heartbeat Interval	PQ	QRST	QRST-PQ	RT (secs)
533.33	196.66	336.66	140	4.04
506.66	203.33	303.33	100	4.58
556.66	193.33	363.33	170	5.08
536.66	206.66	330	123.33	5.64
526.66	200	326.66	126.66	6.18
550	206.66	343.33	136.66	6.70
540	210	330	120	7.25
540	216.66	323.33	106.66	7.79
500	210	290	80	8.33
580	226.66	353.33	126.66	8.83
523.33	210	313.33	103.33	9.41
540	216.66	323.33	106.66	9.94
516.66	220	296.66	76.66	10.48
563.33	236.66	326.66	90	10.99
543.33	210	333.33	123.33	11.56
540	203.33	336.66	133.33	12.10
513.33	196.66	316.66	120	12.64
556.66	216.66	340	123.33	13.15

Table 1 Heartbeat interval, PQ time, QRST time and the CSP rise time RT for a 9-second inflation cycle

3. Results

All results were obtained by studying 3 groups of animals, namely sheep, pigs and dogs. The group size was 5, 5 and 3 respectively. All animals were pre-medicated with two ampoules atropine intramuscular and anesthetized before the catheters were placed into the

right ear artery for arterial pressure monitoring and/or the right ear vein for intravenous infusions.

At each step of the experiment the animals were monitored online to avoid any unnecessary suffering and to ensure anesthesia was still effective. The experiment was terminated during complete anesthesia with a high dosage of potassium chloride injection. All measurements were recorded on a computerised data acquisition system (monitoring and long time storage) for biological signals.

The results comprise a preliminary investigation of the spread of the derived quantities observed during PICSO cycles. The systolic plateau and its rise time were calculated for 10 PICSO cycles of approximately 14-seconds each (10 inflation + 4 deflation).

The systolic plateau and its rise time were used to compare the calculations from T90 method and pattern recognition method. Figure 6 shows the results of both calculations; it is clear that the systolic plateau of pattern recognition method is higher than its T90 counterpart.

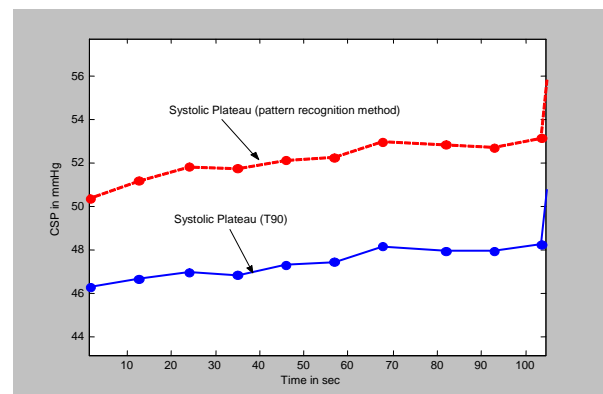


Figure 6 Comparison of Systolic Plateaus for 10 PICSO cycles using T90 and pattern recognition methods

The new technique can identify accurate occlusion and release points within PICSO cycles, thus helping to achieve an increase of Coronary Sinus Pressure CSP to yield higher blood pressure values resulting in better distribution of the blood flow through the ischemic area.

4. References

- [1] Alzubaidi L. Accurate methods of calculating the Coronary Sinus Pressure plateau. *J IJCSI* 2011; Vol. 8, Issue 1, pp.138-140;
- [2] Alzubaidi L, Mohl W, Rattay F. Automatic Computation for Pressure Controlled Intermittent Coronary Sinus. *J IJCSI* 2010; Vol. 7, Issue 6, pp.285-289;
- [3] Mohl W, Gueggi M, Haberzeth K, Losert U, Pachinger O, Schabart A. Effects of intermittent coronary sinus occlusion (ICSO) on tissue parameters after ligation of LAD. *Bibliotheca Anatomica* 1980; 20: 517-521.
- [4] Glogar D, Mohl W, Mayr H, Losert U, Sochor H, Wolner E. Pressure-controlled intermittent coronary sinus occlusion reduces myocardial necrosis (Abstract). *Am J Cardiol* 1982; 49: 1017.
- [5] Schreiner W, Neumann F, Schuster J, Froehlich KC, Mohl W. Computation of derived diagnostic quantities during intermittent coronary sinus occlusion in dogs. *Cardiovasc Res* 1988; 22(4): 265-276.
- [6] Schreiner W, Mohl W, Neumann F, Schuster J. Model of the haemodynamic reactions to intermittent coronary sinus occlusion. *J Biomed Eng* 1987; 9(2): 141-147.
- [7] Kenner T, Moser M, Mohl W, Tied N. Inflow, outflow and pressures in the coronary microcirculation. In: *CSI - A New Approach to Interventional Cardiology*. Mohl W, Faxon D, Wolner E (editors). Darmstadt: Steinkopff; 1986; 15.
- [8] Neumann F, Mohl W, Schreiner W. Coronary sinus pressure and arterial flow during intermittent coronary sinus occlusion. *Am J Physiol* 1989; 256(3 Pt 2): H906-915.
- [9] Moser M, Mohl W, Gallasch E, Kenner T. Optimization of pressure controlled intermittent coronary sinus occlusion intervals by density measurement. In: *The Coronary Sinus*, Vol. 1. Mohl W, Glogar D, Wolner E (editors). Darmstadt: Steinkopff; 1984; pp.529-536.
- [10] Mohl W, Glogar D, Kenner T, Klepetko W, Moritz A, Moser M. Enhancement of washout induced by pressure controlled intermittent coronary sinus occlusion (PICSO) in the canine and human heart. In: *The Coronary Sinus*, Vol. 1 Mohl W, Glogar D, Wolner E (editors). Darmstadt: Steinkopff; 1984; pp.537-548.

Linear Models and Biomarker Search with Microarray Data

E.M. Rivera¹, M.L. Sánchez-Peña¹, C.E. Isaza^{1,2}, J. Seguel³, M. Cabrera-Ríos¹

¹Bio IE Lab, Industrial Engineering Department, University of Puerto Rico - Mayagüez, Mayagüez, PR, USA

²Biology Department, Universidad Autónoma de Nuevo León, Sn. Nicolás de los Garza, NL, MEX

³Electrical & Computer Engineering Department, University of Puerto Rico - Mayagüez, Mayagüez, PR, USA

Abstract – *High throughput biological experiments such as DNA Microarrays are very powerful tools to understand and characterize multiple illnesses. These types of experiments, however, have also been described as large, complex, expensive and hard to analyze. For these reasons, analyses with linear assumptions are frequently bypassed for more sophisticated procedures with higher complexity. In this work, a search procedure for potential biomarkers using data from microarray experiments is proposed under purely linear assumptions. The method shows a high discrimination rate and does not require the adjustment of parameters by the user, thus preserving analysis objectivity and repeatability. A case study in the identification of potential biomarkers for cervix cancer is presented to illustrate the application of the proposed procedure.*

Keywords: Cancer biomarkers, Microarray Experiments,

1 Introduction

The search for genes whose measured change in expression behavior is an indication of a tissue being in a particular state (e.g. in a state of cancer vs. a state of health) is an important research objective in biology and the medical sciences. These genes are known as biomarkers. Microarray experiments play an important role in the identification of this type of genes. In the successful identification of potential biomarker genes, lies an important characterization of the cell in the presence of cancer. This can lead to enhance disease diagnosis and prognosis capabilities.

Based on our own experience with microarray data, the following challenges regarding microarray experiments can be identified: (1) the available data is highly dimensional in terms of the number of genes to be studied (~104) while showing a scarce number of replicates, (2) there is a rather large variation across replicates, (3) the data is not normally distributed and does not exhibit homogeneous variances, (4) there is a considerable number of missing observations in the majority of experiments, (5) the data is commonly found already being normalized or nonlinearly transformed. All of these complicate the detection of potential biomarkers.

Furthermore, when it comes to data analyses, the following are also important challenges: (i) there is no standard way to compare results for gene selection or identification between studies, (ii) even with the same data (and sometimes with the same technique) different researchers end up with different screening of genes [Ein-dor, et al. 2005] thereby leading to a large number of potential biomarkers to be investigated, the research of which could prove lengthy and very expensive.

Truly integrated work across disciplines is not frequent in most microarray analysis works. Biology and Medicine experts are usually left with the burden of using coded analysis tools with a series of parameters -of statistical, computational or mathematical nature/ that significantly affect the outcome of the software packages [Pan, 2002]. This leads to issues in results reproducibility and comparability between studies.

These challenges motivate the search for microarray analysis techniques from which consistent results can be achieved across several experiments and users, particularly for the identification of potential biomarkers.

The purpose of this work is to introduce an approach to identify potential biomarkers from the analysis of microarray experiments based solely on linear models and assumptions. Although an initial purpose on the design of the method was to establish a baseline of comparison for the many sophisticated methods with underlying nonlinear assumptions, it soon became apparent that a very effective strategy might be based on linearity.

2 The Analysis Strategy

Figure 1 schematically shows the strategy proposed in this work. Each step is explained below.

Step 1: Microarray Experiment. The process begins with a microarray experiment with m_1 tissues in state one (Healthy) and m_2 tissues in state two (Cancer) characterized in n genes. In the intersection of each of the n genes with each of the m_1+m_2 tissues, the relative expression of that particular gene in the selected tissue is quantified.

Step 2: Represent each gene with multiple performance measures. In this work, the use of a p _values is advocated to represent each gene. A p _value can be computed from the application of a statistical comparison test, like the Mann-Whitney nonparametric test for difference of medians. A different p _value for the same gene can be obtained by removing a couple of tissues from the microarray experiment under analysis. In a comparison of medians, a low p _value indicates a high probability for the medians to be significantly different.

Step 3: Apply Data Envelopment Analysis. Data Envelopment Analysis (DEA) finds the convex envelop of a particular data set consistently and without the need of varying parameters manually. If, for example, two p _values were used to represent each of the n genes in the experiment, then DEA can be used to find the envelope conformed by the dominating genes following the minimization direction of both p _values. Finding such envelope is done through the application of a linear programming formulation, which is the first instance where linearity becomes useful.

Step 4: Select genes in a series of efficient frontiers. The envelopes found through DEA are formally known as efficient frontiers. When an efficient frontier is found, then the solutions lying on it can be removed (as a layer of an onion), to then find the efficient frontier right underneath it. Following this scheme, several layers can be chosen containing different numbers of genes. These genes, having been found through the minimization of their p _values, are the most likely candidates to be biomarkers. These will be referred to as efficient genes.

Step 5: Create an experimental design to vary the efficient genes. An experimental design using as controllable variables the presence of the genes can be constructed. Each variable can take a value of 0 or 1 (0 for absence of the gene). This experimental design will prescribe a limited number of runs to measure a particular response of interest. In this case, one run corresponds to a combination of efficient genes.

Step 6: At each experimental design point, measure classification performance through linear discriminant analysis. Using the experimental design from the previous step, at each combination of efficient genes it is possible to obtain a measure of classification performance using a linear classifier through linear discriminant analysis. A linear classifier of this kind will always converge to the same position, thus preserving consistent results. At this point, then, a complete experimental design relating the classification rate with the absence or presence of the potential biomarkers is available.

Step 7: Fit a 1st order linear regression model. With the complete experimental design, it is possible to fit a 1st order linear regression model. This model will relate classification performance (response) to the absence or presence of the efficient genes (independent variables).

Step 8: Apply integer linear programming to choose the potential biomarkers that maximize classification performance. An optimization problem can be set up in this stage. This problem entails finding the combination of efficient genes –recall that each gene is represented by a variable that can take values of 0 or 1 to indicate absence or presence of that gene–, that maximizes the classification performance, i.e. choose the genes that maximize the regression model from the previous step.

This procedure, as it was explained, uses only linear models. Because of the techniques chosen in the strategy, the results are consistent. Furthermore, the selected genes do not depend upon the setting of any parameters by the user. This favors the repeatability and auditability of the analysis.

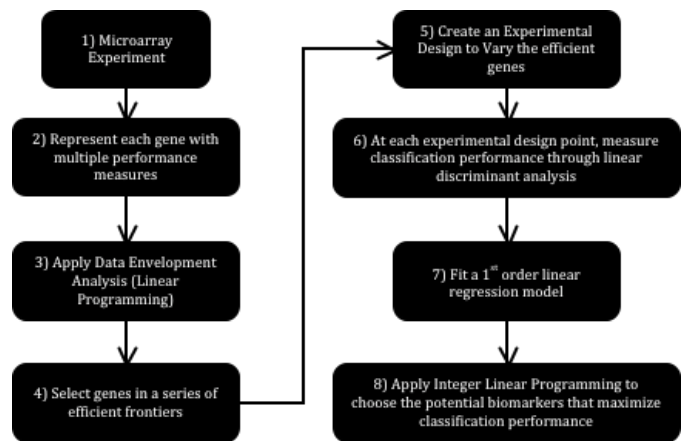


Figure 1. Analysis Strategy based on Linear Models

3 Case Study on Cervix Cancer

This case study helps to illustrate the application and the performance of the proposed procedure.

Step 1. The microarray database under analysis is related to cervix cancer and was compiled by Wong et al [3]. The database consists of 8 healthy tissues and 25 cervix cancer tissues, all of them with expression level readings for 10,690 genes.

Step 2. The Mann-Whitney nonparametric two-sided test for comparison of medians was used to generate two different p _values per gene, following a leave-one-tissue-out strategy, which focuses on extracting a particular tissue associated with one state. By removing a vector, a replicate is deleted from the set, thereby forcing a p _value that is different to the original one. Thus, two different p _values are effectively created. The selection of the tissue to be removed to create a distinct matrix is performed randomly as a first approach.

Step 3. The Data Envelopment Analysis model used for this case study was the Banks-Charnes-Cooper (BCC) model [4].

This is a linear programming model with the following associated formulations:

$$\begin{aligned}
 &\text{Find } \boldsymbol{\mu}, \mathbf{v}, \mu_0^+, \mu_0^- \quad \text{to} \\
 &\text{Maximize } \boldsymbol{\mu}^T \mathbf{Y}_0^{\max} + \mu_0^+ - \mu_0^- \\
 &\text{Subject to} \\
 &\quad \mathbf{v}^T \mathbf{Y}_0^{\min} = 1 \\
 &\quad \boldsymbol{\mu}^T \mathbf{Y}_j^{\max} - \mathbf{v}^T \mathbf{Y}_j^{\min} + \mu_0^+ - \mu_0^- \leq 0 \quad j = 1, \dots, n \\
 &\quad \boldsymbol{\mu}^T \geq \varepsilon \cdot \mathbf{1} \\
 &\quad \mathbf{v}^T \geq \varepsilon \cdot \mathbf{1} \\
 &\quad \mu_0^+, \mu_0^- \geq 0
 \end{aligned}$$

$$\begin{aligned}
 &\text{Find } \mathbf{v}, \boldsymbol{\mu}, v_0^+, v_0^- \quad \text{to} \\
 &\text{Minimize } \mathbf{v}^T \mathbf{Y}_0^{\min} + v_0^+ - v_0^- \\
 &\text{Subject to} \\
 &\quad \boldsymbol{\mu}^T \mathbf{Y}_0^{\max} = 1 \\
 &\quad \mathbf{v}^T \mathbf{Y}_j^{\min} - \boldsymbol{\mu}^T \mathbf{Y}_j^{\max} + v_0^+ - v_0^- \geq 0 \quad j = 1, \dots, n \\
 &\quad \mathbf{v}^T \geq \varepsilon \cdot \mathbf{1} \\
 &\quad \boldsymbol{\mu}^T \geq \varepsilon \cdot \mathbf{1} \\
 &\quad v_0^+, v_0^- \geq 0
 \end{aligned}$$

The optimal values of the decision variables correspond to the intercept and the partial first derivatives (with respect of each performance measure involved) of a supporting hyperplane lying on top of extreme points of the data set under analysis. At the end of the analysis, a piece-wise frontier is distinguishable as shown in Figure 3.

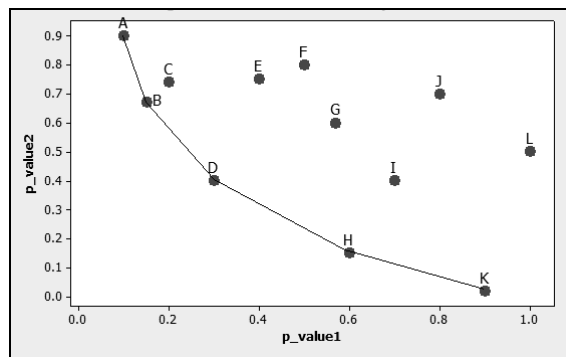


Figure 3. Representation of genes characterized through two different p_values. Only the case with 2 p_values has a convenient graphical representation, but the analysis can be extended to as many dimensions as performance measures selected.

Step 4. The first ten frontiers were kept for this analysis containing a total of 28 genes. It is important to note the discrimination rate shown by the method already at this point: a reduction of four orders of magnitude in the number of genes to analyze.

Step 5. A composite experimental design involving 28 binary variables (one per gene in the shortlist from the previous step), was used. Three different experimental designs form the composite with 123 runs. The first design is an orthogonal array consisting on 47 runs with between 10 to 18 genes each; the second design has 48 runs with between 1 to 26 genes generated randomly; and the third design consisted of 28 runs, each with only one gene. Figures 4, 5 and 6 show the resulting designs.

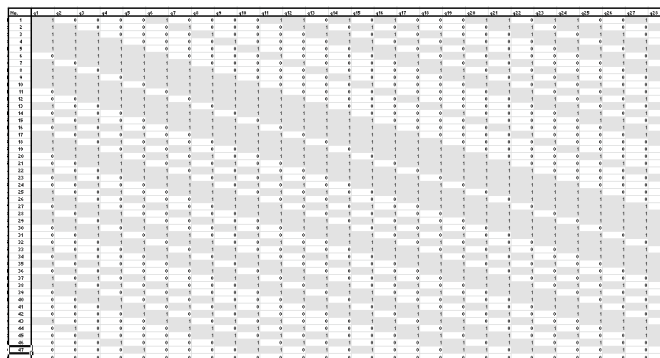


Figure 5. Design of Experiment 1. Shaded in gray are the values of 1.

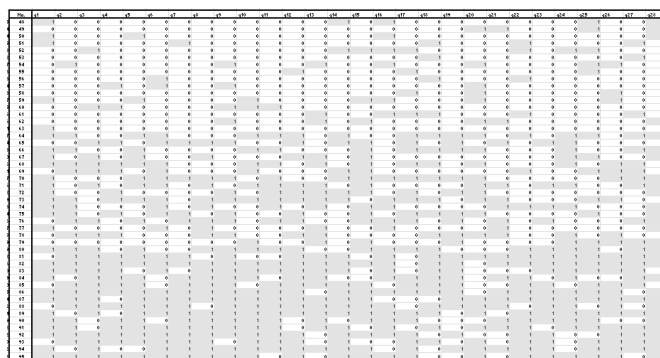


Figure 6. Design of Experiment 2 (Runs 1-16: 20% of total number of genes, runs 16-32: 50% of total number of genes, runs 33-48: 80% of total number of genes). Shaded in gray are the values of 1.

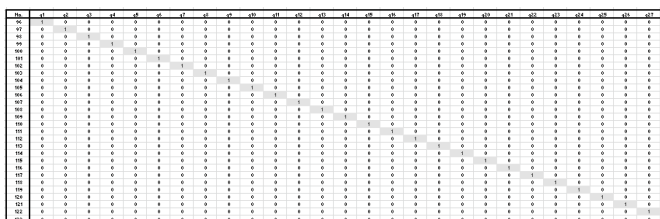


Figure 7. Design of Experiment 3. Shaded in gray are the values of 1.

Step 6. A linear discriminant analysis was carried out using the combination of genes prescribed by each run of the composite design to record the classification performance of a linear classifier.

Step 7. With the experimental design complete, a linear regression of the classification performance as a function of

the presence or absence of the 28 genes is built as shown in Table 1.

Variable	Coefficient Symbol	Regression Coefficient
	b0	0.8868
g1	b1	0.0152
g2	b2	0.0027
g3	b3	0.0097
g4	b4	0.0146
g5	b5	0.0030
g6	b6	0.0083
g7	b7	-0.0034
g8	b8	0.0051
g9	b9	0.0001
g10	b10	0.0054
g11	b11	0.0008
g12	b12	-0.0020
g13	b13	0.0120
g14	b14	-0.0027
g15	b15	0.0138
g16	b16	0.0089
g17	b17	0.0166
g18	b18	0.0145
g19	b19	0.0089
g20	b20	0.0120
g21	b21	0.0137
g22	b22	0.0105
g23	b23	-0.0068
g24	b24	-0.0025
g25	b25	0.0093
g26	b26	0.0050
g27	b27	0.0079
g28	b28	0.0158

Table 1. Linear Regression Model using 123 experimental designs.

Step 8. Using the linear regression model from Table 1, the optimization model is to find the combination of genes (through the use of binary variables) to maximize the predicted classification performance. Such optimization resulted in the identification of 23 important genes, that is, potential cervix cancer biomarkers. These are shown in Table 2.

Currently, our group is working on the validation of these potential biomarkers, as well as on their representation in a hierarchical list or a relationship network.

Index	Frontier	Accession Number	Optimization Selection
1	1	AA488645	X
2	2	H22826	X
3	3	A1553969	X
4	3	T71316	X
5	3	AA243749	X
6	3	AA460827	X
7	4	AA454831	
8	4	AA913408, AA913864	X
9	5	AA487237	X
10	5	AA446565	X
11	6	H23187	X
12	7	A1221445	
13	7	R36086	X
14	7	AA282537	
15	8	N93686	X
16	8	R91078	X
17	8	R44822	X
18	9	A1334914	X
19	9	R93394	X
20	9	AA621155	x
21	9	AA705112	x
22	9	R52794	x
23	10	AA424344	
24	10	H69876	
25	10	H55909	x
26	10	W74657	x
27	10	AI017398	x
28	10	H99699	x

Table 2. The procedure selected 23 potential biomarkers through the maximization of the expected classification performance.

4 Conclusions

In this work, a strategy to detect potential biomarkers from the analysis of microarray experiments is proposed. The

strategy is based solely on linear models and assumptions. Its consistent convergence and lack of parameter setting by the users, make this method a very competitive and attractive one for repeatability and auditability. This is especially important in high throughput experiments and in a highly interdisciplinary field like bioinformatics. A case study involving the analysis of a microarray database on cervix cancer was presented to demonstrate the capabilities of the strategy. Indeed, in this case study it was possible to discriminate among more than 10,000 genes to converge to 23 potential cervix cancer biomarkers. These are currently under analysis for validation in our research group.

5 Acknowledgement

This work was made possible thanks to the NIH-MARC grant "Assisting Bioinformatics Efforts at Minority Institutions" PAR-03-026 and BioSEI UPRM grant 330103080301.

6 References

- [1] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*. 2005 Jan 15;21(2):171-178.
- [2] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*. 2002 Apr 1;18(4):546-554.
- [3] Wong YF, Selvanayagam ZE, Wei N, Porter J, Vittal R, Hu R, Lin Y, Liao J, Shih JW, Cheung TH, Lo KW, Yim SF, Yip SK, Ngong DT, Siu N, Chan LK, Chan CS, Kong T, Kutlina E, McKinnon RD, Denhardt DT, Chin KV, Chung TK. Expression Genomics of Cervical Cancer: Molecular Classification and Prediction of Radiotherapy Response by DNA Microarray. *Clin Cancer Res*. 2003 Nov 15;9(15):5486-92.
- [4] Charnes A, Cooper WW, Lewin AY, Seiford LM. *Data Envelopment Analysis: Theory, Methodology, and Applications*. Boston MA, USA: Kluwer Academic Publishers.1993.

Comparative Analysis of Krylov Iterative Methods in Support Vector Machines

Matthew Freed, Joseph Collins, and Jeonghwa Lee

Department of Computer Science, Shippensburg University, Shippensburg, PA, U.S.A

Abstract—*Data mining and classification is a growing and important field in bioinformatics. Machine learning algorithms such as support vector machines can be used with genetic information to predict disease susceptibility. In particular, single nucleotide polymorphisms have been analyzed to classify an individual into "sick" or "healthy" categories for a specific genetic disorder. The most computationally intensive part of the support vector machine algorithm involves solving a quadratic programming problem through the use of an iterative solver. This research examines various iterative solving methods that are utilized within support vector machines. In such a solver, the solution of the problem is obtained through successively converging on an optimal result. These solvers are analyzed based on efficiency and the accuracy of the classification.*

Keywords: Data mining, gene classification, Krylov iterative methods, support vector machine

1. Introduction

With the development of the deoxyribonucleic acid (DNA) microarray technique, it has become possible to gather genetic information at lower costs [2]. The greater amount of information available has led to an effort to apply data mining and classification techniques to this information [11]. The end goal is to develop an algorithm that, when given genetic information as input, can predict an individual's susceptibility to disease.

Single nucleotide polymorphisms (SNPs) show much promise in this search. SNPs are single base changes of one nucleotide in a strand of DNA. Sets of SNPs present in a single block of DNA can be gathered together in a genotype [4]. One method that has been used to analyze genetic information is the support vector machine (SVM). This classification algorithm treats each SNP genotype as a feature vector. The SVM builds a model after reading in sets of genotypes from individuals in the "healthy" and "sick" categories [6].

In building the model, the SVM constructs a hyperplane that best separates the data points into the two categories. To do this, it solves a quadratic programming problem [6]. When dealing with genetic information, the dimensionality of the data can be very large. There may be hundreds or thousands of SNPs in each feature vector. The resulting quadratic

programming (QP) problem can be computationally intensive, and not feasibly solvable with a direct solver [12]. To overcome this, iterative solvers can be used. Iterative solvers approach a solution over many iterations to provide an approximation. In many cases, this approximation is good enough for practical purposes [12].

This paper is organized as follows: Section 2 gives a concise introduction to the implementation of the SVM. In Section 3, a selection of Krylov iterative methods are discussed in detail. Section 4 presents the numerical results of our experiments. Concluding remarks are made in Section 5.

2. Support Vector Machines

The SVM, first introduced by Vladimir Vapnik in 1992, has been established as a powerful algorithmic approach to the problem of classification, which belongs to the larger context known as supervised learning [12]. Within this supervised learning problem of classification, one is given a set of training data consisting of n individual points,

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n, \quad (1)$$

where y_i may be the value of ± 1 , which indicates the class for which \mathbf{x}_i belongs—either in (+1) or out (−1) of the set that one wishes to learn to recognize [3]. Each \mathbf{x}_i from Eq. (1) is a real vector in p -dimensions that describes the data point. The initial goal of the SVM is to locate the maximum margin hyperplane that divides the points described by $y_i = 1$ from those as $y_i = -1$. A hyperplane may be represented as the set of points \mathbf{x} which satisfies the following decision rule:

$$f(x) \equiv \mathbf{w} \cdot \mathbf{x} - b = 0, \quad (2)$$

where \mathbf{w} is a normal vector perpendicular to the hyperplane, and all training points with $y_i = 1$ lie on one side of the hyperplane, while all the training points with $y_i = -1$ lie on the other side [12]. SVMs aim to choose \mathbf{w} (a normal vector to the hyperplane) and b (some offset) to maximize the distance between the parallel hyperplanes such that they are as far apart as possible while still separating the data, hence establishing $f(\mathbf{x})$ as the decision rule. Using Eq. (2), these parallel hyperplanes may be described as follows:

$$\mathbf{w} \cdot \mathbf{x} - b = 1, \quad (3)$$

and

$$\mathbf{w} \cdot \mathbf{x} - b = -1. \quad (4)$$

For training data that are linearly separable—that is, two sets of points in p -dimensions that may be separated by a hyperplane—one may select the two hyperplanes of the margin in such a way that there are no points between them and then try to maximize their distance. As one increases the size of the margin, one must prevent data points from falling into it. To ensure that this does not occur, one must utilize the following constraints on Eq. (3) – (4):

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \quad \text{when } y_i = +1 \quad (5)$$

and

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \quad \text{when } y_i = -1. \quad (6)$$

Eq. (5) – (6) represent parallel hyperplanes that separate the data, which—together—are referred to as the fat plane [12]. Eq. (5) – (6) may be rewritten as

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \text{for all } 1 \leq i \leq n. \quad (7)$$

Using geometry, the perpendicular distance between these parallel hyperplanes (twice the margin) is

$$2 \times \text{margin} = 2(\mathbf{w} \cdot \mathbf{w})^{-\frac{1}{2}}. \quad (8)$$

Utilizing Eq. (7) – (8), one may construct the fattest possible fat plane, known as the maximum margin SVM [12], by solving a particular problem in quadratic programming:

$$\text{minimize: } \frac{1}{2} \mathbf{w} \cdot \mathbf{w} \quad (9a)$$

$$\text{subject to: } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \quad i = 1, \dots, m. \quad (9b)$$

When arriving to the solution of Eq. (9), some of the training data points will lie on the extreme boundaries of the fat plane, denoted the support vectors [12]. The Krylov iterative methods for solving this quadratic programming problem are discussed in the following section.

3. Krylov Iterative Methods

Given a square system of n linear equations with a vector of unknowns \mathbf{x} , we may construct the following matrix equation:

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad (10)$$

where the components of Eq. (10) may be represented as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}. \quad (11)$$

From Eq. (11), \mathbf{A} may then be decomposed into a diagonal component \mathbf{D} and strictly lower and upper triangular components \mathbf{L} and \mathbf{U} :

$$\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}, \quad (12)$$

where the components of Eq. (12) may be represented as

$$\mathbf{D} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}, \quad (13a)$$

$$\mathbf{U} = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}. \quad (13b)$$

The system of linear equations from Eq. (10) – (13) may be rewritten as

$$(\mathbf{D} + \omega\mathbf{L})\mathbf{x} = \omega\mathbf{b} - [\omega\mathbf{U} + (\omega - 1)\mathbf{D}]\mathbf{x} \quad (14)$$

for a constant $\omega > 1$ [1]. Various iterative methods exist to solve the expression of Eq. (14).

3.1 Successive Overrelaxation Method

The successive overrelaxation (SOR) method is derived by extrapolating the Gauss-Seidel method [8]. This extrapolation takes the form of a weighted average between the previous iterate and the computed Gauss-Seidel iterate successively for each component:

$$x_i^k = \omega \bar{x}_i^k + (1 - \omega)x_i^{k-1}, \quad (15)$$

where \bar{x} represents a Gauss-Seidel iterate and ω is the extrapolation factor [1].

In matrix terms, the SOR algorithm may be written as follows:

$$\mathbf{x}^k = (\mathbf{D} - \omega\mathbf{L})^{-1}[\omega\mathbf{U} + (1 - \omega)\mathbf{D}]\mathbf{x}^{k-1} + \omega(\mathbf{D} - \omega\mathbf{L})^{-1}\mathbf{b}, \quad (16)$$

where the matrices \mathbf{D} , \mathbf{L} and \mathbf{U} represent the diagonal, strictly lower-triangular and strictly upper-triangular parts of \mathbf{A} from Eq. (13), respectively [1].

The underlying success behind SOR is to choose a value for ω that accelerates the rate of convergence. When $\omega = 1$, the SOR method simplifies to the Gauss-Seidel method [1], yet it will fail to converge if $\omega \notin \{0, 2\}$ [8]. Generally speaking, it is impossible to choose the most desirable value for ω in advance, thus it is common to utilize the following heuristic estimate:

$$\omega = 2 - O(h), \quad (17)$$

where h is the mesh spacing of the discretization of the underlying physical domain [1].

3.2 Quasi-Minimal Residual Method

Iterative methods often exhibit irregular convergence behaviors. A related algorithm, known as the quasi-minimal residual (QMR) method [5], attempts to overcome this problem. The underlying idea behind this algorithm is to

solve the reduced tridiagonal system in a least squares sense. QMR also uses look-ahead techniques to avoid breakdowns in its Lanczos process, which makes it more robust than SOR [1].

3.3 Biconjugate Gradient Method

The biconjugate gradient (BiCG) method [9] is commonly used in solving systems of linear equations. It is a generalized form of the conjugate gradient method, in that it can be applied to matrices that are non-symmetric. To formulate the biconjugate gradient method as an iterative method, it is necessary to use a metric at each iteration to determine if the approximation vector \mathbf{x} is closer to the solution \mathbf{x}_* . It has been shown that this solution is also the minimizer for the quadratic function [13]:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} - \mathbf{x}^T \mathbf{b}, \quad \mathbf{x} \in \mathbf{R}^n. \quad (18)$$

Therefore, as Eq. (18) is smaller than the previous iteration, the value of \mathbf{x} is closer to the solution. Starting with an initial guess \mathbf{x}_0 , the gradient of the function will be $\mathbf{A}\mathbf{x}_0 - \mathbf{b}$. If \mathbf{x}_0 is assumed to start at 0, the first basis vector p_1 will equal \mathbf{b} . Each of the other vectors in the basis is conjugate to the gradient of the function.

The error at each iteration is measured as the residual. This residual is defined as

$$r_k = b - Ax_k. \quad (19)$$

In calculating the basis vectors at each step, this residual from Eq. (19) is taken into account in order to move the approximation towards the solution. The value of x therefore is updated during each iteration to

$$x = x + \alpha \times p, \quad (20)$$

where α is based on the residual divided by $A \times p$.

The resulting algorithm involves two matrix vector products, including one transpose product. This is the major computational cost of the biconjugate gradient method [10].

3.4 Biconjugate Gradient Stabilized Method

The biconjugate gradient stabilized (BiCGSTAB) method [14] is a variation of the biconjugate gradient and conjugate gradient squared (CGS) methods. BiCGSTAB makes several improvements, most importantly in that it stabilizes the algorithm. CGS relies on the squaring of the residual, which can result in rounding errors that affect the approximation in greater amounts at each iteration. As a result, the convergence pattern may be irregular. BiCGSTAB smooths this convergence by updating the way the approximation x is determined at each step:

$$x = x + \alpha \times p + \omega \times s. \quad (21)$$

Here, ω from Eq. (21) is a scaling factor that allows the distance that the approximation changes to vary. Larger steps may be taken during iterations, which assists in speeding up

convergence. The stabilizer s is what allows for a smoother convergence. It is based on the residual and the matrix:

$$s = r - \alpha A \times p. \quad (22)$$

Eq. (22) results in avoiding the irregular convergence that is associated with BiCG. Further, there is no transpose involved in this algorithm, which is often desirable for solving certain matrices [13].

4. Numerical Results

The dataset used with the SVM was taken from publicly available genetic information. It is derived from the 616 kilobase region on Chromosome 5q31 [4]. Within this region may contain the genetic variation that is responsible for Crohn's disease [3]. The data contains a total of 103 genotyped single nucleotide polymorphisms for each of the 387 genotypes. Of this, 144 of them are case and 243 are control.

The SVM package chosen was the Mangasarian-Musicant variation due to its brevity [12]. Each of the genotypes was treated as a feature vector and read as input to the SVM. The kernel function increased the dimensionality of the dataset using a linear kernel. Half of the genotypes were used as the training set. During the solving of the SVM, each of the four different iterative solvers was used to solve the quadratic programming portion. The other half of the genotypes was then classified using the training data. Table 1 shows the resulting data that was collected.

For each of the solvers, we measured the efficiency of the solver as well as the accuracy of the classification resulting from the solution. The number of iterations each solver took to converge as well as the time it took can be considered a measure of its efficiency. For the classification, a simple accuracy measurement consisting of the percent of genotypes correctly classified. The sensitivity and specificity of the data was also taken. The sensitivity is the proportion of individuals who have the disease and are correctly identified as such. The specificity is the proportion of individuals who do not have the disease and are correctly identified.

All of the solvers except QMR were able to achieve convergence within the maximum number of iterations. QMR terminated after 43 iterations as a result of a breakdown in the gamma variable. However, it still was able to produce a viable classification. The QMR algorithm is not as robust as some of the other methods, and the algorithm failed on this particular matrix.

Out of the four solvers, BiCG was the most accurate, correctly placing 64.1% of the SNPs into the proper category. The classification of BiCGSTAB was similar with a 62.1% accuracy. SOR and QMR both resulted in classifications with 60.2% accuracies. The accuracies of each of the solvers were relatively similar. Further, the results were comparable to the same dataset used with other SVM packages. In particular,

Table 1: Classification results of the QP solvers

Solver	Iterations	Solve Time (sec)	Accuracy (%)	Specificity (%)	Sensitivity (%)
SOR	60	0.0500	60.2	51.9	68.6
QMR	43	0.0050	60.2	51.9	68.6
BiCG	149	0.0020	64.1	61.5	66.7
BiCGSTAB	63	0.0012	62.1	53.8	70.5

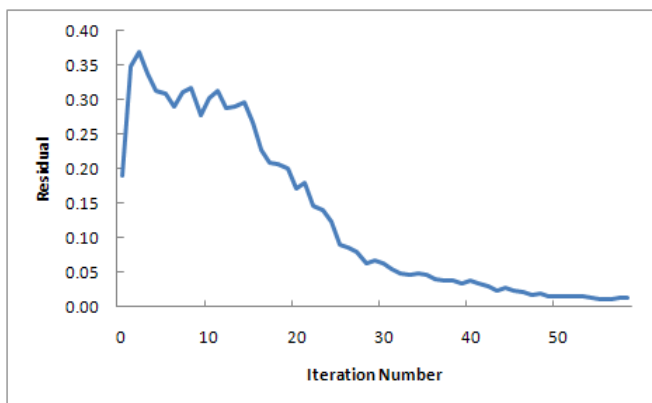


Fig. 1. Convergence history of the SOR solver

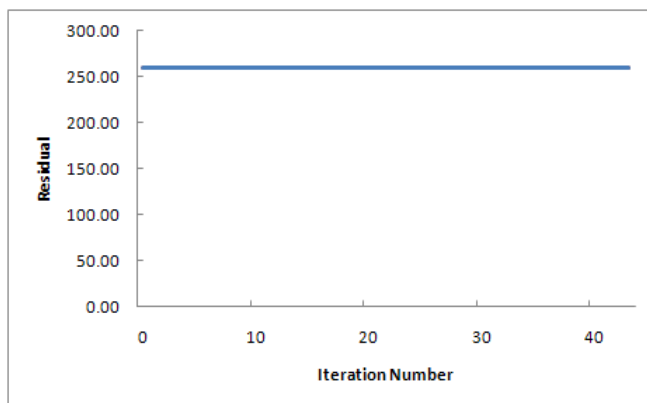


Fig. 2. Convergence history of the QMR solver

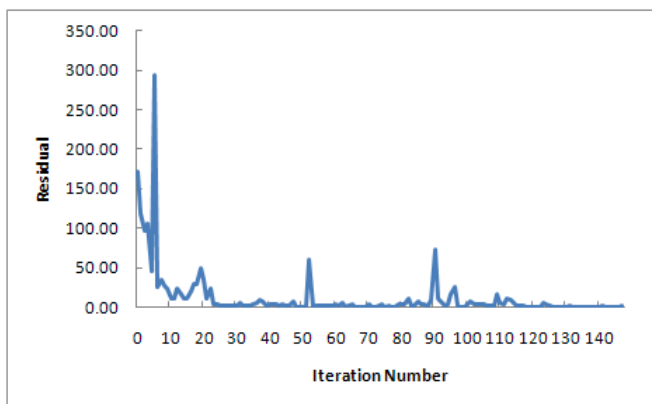


Fig. 3. Convergence history of the BiCG solver

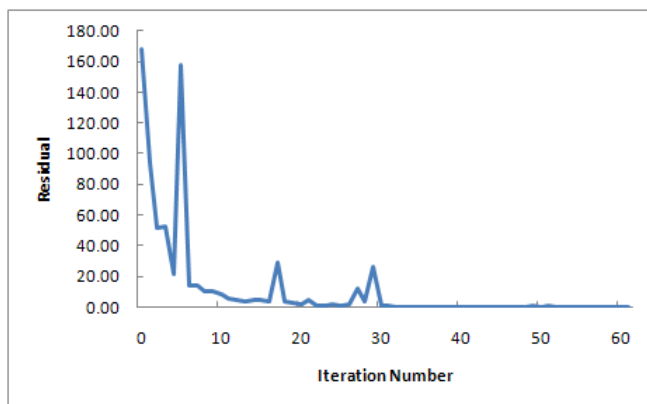


Fig. 4. Convergence history of the BiCGSTAB solver

the commonly used SVM-Light package resulted in 62% accuracy [7].

The convergence histories of the solvers can be seen in Figures 1–4. The failure of the QMR method can be seen as no change in the residual. SOR converges slower than BiCG and BiCGSTAB. These two solvers quickly reached a low residual value, and then slowly converged within the tolerance. The stabilizing effect of the BiCGSTAB algorithm over the BiCG can clearly be seen. The convergence history of BiCG is erratic, with many increases in the residual. The BiCGSTAB smooths this, resulting in a much more stable convergence.

In terms of time, BiCGSTAB was clearly the most effi-

cient solver, converging in 0.0012s. Of note is that not all times correlated with the number of iterations. For example, SOR iterated a similar number of times as BiCGSTAB. However, each iteration took a greater amount of time, and as a result took longer to converge. The BiCG took the greatest number of iterations. The time per iteration was smaller than QMR and SOR, and as a result converged in less time.

5. Conclusion

Support vector machines can be applied confidently to the problem of classification of genetic data. As more and more genetic information becomes available, classification algorithms such as the SVM can be used to make useful

models based on the data. With the addition of sophisticated iterative methods, an accurate solution can be achieved in less time.

The numerical results demonstrate the efficiency of various iterative solvers. As can be seen with the failure of QMR, certain methods may not be applicable with certain matrices. The BiCGSTAB provided a model with high classification accuracy. It is also the most efficient of the methods examined in terms of time. Overall, the results suggest that BiCGSTAB is a robust algorithm that is a good choice for solving large quadratic programming problems. This experiment may assist researchers in selecting an iterative method when dealing with data mining using genetic information.

References

- [1] Barrett, R., Berry, M., Chan, T. F.; Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C. & van der Vorst, H., *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, 2nd ed. Society for Industrial and Applied Mathematics, Philadelphia, PA, (1994).
- [2] Caron, L., Bell, J., *Association Study Designs for Complex Diseases*, Nature Reviews: Genetics, 91–98, (2001).
- [3] Cortest, C. & Vapnik, V., *Support-Vector Networks*, Machine Learning, Vol. 20, No. 3, 273–297, (1995).
- [4] Daly, M., Rioux, J., Schaffner, S., Hudson, T. & Lander, E., *High Resolution Haplotype Structure in the Human Genome*, Nature Genetics, Vol. 29, 229–232, (2001).
- [5] Freund, R. & Nachtigal, N., *QMR: A Quasi-Minimal Residual Method for Non-Hermitian Linear Systems*, Numer. Math., Vol. 60, 315–339, (1991).
- [6] Guyon, I., Weston, J., Barnhill, S., *Gene Selection for Cancer Classification using Support Vector Machines*, Machine Learning, 389–422, (2002).
- [7] Joachims, T. *Making Large Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning*, MIT (1999).
- [8] Kahan, W., *Gauss-Seidel Methods of Solving Large Systems of Linear Equations*, Ph.D. Thesis, Toronto, Canada, University of Toronto, (1958).
- [9] Lanczos, C., *Solution of Systems of Linear Equations by Minimized Iterations*, J. Research National Bureau Standards, Vol. 49, 33–53, (1952).
- [10] Lee, J., Zhang, J. & Lu, C., *Performance of Preconditioned Krylov Iterative Methods for Solving Hybrid Integral Equations in Electromagnetics*, Journal of Applied Computational Electromagnetics Society, Vol. 18, No. 4, 54–61, (2003).
- [11] Mao, W., Lee, J., *A Combinatorial Analysis of Genetic Data for Crohn's Disease*, Journal of Biomedical Science and Engineering, 52–58, (2008).
- [12] Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P., *Numerical Recipes: The Art of Scientific Computing*, 3rd ed., Cambridge University Press, Cambridge, (2007).
- [13] Saad, Y., *Iterative Methods for Sparse Linear Systems*, 2nd ed., Society for Industrial and Applied Mathematics, (2003).
- [14] van der Vorst, H., *Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems*, SIAM J. Sci. Stat. Comput., Vol. 13, 631–644, (1992).

Fast Splice Site Classification Using Support Vector Machines in Imbalanced Data-sets

Jair Cervantes¹, Asdrúbal López Chau², Adrian Trueba Espinoza¹ and José Sergio Ruiz Castilla¹

¹Department of Computer Sciences, UAEM-Textcoco, Textcoco, MX 56259, México
chazarra17@gmail.com

²Department of Computer Sciences, CINVESTAV-IPN, México D.F. 07360, México.
achau@computacion.cs.cinvestav.mx

Abstract—*Splice sites prediction is an important objective of genome sequencing. In last years, careful attention has been paid in order to the improve the performance of the algorithms used, but the study of most feasible methods to improve the performance in large and imbalanced data-sets is still of immense importance. This paper presents a novel SVMs classification method which works with gene data, the proposed method reduces significantly the training time and obtain a high accuracy on huge and imbalanced data-sets. Experimental results show that the accuracy obtained by the proposed algorithm is slightly better (98.9%) in comparison with other SVMs implementations such as SMO (98.6%), LibSVM (98.6%), and Simple SVM (98.2%). Furthermore the proposed approach can be used in large and imbalanced data-sets obtaining high classification accuracy.*

Keywords: SVM, Splicing, Imbalanced data-sets

1. Introduction

The advances and development in DNA sequencing technologies have resulted in a impressive increase in the size of genomic sequences. This growth of sequence data demands effective techniques to processing huge amounts of biological information. Identifying genes is an important issue in bioinformatics, and the accurate identification of splice sites in DNA sequences plays one of the central roles of gene structural prediction in eukaryotic cells. An effective detection of splice sites requires the knowledge of characteristics, dependencies, relationship of nucleotides in the splice site surrounding region and an effective encoding method.

The classification of gene sequence into regions that code for genetic material and regions that do not is a challenging task in DNA sequence analysis. It is not an easy challenge. It is due to size of DNA sequences and sometimes regions that encode in proteins (exons) can be interrupted by regions that do not encode (introns). These sequences are characterized, however they are not clearly defined by local characteristics at splicing sites. Identifying exons into DNA sequences presents a computational challenge. In some organisms the introns are small regions and the splicing sites are fully characterized. However, in some other sequences, including

human genome, it is a great problem to localize the correct transition between the regions that encode and the ones that not. Furthermore, the genes in many organisms splice of different way, which complicates considerably the task. On the other hand, splice sites fall into two categories: donor sites of introns and acceptor sites of introns. These sites display some characteristic patterns, e.g. 99% of donor sites begin with base pairs GT while 99% acceptor sites end with based pairs AG. However, not all locations with base pairs GT or AG are necessarily splice sites. Some occurrences of AG or GT occur outside of a gene or inside an exon. These are called decoys, because they do not indicate the presence of a splice site. Furthermore, the majority of gene data-sets are imbalanced and the bulk of classifiers generally perform poorly on imbalanced data-sets because making the classifier too specific may make it too sensitive to noise and more prone to learn an erroneous hypothesis. Another factor is that in imbalanced data-sets an instance can be treated as noise and ignored completely by the classifier. Due to it, efficient methods and fast techniques that aims to tackle this problem are necessary.

In this paper, we use a novel approach for train and predict acceptor and donor splice sites in huge and imbalanced data-sets using Support Vector Machines (SVM). SVM has received considerable attention due to its optimal solution, discriminative power and performance. Lately some SVM classification algorithms have been used in splice site detection with acceptable accuracies [1] [2] [3] [10] [12] [14]. Cheng et al [2] use SVMs in order to predict mRNA polyadenylation sites [poly(A) sites] the method can help identify genes, define gene boundaries, and elucidate regulatory mechanisms. Damaevicius [3] and Xia [12] use SVMs in order to detect splice-junction (intron-exon or exon-intron) sites in DNA sequences. In [14] the authors use a SVM in order to discover sequence information that could be used to distinguish real exons from pseudo exons. Baten et al. [1] make use of SVM with polynomial kernel in order to obtain an effective detection of splice sites, the authors used a first order Markov model as a pre-processing step of DNA sequences. Some authors have been using SVM for the detection of splicing sites. However, when faced SVM with imbalanced data-sets the performance of SVM drops

significantly. Other important disadvantage of SVMs is due to memory requirements grows with square of input data points, so training complexity of SVMs is highly dependent on the size of a data-set.

This paper presents a novel splice sites fast classification model using SVM for imbalanced data-sets. The proposed method reduces intelligently the input data-set, tackling the problem of imbalanced data-sets with SVM and reducing significantly the training time. The rest of the paper is organized as following: Section II reviews some preliminaries of SVM. Section III focuses on explaining the methodology of proposed SVM classification algorithm. Section IV shows experimental results. Conclusions are given in Section V.

2. Preliminaries

2.1 Support Vector Machines

Support Vector Machines aim at estimating an optimal classification function using labeled training data from X_{tr} such that f will correctly classify unseen examples (test data). In our case, input space X will contain simple representations of sequences A, C, G, T while corresponds to true splice and decoy sites, respectively. Considering binary classification, we assume that a training set X_{tr} is given as:

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_n, \mathbf{y}_n) \quad (1)$$

i.e. $X_{tr} = \{x_i, y_i\}_{i=1}^n$ where $x_i \in R^d$ and $y_i \in \{+1, -1\}$ is the label of example x_i . The generated classification function can be written as

$$g(x) = \text{sign} \left(\sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b \right) \quad (2)$$

where $\mathbf{x} = [x_1, x_2, \dots, x_l]$ is the input data, α_i and y_i are Lagrange multipliers. SVM training obtain a set of real-valued weights $\alpha_i \geq 0$ such the normal vector can be expressed as a linear combination of input vectors, $w = \sum_{i=1}^n y_i \alpha_i x_i$. Input vectors x_i having non-zero weight are called support vectors and they determine the SVM solution. Once the SVM is trained, a new object x can be classified using (2). The vector \mathbf{x}_i is shown only in the way of inner product. The α_i s are Lagrange multipliers and b is the usual bias which are the result of SVM training.

The principal disadvantage of SVMs is due to complexity that grows with square of input data points. Sequential minimal optimization (SMO) breaks the large Quadratic Programming (QP) problem into a series of smallest possible QP problems [9]. These small QP problems can be solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. The memory required by SMO is linear in the training set size, which allows SMO to handle very large training sets [9]. A requirement in (3) is $\sum_{i=1}^n \alpha_i y_i = 0$, it is enforced throughout the iterations and implies that the smallest number of multipliers can be

optimized at each step is two. At each step SMO chooses two elements α_i and α_j to jointly optimize, it finds the optimal values for these two parameters while all others are fixed. The choice of the two points is determined by a heuristic algorithm, the optimization of the two multipliers is performed analytically.

2.2 Methods for imbalanced classification

The classification of imbalanced data-sets is a crucial problem in machine learning because it normally causes negative effects on the performance of a classification method. There are two methods to tackle this problem. At the data level, re-sampling training data is a popular solution to classification of imbalanced data-sets, the most important techniques used at the data level or by preprocessing data exist are Over-sampling and Under-sampling.

2.2.1 Over-sampling

This technique over samples the minority class to balance the class distribution of a training data-set. Specifically, the minority class is over sampled until the size is equal to the size of the maximum class. Over sampling is a popular technique tackle some imbalanced classification problems. However in SVM increases significantly the training time.

2.2.2 Under-sampling

This technique under samples the majority class to balance the class distribution of a training data-set. Specifically, the majority class is under sampled until the size is equal to the size of the minimum class. Some previous studies showed that under sampling was better than over sampling in classification of imbalanced data-sets. It should also noted that under sampling usually reduces the training time but discard some potentially useful training examples and may degrade the performance of the classifier.

On the other hand, at the algorithmic level, weighting training data assign a larger weight to the minority class in order to balance the input data-set.

3. Methodology

In the following, we describe the methodology for splice sites recognition. Given a sequence, the proposed algorithm starts by encoding the DNA sequences. DNA encoding is crucial to successful intron/exon prediction. The next step is done by training SVMs on the training data and tuning their hyperparameters on the validation data.

3.1 DNA Encoding

DNA encoding has been extensively researched in recent years [5][8]. Each technique is based on the most important features to be shown. Sparse encoding is a widely used encoding schema which represents each nucleotide with 4 bits: $A \rightarrow 1000, C \rightarrow 0100, G \rightarrow 0010$ and $T \rightarrow 0001$ [7].

Suppose we have a DNA sequence of AGGCGTATGAGG. With the sparse encoding, the sequence is represented as: 1000 | 0010 | 0010 | 0100 | 0010 | 0001 | 1000 | 0001 | 0010 | 1000 | 0010 | 0010. where | is a virtual separator used to illustrate the example.

We use 18 additional features with the sparse encoding schema. The first 16 components define the nucleotide pairs into a DNA sequence, which are defined as $\beta = \{(x_{AA}), (x_{AC}), (x_{AG}), (x_{AT}), \dots, (x_{TA}), (x_{TC}), (x_{TG}), (x_{TT})\}$. When some nucleotide pair is in the sequence, it is marked with 1 and an absence of this pair is marked with 0. The DNA sequence, AGGCGTATGAGG can be encoding by this schema as: 0 0 1 1 0 0 1 0 1 1 1 1 1 0 1 0.

The last two components correspond to the informative function of each triples in the sequence ranked by their *F-value*. For each triple, we specify its location relative (pre and post) and its mean frequency among exons and decoys $\mu_k^+ - \mu_k^-$ respectively.

The *F-value* criterium is that used by Golub et al [6]. For each triple $x_k, k = 1, \dots, n$, we calculated the mean $\mu_k^+(\mu_k^-)$ and the standard deviation $\sigma_k^+(\sigma_k^-)$ using positive and negative examples. The *F-value* criterium is given by

$$F(x_k) = \left| \frac{\mu_k^+ - \mu_k^-}{\sigma_k^+ + \sigma_k^-} \right| \quad (3)$$

where x_k is the k -esime triple, the *F-value* serves as a simple heuristic for ranking the triples according to how well they discriminate. The last point in the vector is represented by the relative presence of each triple of nucleotides. If this sequence AGGCGTATGAGG belong to data-set of example 1 can be encoding by this schema as: $\gamma = \{f_{AGG}, f_{AGG}\} = \{0.231, 0.231\}$, where γ is computed using the *F-value* criterium. The *F-value* is repeated because the triple AGG is in the sequence pre and post (AGG...CGTATG... AGG).

The proposed encoding schema allows to obtain the nucleotides of each sequence, encoding the pairs show the importance of some pairs in the sequence, and obtain the importance of each triple at the begin and at the end of each sequence. The previous DNA sequence can be encoding by the complete schema as: 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0 | 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0 | 0.231, 0.231. Where | is a virtual separator which objective is just illustrate the three techniques used. With the proposed encoding schema SVM can use the features and discriminate between the categories.

1000 | 0010 | 0010 | 0100 | 0010 | 0001 | 1000 | 0001 | 0010 | 1000 | 0010 | 0010.

3.2 Classification algorithm

SVM classification aim at estimating a classification function $H : X \rightarrow \{\pm 1\}$ using labeled training data from $X \times \{\pm 1\}$ such that H will correctly classify unseen examples (testing data). In our case, input space X will contain

simple representations of sequences $\{A, C, G, T\}^N$, while ± 1 corresponds to true splice and decoy sites, respectively.

Learning with imbalanced data is one of the recent challenges in machine learning. There are some techniques proposed in order to find a solution for this problem, such as the application of a preprocessing stage focused on balancing data, in preprocessing data two tendencies exist: reduce the set of examples (under-sampling) or replicate minority class examples (over-sampling). Over-sampling of minority classes can be done by re-sampling the examples from minority classes thus increasing the bias of the learned classifier towards them and increasing the accuracy on minority classes. Under-sampling with imbalanced datasets could be considered as a prototype selection procedure which the majority class can reduce the bias of the learned classifier towards it and thus improve the accuracy on the minority classes. In this paper, we used under-sampling, the selection process under-sample the majority class in order to remove noisy and redundant training instances however the proposed algorithm recover the most important data points and the outliers keeping all the information in the training data-set. Our goal in this case is to retain and use this information, because even though under-sampling the majority class provokes an inherent loss of valuable information.

INPUT: X_{EDS}

// X_{EDS} ; Entire Imbalanced data-set

OUTPUT: $H_f : \{x_i \in x_{EDS} : x_i \in SVS\}$;

Initialization;

1. $X_r^+ \leftarrow 0$ /* training data-set with positive labels begins empty */
2. $X_r^- \leftarrow 0$ /* training data-set with negative labels begins empty */
3. $X_r^+ \leftarrow \{x_i \in x_{EDS} : y_i = +1\}$, $i = 1, 2, \dots, p$;
4. $X^- \leftarrow$
get_RandomSampling $\{x_i \in x_{EDS} : y_i = -1\}$, $i = 1, 2, \dots, p$;
5. Obtain outliers (O^+, O^-) using Algorithm 2;
6. Obtain (X_f^+, X_f^-) using Algorithm 2;
7. $X_{RD}^+ \leftarrow (X_f^+ \cup O^+)$;
8. $X_{RD}^- \leftarrow (X_f^- \cup O^-)$;
9. $H_f(X_{RD}^+, X_{RD}^-) \leftarrow \text{trainSVM}(X_{RD}^+, X_{RD}^-)$;
10. **return** $H_f(X_{RD}^+, X_{RD}^-)$

Algorithm 1: SVM training

In this paper, we propose a fast SVM algorithm to work with imbalanced data-sets. The proposed algorithm is based in the sparse property of SVM When using SVM for classification, in most cases has been found that after the

training, the number of SV is very small compared with the number of elements of the training data-set, so taking advantage of this fact, the basic idea behind the reduction of the training data-set strategy is to select elements most likely to be SV. The Algorithm 1 shows the general process to detect splices sites or decoys by our technique.

The first step in the proposed algorithm consists in obtain the minority class which contains p instances, in the imbalanced data-set and label them as positive X_r^+ , we also randomly select from the entire data-set X_{EDS} and label them as negative X_r^- .

X_r^+ and X_r^- are used by the algorithm 2 in order to find an introductory hyperplane $H_1(X_r^+, X_r^-)$, from H_1 we obtain SV, non-SV and $O^+ \cup O^-$ by testing the hyperplane obtained in the entire data-set, the data-set $O^+ \cup O^-$ contains all data points that are misclassified with H_1 and contains valuable information in this process. In order to obtain the most important data points in the entire data-set we train a SVM and obtain $H_2(X_{ch}^+, X_{ch}^-)$ where X_{ch}^+ and X_{ch}^- represent the data points that are SV and non SV with H_1 respectively. Testing H_2 in the entire data-set we obtain the most important data points and eliminate redundant training instances.

The small size of (X_{RD}^+, X_{RD}^-) contributes to speed up the training of the proposed method. Furthermore, the reduced data-set obtained contains the most important data points in the entire data-set.

∩

INPUT: X_r^+, X_r^-

// X_{Tr} ; Training data-set

OUTPUT: X_f^+, X_f^-, O^+, O^- ;

Initialization;

1. $H_1(X_r^+, X_r^-) \leftarrow \text{trainSVM}(X_r^+, X_r^-)$;
2. $SV \leftarrow \text{get_SV}(H_1(X_r^+, X_r^-))$;
3. $\text{nonSV} \leftarrow \text{get_nonSV}(H_1(X_r^+, X_r^-))$;
4. $X_r^+ \leftarrow 0$ /* positive outliers or misclassified data points with H_1 are empty */;
5. $X_r^- \leftarrow 0$ /* negative outliers or misclassified data points with H_1 are empty */;
6. $O^+ \cup O^- \leftarrow \text{testing_SVM}H_1(X_r^+, X_r^-)$;
7. $X_{ch}^+ \leftarrow SV$;
8. $X_{ch}^- \leftarrow \text{nonSV}$;
9. $H_2(X_{ch}^+, X_{ch}^-) \leftarrow \text{trainSVM}(X_{ch}^+, X_{ch}^-)$;
10. $(X_f^+, X_f^-) \leftarrow \text{testing_SVM}H_2(X_{ch}^+, X_{ch}^-)$;
11. **return** X_f^+, X_f^-, O^+, O^- .

Algorithm 2: Proposed under-sampling algorithm

The main advantages of proposed model include a) it can make use of the discriminative features (features which show relevant differences between true splices sites and decoys),

reducing the influence of some irrelevant and redundant features; b) it can work on imbalanced data-sets, the algorithm implements an undersampling technique in order to balance the data points and recover the most important data points in the data-set, retain valuable information with the proposed process; c) The training time obtained with the proposed method is very fast in comparison with other fast SVM implementations.

4. Experimental Results

In this section, we describe the methodology used and show the results obtained with the proposed algorithm,

4.1 Metrics for Imbalanced Classification

In order to evaluate classifiers on highly imbalanced data-sets, is necessary to use an adequate metric. With highly skewed data distribution, the overall accuracy metric is not sufficient any more. This is because with an imbalance of 99 to 1, a classifier that classifies everything negative will be 99% accurate, but it will be completely useless as a classifier to detect rare positive samples.

The medical community, and increasingly the machine learning community, use two metrics, the sensitivity and the specificity, when evaluating the performance of various tests. The sensitivity is the performance of proposed SVM to calculate the proportion of noncoding nucleotides that have been correctly predicted as noncoding and it is evaluated as

$$S_n^{false} = \frac{T_N}{T_N + F_P} \quad (4)$$

S_n is the proportion of candidate sites in the testing data-set that have been correctly predicted and it is expressed as

$$S_n = \frac{N_c}{N_t} \quad (5)$$

S_n^{true} is the proportion of coding nucleotides that have been correctly predicted as coding, i.e.,

$$S_n^{true} = \frac{T_P}{T_P + F_N} \quad (6)$$

where T_P is the number of sequences with real splice sites which are predicted to be true (true positives), T_N is the number of sequences without real splice sites which are predicted to be false (true negatives), F_P is the number of sequences without real splice sites which are predicted to be true (false positives) F_N is the number of sequences with real splice sites which are predicted to be false (false negatives), N_c is the number of exons that have been correctly predicted in the testing data-set, and N_t is the total number of exons sites in the testing data-set.

The receiver operator characteristic curve (ROC) analysis describes the sensitivity and specificity of a classification model using graphics. It is considered as an effective method to assess the performance of a classification method. We also used this metric to evaluate our classifier. We also list the

sensitivity and specificity separately to give the reader an even better idea of the performance of our classifier.

4.2 Model selection

SVM training involves to fixing several parameters. The parameters chosen have a crucial effect of the performance of the trained classifier. To be able to apply the SVM, we select the radial basis function (RBF) kernel function to train the SVM. The RBF kernel function is defined as

$$K(x_i - x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \quad (7)$$

we have to find the complexity parameter C and γ , controlling the tradeoff between training error and complexity, and the kernel parameters. In order to identify an optimal hyperparameter set, we applied a “grid search” on C and γ using cross-validation.

4.3 Examples

In order to show the experimental results of the proposed method, we use two examples. First example is a small data-set with balance data-set, but the second example is an imbalanced and large data-set example.

4.3.1 Example 1

We use Primate splice-junction gene sequences(DNA) taken from Genbank64.1 (ftp site: genbank.bio.net).The DNA data-set contains 3190 DNA sequences with 62 descriptors for each sequence, 767exon/intron boundaries(referred to as EI sites), 768 intron/exon boundaries(referred to as IE sites) and 1655 neither.

In this example, we use 80% of the input data to train the SVM and 20% to test. The SVM was trained and evaluated 20 times, the experimental results are shown in the Table I. It shows the experimental results obtained with the proposed approach with the average accuracy (Acc) and the standard deviation(SD). The results obtained with S_n^{false} , S_n and S_n^{true} provide a good measure of the classifier. However, in this case the data-set is very small, the training time is almost the same with some SVM implementations like SimpleSVM, Libsvm, Sequential Minimal Optimization(SMO), but when the training data-set is large the training time grows exponentially.

Table I

Genbank 64.1 data-set			
	Av_EI	Av_IE	Av_Neither
Acc	99.37	99.18	97.8
SD	0.16	0.27	0.24
S_n^{true}	0.99	0.98	0.97
S_n^{false}	0.99	0.98	0.97
S_n	0.99	0.99	0.97

Acc.-average accuracy, SD.- standard deviation.

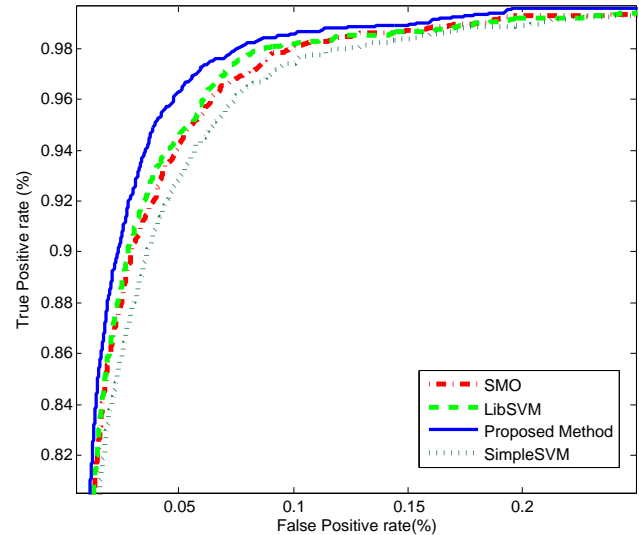


Fig. 1: ROC curves of the four classifiers. The proposed method, LibSVM, SMO and SimpleSVM.

4.3.2 Example 2

The second example is acceptor/donor data-set which was obtained from <http://www2.fml.tuebingen.mpg.de/raetsch/projects/>.

The data-set contains 91546 training data points and 75905(2132 true sites) testing data points for acceptors and 89163 training data points and 73784(2132 true sites) testing data points for donors. In this example we show the difference of training time between the proposed approach and other fast SVM implementations.

The Figure 1 shows the ROC curves obtained with the proposed algorithm, The AUC for the proposed method, LibSVM, SMO and SimpleSVM are 0.9894, 0.9860, 9865 and 9823 respectively. The Figure 2 shows the discriminative power of the proposed method, in the Figure 2 are shown the AUC of LibSVM with only the sparse encoding and the AUC of proposed method. It is clear that, a set of highly discriminative features could significantly improve the classification accuracy. Some features were added with the purpose of enhancing the classifier performance. Moreover, not only in the performance measure is more robust, but also we get a small training time as can we see in the Table II.

Table II

Algorithm	Acceptor data-set		Donnor data-set	
	t	AUC	t	Acc
Proposed App	469	98.9	673	98.7
LIBSVM	6371	98.6	4924	98.5
SMO	123493	98.6	104525	98.4
SimpleSVM	432919	98.2	381049	98.1

traininig data, t training time in seconds, Acc accuracy.

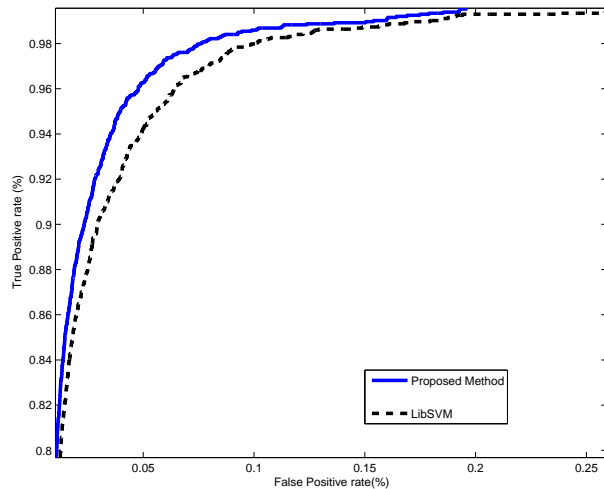


Fig. 2: ROC curves of the four classifiers. The proposed method, LibSVM, SMO and SimpleSVM.

5. Conclusions

In this paper we present a novel SVM classification approach for large data-sets using imbalanced data-sets. In order to reduce SVM training time for large data-sets, we use a modified algorithm which overcomes the drawback that only part of the original data near the support vectors are trained. Experiments done with real world data-sets, show that the proposed method has advantage in large data-sets. Furthermore, not only in the training time is more robust, but also we get much area under the ROC curve, providing an adequate measure for the quality of the classifier. Some features have been proposed for the classification Donor/acceptor. introducing a new encoding method. However, not all features are equally effective for the classification task. Therefore, the careful choice of features is crucial for building accurate splice detectors and if an appropriate system for imbalanced data-sets is implemented, the SVM classifier easily outperform previously proposed methods. Choosing a set of highly discriminative features could significantly improve the classification accuracy. In this work, we study the some features with the purpose of enhancing the classifier performance, and improve significantly the training time used with other fast SVM implementations.

References

- [1] AKMA Baten, BCH Chang, SK Halgamuge and Jason Li. "Splice site identification using probabilistic parameters and SVM classification", *BMC Bioinformatics*, Vol. 7, S15, 2006.
- [2] Yiming Cheng, Robert M. Miura and Bin Tian *Prediction of mRNA polyadenylation sites by support vector machine*, *Bioinformatics*, Vol. 22 no. 19, pp 2320-2325, 2006.
- [3] Robertas Damaevicius *Splice Site Recognition in DNA Sequences Using K-mer Frequency Based Mapping for Support Vector Machine with Power Series Kernel*, International Conference on Complex, Intelligent and Software Intensive Systems, pp 687-692, 2008.
- [4] Gideon Dror, Rotem Sorek and Ron Shamir *Accurate identification of alternatively spliced exons using support vector machine*, *Bioinformatics*, Vol. 21 no. 7, pp 897-901, 2005.
- [5] Fickett, J. W. *Finding genes by computer: the state of the art*, *Trends Genet*, Vol. 12 no.8 pp 316-320, 1996.
- [6] T. R. Golub and D. K. Slonim and P. Tamayo and C. Huard and M. Gaasenbeek and J. P. Mesirov and H. Coller and M. L. Loh and J. R. Downing and M. A. Caligiuri and C. D. Bloomfield, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, *Science*, vol. 286, pp. 531-537, 1999.
- [7] Jones, D. and Watkins, C. *Comparing kernels using synthetic dna and genomic data..* Technical report, University of London, UK, 2000.
- [8] Liew, A. W.-C., Wu, Y., and Yan, H. *Selection of statistical features based on mutual information for classification of human coding and non-coding dna sequences*. *Bioinformatics technologies*, In ICPR04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR04) Volume 3, pages 766-769, Washington, DC, USA, 2004.
- [9] Platt J., "Fast Training of support vector machine using sequential minimal optimization. In A.S.B. Scholkopf, C. Burges, editor, *Advances in Kernel Methods: support vector machine* . MIT Press, Cambridge, MA 1998.
- [10] Pritish Varadwaj, Neetesh Purohit and Bhumika Arora., "Detection of Splice Sites Using Support Vector Machine . In *Contemporary Computing, Second International Conference, Proceedings*. Vol. 40, pp. 493-502, 2009.
- [11] Jing Xia, Doina Caragea and Susan Brown *Exploring Alternative Splicing Features Using Support Vector Machines*, Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine, pp 231-238, 2008.
- [12] Jing Xia, Doina Caragea and Susan Brown., "Exploring Alternative Splicing Features Using Support Vector Machines. In *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*. pp. 231-238, 2008.
- [13] Xiang H-F. Zhang, Katherine A. Heller, Ilana Hefter, Christina S. Leslie and Lawrence A. Chasin *Sequence Information for the Splicing of Human Pre-mRNA Identified by Support Vector Machine Classification*, *Genome Research*, Vol.13, pp. 2637-2650, 2003.
- [14] Xiang H-F. Zhang, K.A. Heller, I. Hefter, Ch. S. Leslie and Lawrence A. Chasin, *Sequence Information for the Splicing of Human Pre-mRNA Identified by Support Vector Machine Classification*, *Genome Research*, Vol. 13. pp 2637-2650, 2003.

A Study on Acupressure Points Online Database

Z. Wang and T. Santos

Department of Math/CS, Virginia Wesleyan College, Norfolk, VA, USA

Abstract: *Acupuncture points or acupressure points have been popularly used for thousands years for Chinese to cure various illness or pains. The method comes with less or no side effects, comparing the western medicine chemicals. With the new computer information technologies, the ancient Chinese healing tool could be very convenient for us to take care of ourselves at home or work. In this paper, we present our online database system for locating the acupressure points. A case study shows the details of the design and structure of the online system.*

Keywords: Acupressure points, databases, information systems, PHP/MySQL

1 Introduction

The vast applications of information systems have been developed very fast in the past decades. Typically, common information systems consist of people, procedures, data, software, and hardware that are integrated together in order to serve the objectives. Specifically computer-based information systems are corresponding networks of hardware/software that people and organizations use to collect, filter, process, create, and distribute. As computers grew in speed and capability, a number of general-purpose database systems emerged. Databases are designed to offer an organized mechanism for storing, managing, and fetching information in an efficient manner. In many cases, PHP and MySQL are used widely to create online databases that help us in need [1].

PHP is a general-purpose scripting language that is especially designed for Web development and can be embedded into HTML. It supports most of modern databases such as Informix, Oracle, and Sybase. It is open source software, meaning PHP is free to download and use. It can be used for both command-line scripting and client-side GUI applications. With PHP, embed dynamic Web design and programming become easy to handle. There is also unlimited control over the web server when using PHP. Whether you need to modify HTML on the fly, process a credit card, add user details to a database, or fetch information from a third-party website, you can do it all from within the same PHP files which the HTML itself is also located.

With more than ten million installations, MySQL database is one of the world's most popular database management systems for dynamic Web applications. It was developed in the mid 1990s and is now becoming a mature technology that powers wide ranges of Internet sites. MySQL is a popularly used not just because of its open source and free to use, but also its excellent performance, high reliability, and ease of use. Furthermore, it can even run on the most basic of hardware, and hardly puts a dent in system resources. So MySQL is highly scalable, meaning a

website using MySQL has the potential to grow [3]. In fact, in a comparison of several databases by *eWeek*, MySQL and Oracle tied for both best performance and for greatest scalability.

With PHP/MySQL, we construct an online information system that allows the user to select an illness and/or uncomfortable body part, and then present all the related acupressure points (Figure 1) that cure the particular symptom or illness. Beside these points are links to the location of the point with an image as well as directions on how to massage the specific point. Also, when an administrator logs in with the correct password and username, the administrator has the option of adding a record to the database.

Acupressure and acupuncture share the same active points (also called trigger points). Over 5,000 years ago, the ancient Chinese developed this system of active points stimulation. These active points are located on imaginary lines called meridians. Accordingly, the points are referred to by the meridian they are located on and consecutive number of point on that meridian [3].

In the next section, we will present a case study that details the system design and structure, including the files, tables, and examples of usages. And then conclusions follow.

2 Case Study

The URL of the online system is <http://zwang.vwc.edu/~tasantos>.

There are following nine PHP files to interact with the MySQL database.

- **directions.php** – lists all the directions on how to massage the point of concern.
- **header.php** – creates a banner and menu options on every page it is posted on.
- **index.php** – the homepage that contains project objectives and links to resources.
- **insert.php** – text boxes for the administrator to input another record.
- **login.php** – text boxes for administrator username and password.
- **login1.php** – lets administrator know if login was successful.
- **output.php** – adding new acupressure points to the points database. This file lets the administrator know if adding the record was successful. The INSERT INTO method is used

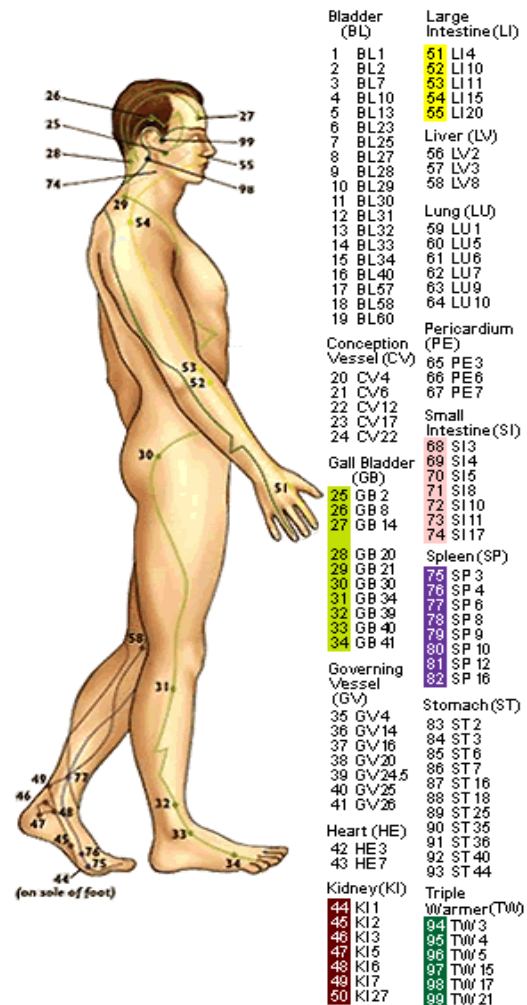


Figure 1 Acupressure points

to obtain the information to create additional database entries. It is only available for the administrator to perform operations, and the code is as follows.

```
<?php
    $server = 'localhost';
    $user = '*****';
    $pass = '*****';
    $mydb = '*****';
    $connect = mysql_connect($server, $user, $pass);
    $table_name = 'points';
    Print "Table $table_name Data<br>";
    $query = "INSERT INTO $table_name VALUES '0', '$name', '$chinesename',
'$linkurl', '$headache', '$hangover', '$sorethroat', '$heartburn',
'$weightloss', '$depression', '$insomnia', '$memory_and_concentration',
'$hiccoughs', '$high_blood_pressure)";

    Print "The Query is <i>$query</i><br>";
    mysql_select_db($mydb);
    print '<br><font size="4" color="blue">';
    if (mysql_query($query, $connect))
    { print "Insert into $mydb was successful!</font>"; }
    else
    { print "Insert into $mydb failed!</font>"; }
mysql_close($connect);
```

- **search1.php** – this form selects the symptom(s) and searches for the corresponding points. See sample of search1.php code below:

```
<?php include 'header.php'; ?>
<font face = papyrus>
<p>
<FORM ACTION=test.php METHOD=post>
<?php
    $menu = array('Headache', 'Hangover', 'Sore Throat', 'Heartburn',
'Weightloss', 'Depression', 'Insomnia', 'Memory & Concentration',
'Hiccoughs', 'High Blood Pressure');
    PRINT '<b>Please select your symptom(s):</b> <BR>';
    for($i=0; $i < count ($menu); $i++)
    { echo "<INPUT type=checkbox name=symptom[] value=$i> menu[$i]";
      echo "<BR>";
    }
?>
<p>
<INPUT type=submit value="Submit">
<INPUT type=reset value="Reset"></font></FORM></BODY>
```

- **test.php** – the action code for search1.php; outputs the points, including Chinese name, and hyperlink of its location.

When this code is submitted, all of the tables that are present in my database will show. Figure 2 shows the structure and content of the *points* table.


```

mysql> describe points;
+-----+-----+-----+-----+-----+-----+
| Field          | Type          | Null | Key | Default | Extra          |
+-----+-----+-----+-----+-----+-----+
| id             | int(11)       | NO   | PRI | NULL    | auto_increment |
| name           | varchar(8)    | YES  |     | NULL    |                |
| chinesename    | varchar(20)   | YES  |     | NULL    |                |
| linkurl        | varchar(100)  | YES  |     | NULL    |                |
| headache       | int(11)       | YES  |     | NULL    |                |
| hangover       | int(11)       | YES  |     | NULL    |                |
| sorethroat     | int(11)       | YES  |     | NULL    |                |
| heartburn      | int(11)       | YES  |     | NULL    |                |
| weightloss     | int(11)       | YES  |     | NULL    |                |
| depression     | int(11)       | YES  |     | NULL    |                |
| insomnia       | int(11)       | YES  |     | NULL    |                |
| memory_and_concentration | int(11)       | YES  |     | NULL    |                |
| hiccoughs      | int(11)       | YES  |     | NULL    |                |
| high_blood_pressure | int(11)       | YES  |     | NULL    |                |
+-----+-----+-----+-----+-----+-----+
14 rows in set (0.00 sec)

mysql>

```

Figure 2 The structure of Points table

When this code is typed in, a display is outputted of all the table's fields and their formats. Let's go through the design of the website and what the site has to offer to those who have a headache. When the user clicks on the link to go to the Acupressure Points System, the following page would then present itself.

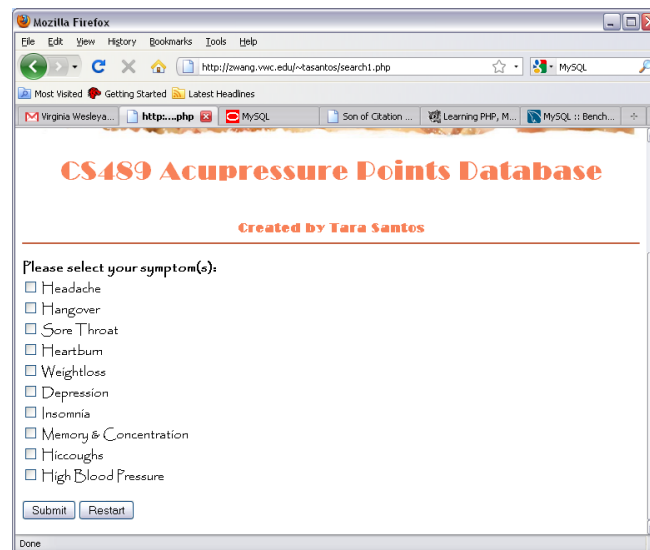


Figure 3 The layout of the online acupressure points database system

The Figure 3 page is created in the *search1.php* file which simply lists the symptoms currently listed in the database. The symptoms the user may select include headache, hangover, sore throat, heartburn, weight loss, depression, insomnia, improve memory and concentration, hiccoughs, and high blood pressure. Figure 4 shows the MySQL data in which the table information is stored.

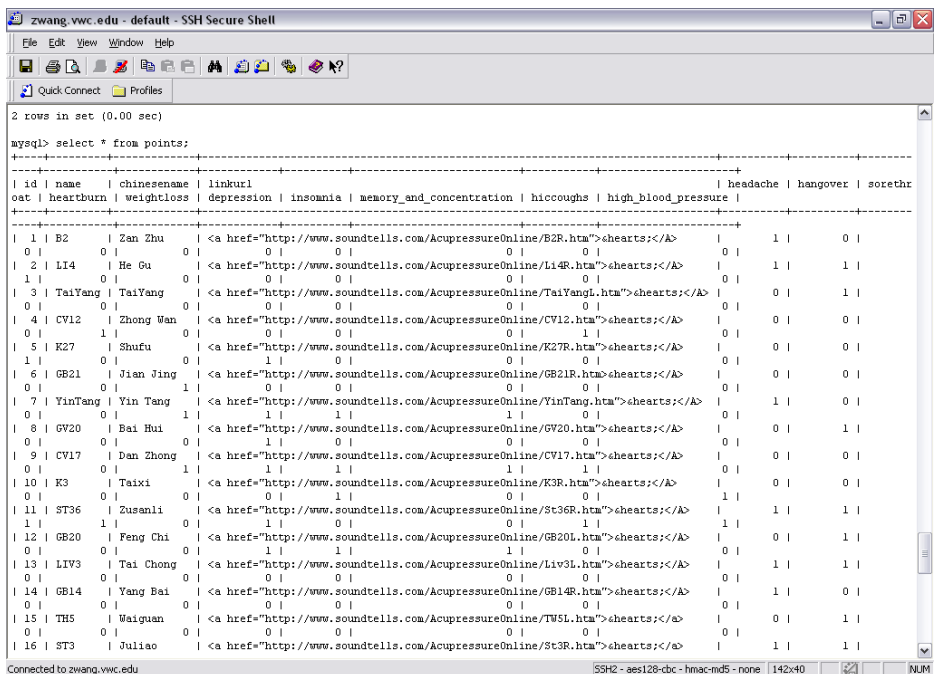


Figure 4 The content of Points table

If the user chooses headache as his or her symptom. The results would be shown in Figure 5.

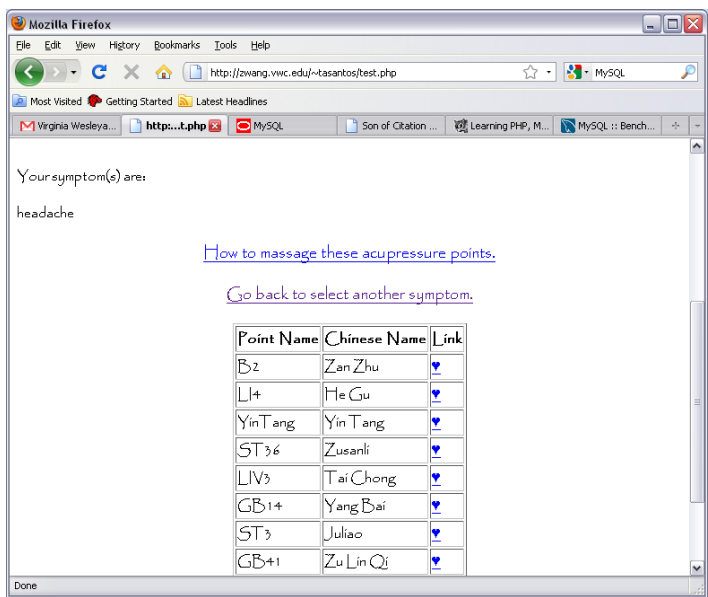


Figure 5 The display of output results

The page then outputs what the user selected as his or her symptom and lists all the related points to cure the headache. When the user clicks on the linking (a blue heart image), a Web page will appear that shows the points location, as Figure 6. This is the location of the first point B2 in Figure 5.

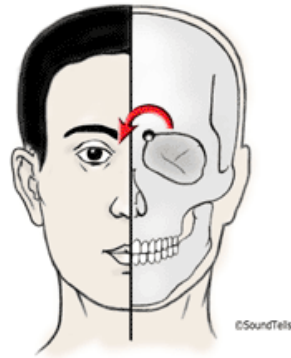


Figure 6 The exact acupressure point location

3 Conclusions

The paper presents the design and structure of online information system that is an easy-to-use and efficient way to help users to cure an individual's ailments naturally. The system was implemented by using a server side and open source scripting language and database PHP/MySQL. The results of system consist of several different acupressure points along with a linking Web page of the points' location, as well as detailed descriptions of how to massage these points. By using the system, the user can find ways to cure or at least relieve his/her specific symptom.

4 References

- [1] Robin Nixon. "Learning PHP, MySQL, & JavaScript". O'Reilly Media, 2009.
- [2] J. Ahigian and Z. Wang. "A low-cost online inventory database system". Proceedings of the 2009 International Conference on E-Learning, E-Business, Enterprise Information Systems, & E-Government (EEE'09), 276-280, CSREA Press, 2009.
- [3] Aaron Stein. "Acupressure Guide: Alleviate Headaches, Neck and Joint Pain, Anxiety Attacks, and Other Ailments". 2nd Ed. SoundTells, 2009.

Human Identification via Neural Network

By: Parviz Eshaghi
Tehran- Iran

Abstract

This paper presents a novel approach of iris verification based on Learning Vector Quantization Neural Network. The features used in this approach are based on the differences between the lines, rakes, and vessels of each iris considered as being non identical with any other one in the world. And for extracting these features, equipments like edge detection and discrete cosine transform (DCT) are used. The recognition obtained is 98% in small size database given via Learning Vector Quantization Neural Network.

Key words: Canny edge detection, discrete cosine transform, Learning Vector Quantization

1. Introduction

There are variable ways of human verification through out the world, as it is of great importance for all organizations, and different centers. Nowadays, the most important ways of human verification are recognition via DNA, face, fingerprint, signature, speech, and iris.

Among all, one of the recent, reliable, and technological methods is iris recognition which is practiced by some organizations today, and its wide usage in the future is of no doubt. Iris is a non identical organism made of colorful muscles including robots with shaped lines. These lines are the main causes of making every one's iris non identical. Even the irises of a pair of eyes of one person are completely different from one another. Even in the case of identical twins irises are completely different. Each iris is specialized by very narrow lines, rakes, and vessels in different people.

The precision of identification via iris is increased by using more and more details. It has been proven that iris patterns are never changed nearly from the time the child is one year old through out all his life.

Over the past few years there has been considerable interest in the development of neural network based pattern recognition systems, because of their ability to classify data. The kind of neural network practiced by the researcher is the Learning Vector Quantization which is a competitive network functional in the field of classification of the patterns.

The iris images prepared as the database is in the form of PNG (portable network graphics)

pattern, meanwhile they must be preprocessed through which the boundary of the iris is recognized and their features are extracted. For doing so, edge detection is done by the usage of Canny approach. For more diverse and feature extraction of iris images DCT transform is practiced.

2. Feature Extraction

For increasing the precision of our verification of iris system we should extract the features so that they contain the main items of the images for comparison and identification. The extracted features should be in a way that cause the least of flaw in the output of the system and in the ideal condition the output flaw of the system should be zero. The useful features which should be extracted are obtained through edge detection in the first step and the in next step we use DCT transform.

2.1 Edge Detection

The first step locates the iris outer boundary, i.e. border between the iris and the sclera. This is done by performing edge detection on the gray scale iris image. In this work, the edges of the irises are detected using the "Canny method" which finds edges by finding local maxima of the gradient. The gradient is calculated using the derivative of a Gaussian filter. The method uses two thresholds, to detect strong and weak edges, and includes the weak edges in the output only if they are connected to strong edges. This method is robust to additive noise, and able to detect "true" weak edges. Figures 1 and 2 are the original and edge images, respectively.

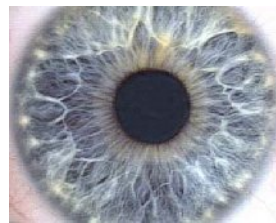


Figure 1: image of a sample iris



Figure 2: edges of a sample iris

Although certain literature has considered the detection of ideal step edges, the edges obtained from natural images are usually not at all ideal step

edges. Instead they are normally affected by one or several of these effects: focal blur caused by a finite depth-of-field and finite point spread function, penumbral blur caused by shadows created by light sources of non-zero radius, shading at a smooth object edge, and local peculiarities or inter reflections in the vicinity of object edges.

Despite the following model does not capture the full variability of real-life edges, the error function (erf) has been used by a number of researchers as the simplest extension of the ideal step edge model for modeling the effects of edge blur in practical applications. Thus, a one-dimensional image (f) which has exactly one edge placed at $x = 0$ may be modeled as:

$$f(x) = \frac{I_r - I_l}{2} \left(\operatorname{erf} \left(\frac{x}{\sqrt{2}\sigma} \right) + 1 \right) + I_l \quad (1)$$

At the left side of the edge, the intensity is $I_l = \lim_{x \rightarrow -\infty} f(x)$, and right of the edge it is $I_r = \lim_{x \rightarrow \infty} f(x)$. The scale parameter σ is called the blur scale of the edge.

2.1.1 Canny Method

The Canny edge detection algorithm is known to many as the optimal edge detector. Canny's intentions were to enhance the many edge detectors already out at the time he started his work. He was very successful in achieving his goal and his ideas and methods can be found in his paper, "A Computational Approach to Edge Detection". In his paper, he followed a list of criteria to improve current methods of edge detection. The first and most obvious is low error rate. It is important that edges existing in images should not be missed and that there be NO responses to non-edges. The second criterion is that the edge points be well localized. In other words, the distance between the edge pixels as found by the detector and the actual edge is to be at a minimum. A third criterion is to have only one response to a single edge. This was implemented because the first 2 were not substantial enough to completely eliminate the possibility of multiple responses to an edge.

The Canny operator works in a multi-stage process. First of all the image is smoothed by Gaussian convolution. Then a simple 2-D first derivative operator (somewhat like the Roberts Cross) is applied to the smoothed image to highlight regions of the image with important spatial derivatives. Edges give rise to ridges in the gradient magnitude image. The algorithm then tracks along the top of these ridges and sets to zero all pixels that are not actually on the ridge top so as to give a thin line in the output, a process known as non-maximal suppression. The tracking process exhibits hysteresis controlled by two thresholds: T1 and T2, with $T1 > T2$. Tracking can only begin at a point on a ridge

higher than T1. Tracking then continues in both directions out from that point until the height of the ridge falls below T2. This hysteresis helps to ensure that noisy edges are not broken up into multiple edge fragments.

An edge in an image may point in a variety of directions, so the Canny algorithm uses four filters to detect horizontal, vertical and diagonal edges in the blurred image. The edge detection operator (Roberts, Prewitt, Sobel for example) returns a value for the first derivative in the horizontal direction (G_y) and the vertical direction (G_x). From this the edge gradient and direction can be determined:

$$G = \sqrt{G_x^2 + G_y^2} \quad (2)$$

$$\theta = \arctan \left(\frac{G_y}{G_x} \right) \quad (3)$$

The edge direction angle (theta) is rounded to one of four angles representing vertical, horizontal and the two diagonals (0, 45, 90 and 135 degrees for example).

2.2 Discrete Cosine Transform

Like any Fourier-related transform, discrete cosine transforms (DCTs) express a function or a signal in terms of a sum of sinusoids with different frequencies and amplitudes. Like the discrete Fourier transform (DFT), a DCT operates on a function at a finite number of discrete data points. The obvious distinction between a DCT and a DFT is that the former uses only cosine functions, while the latter uses both cosines and sinusoids (in the form of complex exponentials). However, this visible difference is merely a consequence of a deeper distinction: a DCT implies different boundary conditions than the DFT or other related transforms.

The Fourier-related transforms that operate on a function over a finite domain, such as the DFT or DCT or a Fourier series, can be thought of as implicitly defining an extension of that function outside the domain. That is, once you write a function $f(x)$ as a sum of sinusoids, you can evaluate that sum at any x , even for x where the original $f(x)$ was not specified. The DFT, like the Fourier series, implies a periodic extension of the original function. A DCT, like a cosine transform, implies an even extension of the original function.

A discrete cosine transform (DCT) expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. DCTs are important to numerous applications in science and engineering, from lossy compression of audio and images (where small high-frequency components can be discarded), to spectral methods for the numerical solution of partial differential equations. The use of cosine rather than sine functions is critical in these applications:

for compression, it turns out that cosine functions are much more efficient (as explained below, fewer are needed to approximate a typical signal), whereas for differential equations the cosines express a particular choice of boundary conditions.

In particular, a DCT is a Fourier-related transform similar to the discrete Fourier transform (DFT), but using only real numbers. DCTs are equivalent to DFTs of roughly twice the length, operating on real data with even symmetry (since the Fourier transform of a real and even function is real and even), where in some variants the input and output data are shifted by half a sample. There are eight standard DCT variants, of which four are common.

The most common variant of discrete cosine transform is the type-II DCT, which is often called simply "the DCT"; its inverse, the type-III DCT, is correspondingly often called simply "the inverse DCT" or "the IDCT". Two related transforms are the discrete sine transform (DST), which is equivalent to a DFT of real and odd functions, and the modified discrete cosine transform (MDCT), which is based on a DCT of overlapping data.

The DCT, and in particular the DCT-II, is often used in signal and image processing, especially for lossy data compression, because it has a strong "energy compaction" property. Most of the signal information tends to be concentrated in a few low-frequency components of the DCT.

$$\text{DCT-II}$$

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right] \quad k = 0, \dots, N-1. \quad (4)$$

This transform is exactly equivalent (up to an overall scale factor of 2) to a DFT of $4N$ real inputs of even symmetry where the even-indexed elements are zero. That is, it is half of the DFT of the $4N$ inputs y_n , where $y_{2n} = 0$, $y_{2n+1} = x_n$ for $0 \leq n < N$, and $y_{4N-n} = x_n$ for $0 < n < 2N$.

The DCT-II implies the boundary conditions: x_n is even around $n=-1/2$ and even around $n=N-1/2$; X_k is even around $k=0$ and odd around $k=N$.

3. Neural Network

In this work one Neural Network structure is used, which is Learning Vector Quantization Neural Network. A brief overview of this network is given below.

3.1 Learning Vector Quantization

Learning Vector Quantization (LVQ) is a supervised version of vector quantization, similar to Self organizing Maps (SOM) based on work of Linde et

al, Gray and Kohonen. It can be applied to pattern recognition, multi-class classification and data compression tasks, e.g. speech recognition, image processing or customer classification. As supervised method, LVQ uses known target output classifications for each input pattern of the form.

LVQ algorithms do not approximate density functions of class samples like Vector Quantization or Probabilistic Neural Networks do, but directly define class boundaries based on prototypes, a nearest-neighbor rule and a winner-takes-it-all paradigm. The main idea is to cover the input space of samples with 'codebook vectors' (CVs), each representing a region labeled with a class. A CV can be seen as a prototype of a class member, localized in the centre of a class or decision region in the input space. A class can be represented by an arbitrarily number of CVs, but one CV represents one class only.

In terms of neural networks a LVQ is a feed forward net with one hidden layer of neurons, fully connected with the input layer. A CV can be seen as a hidden neuron ('Kohonen neuron') or a weight vector of the weights between all input neurons and the regarded Kohonen neuron respectively.

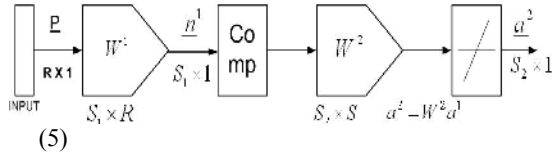
Learning means modifying the weights in accordance with adapting rules and, therefore, changing the position of a CV in the input space. Since class boundaries are built piecewise-linearly as segments of the mid-planes between CVs of neighboring classes, the class boundaries are adjusted during the learning process. The tessellation induced by the set of CVs is optimal if all data within one cell indeed belong to the same class. Classification after learning is based on a presented sample's vicinity to the CVs: the classifier assigns the same class label to all samples that fall into the same tessellation – the label of the cell's prototype (the CV nearest to the sample).

The core of the heuristics is based on a distance function – usually the Euclidean distance is used – for comparison between an input vector and the class representatives. The distance expresses the degree of similarity between presented input vector and CVs. Small distance corresponds with a high degree of similarity and a higher probability for the presented vector to be a member of the class represented by the nearest CV. Therefore, the definition of class boundaries by LVQ is strongly dependent on the distance function, the start positions of CVs, their adjustment rules and the pre-selection of distinctive input features.

Briefly explaining, this network has two layers: a layer of input neurons, and a layer of output neurons. The network is given by prototypes $W=(w(i), \dots, w(n))$. It changes the weights of the network in order to classify the data correctly. For each data point, the prototype (neuron) that is closest to it is determined (called the winner neuron). The weights of the connections to this neuron are then

adapted, i.e. made closer if it correctly classifies the data point or made less similar if it incorrectly classifies it.

3.1.1 Learning Algorithm



Learning Vector Quantization (LVQ) structure

The number of neurons in the first layer (s1) should be equal at least to the number of neurons in the second layer, i.e. $S_1 \geq S_2$

Generally, the neurons in the first layer are more than the second layer. The behavior of the LVQ Neural Network is expressed by the equation below:

$$n_i^1 = -\|W_i^1 - \underline{P}^T\| \quad (6)$$

The pure input of the 1st neuron in the first layer

And also it is written in Vector form as below:

$$n^1 = \begin{bmatrix} \|W_1^1 - \underline{P}^T\| \\ \vdots \\ \|W_{s1}^1 - \underline{P}^T\| \end{bmatrix} \quad (7)$$

The output vector of the first layer is:

$$\underline{a}^1 = \text{comp}(n^1) \quad (8)$$

Therefore, the vector which has the nearest weight to the input vector is equal to 1, and the rest of the neurons have the zero output. The function of the second layer is to compose the subclasses of one class and to create just one class. W^2 matrix in each column has the element 1 and the other elements are zero.

$$W_{ji}^2 = 1 \quad \text{The subclass of } i \text{ belongs to } j \text{ class}$$

Kohonen learning rule is used to organize the parameters of the LVQ NN layer in the form below:

$$W_{i*}^1(k+1) = W_{i*}^1(k) + a(P^T(k+1) - W_{i*}^1(k)), \quad (9)$$

$$\text{If: } a_{j*}^2 = t_{j*}(k+1) = 1 \quad (10)$$

4. Simulation Results

Practically we have done this work for a prepared database for 10 people in which we gave a class for the iris image of the left and right eye of every one and in the long run we obtained 10 classes. It means that in the second layer (s2) of the LVQ NN the number of the neurons is 10, while we put in the first layer (s1) 30 neurons. For each one of these 10 persons we took one image of the left eye iris and another from the right eye iris, and implemented these 20 taken images to the input of neural network after feature extraction by Canny edge detection approach and DCT transform. After learning network for these 20 input images, and testing by the other images from other left and right eyes irises other than the very 20 images we had, finally the true recognition results of our test came to an average of 98%. In this test we also used different noised images.

5. Conclusion

In this paper, a novel technique is proposed for iris verification. The classification is performed using LVQ Neural Network. The neural network based approach is found to be a promising one for iris recognition.

References

- [1] Canny, J., *A Computational Approach to Edge Detection*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, 1986.
- [2] Frigo, Matteo, Steven G. Johnson: FFTW, <http://www.fftw.org/>. A free (GPL) C library that can compute fast DCTs (types I-IV) in one or more dimensions, of arbitrary size.
- [3] Frigo, Matteo, Steven G. Johnson, "The Design and Implementation of FFTW3," Proceedings of the IEEE 93 (2), 216–231 2005.
- [4] "Bibliography on the Self-Organizing Map (SOM) and Learning Vector Quantization (LVQ)", Neural Networks Research Centre, Helsinki University of Technology, 2002.

Telomerase Gene Prediction Using Support Vector Machines

David Luper¹ and Spandana Makeneni²

¹Computer Science Department, UGA, Athens, GA, USA

²Institute of Bioinformatics, Complex Carbohydrate Research Center, UGA, Athens, GA, USA

Abstract - Telomerase genes have been said to be of great importance in various aspects of biology. Currently their composition and purpose is a topic of much research. Finding and validating telomerase genes in different species is of great importance and is also a difficult task that consumes many resources. In this research a method for isolating potential telomerase gene regions within a genome is discussed. A Support Vector Machine will be used to differentiate regions of DNA containing telomerase genes from those that do not. The Support Vector Machine will be trained on identified telomerase genes from related species, and then it will be used to classify sequences encompassing an entire chromosome of a different species as either potential telomerase gene regions or non-telomerase regions. Ultimately, a fast algorithm is presented that can act as an initial filter to remove large portions of a genome, allowing more time intensive routines to better target optimal regions of a genome.

Keywords: Data Mining, Computational Biology, Machine Learning

1 Introduction

Telomerase (Fig. 1), also called telomere terminal transferase, is an enzyme made of protein and RNA subunits that dictates the synthesis of telomere terminal repeats. This mechanism is required for the maintenance of chromosome termini, as the structure and integrity of telomeres are essential for genome stability. Telomere deregulation can lead to cell death, cell senescence, or abnormal cell proliferation. It has been identified that telomerase plays very important roles in aging and cancer. Telomerase activity is detected during development and has a very low, almost undetectable, activity in somatic (body) cells. These somatic cells age as a result of telomerase inactivity. So, if telomerase is activated in a cell, the cell will continue to grow and divide leading to exciting possibilities. In the past several years of research, it has been found that cancer cells are immortal and divide uncontrollably. Such immortal cancer cells have 10-20 times more active telomerase than in normal body cells. Reducing this activity could eventually lead to the death of those cells.

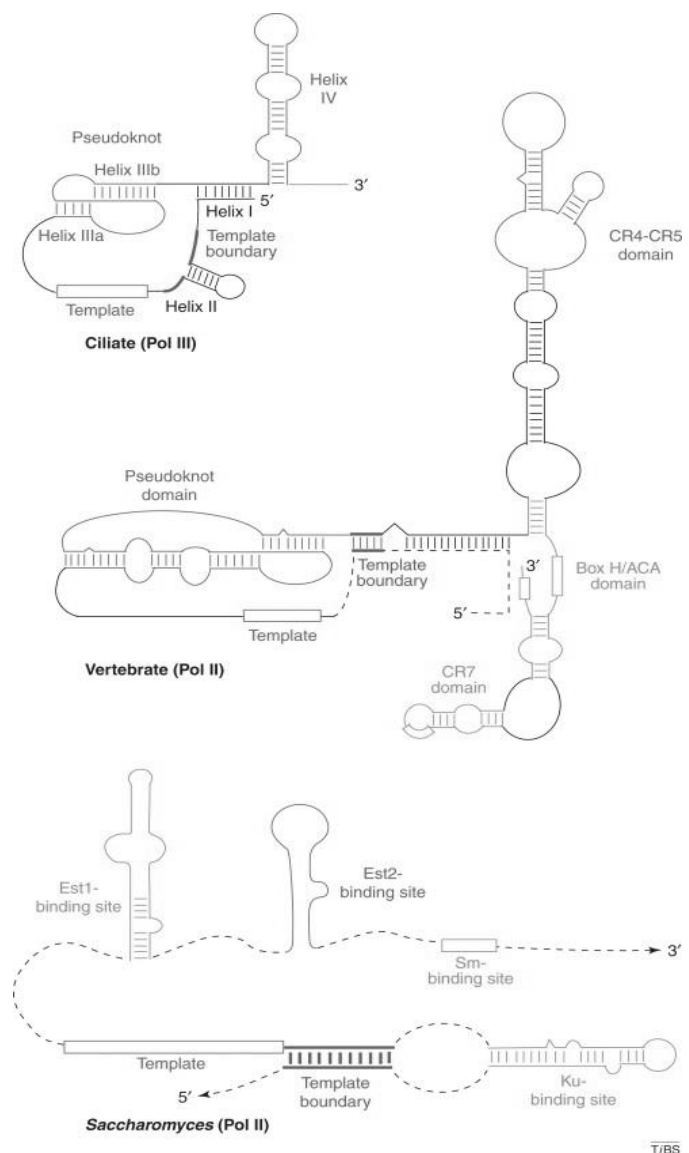


Figure 1. Detailed telomerase RNA secondary structure for humans and yeast

This could be a great therapy especially in the early stages of cancer. In the later stages, inhibition or absence of telomerase may result in cell crisis in cancer cells and tumor regression in cancer patients. Research on telomerase continues to be a very exciting field with potential for discovering many more facts about what might help fight cancer and the aging process.

In this research both yeast telomerase genes and telomerase genes of vertebrates are used to teach a supervised machine learning algorithm what telomerase genes look like. Once the algorithm builds a model to represent these genes, it can look through entire genomes to narrow down the search for new telomerase genes in species they have not been identified. The type of supervised machine learning algorithm used for this research is a Support Vector Machine (SVM). Support Vector Machines have been used for instance classification in complex biological domains with great effectiveness [1] [3]. SVMs have been shown to obtain better results over a wide variety of problems in comparison with other algorithms used in supervised machine learning. This is because they generalize better due to the nature of how they learn.

This paper will show how an SVM can be used to narrow the search for new telomerase genes. It will be laid out in the following manner. First, supervised machine learning and support vector machines will be briefly discussed. After this, the methodology section will outline the steps used in this research to isolate regions of chromosomes labeled as having potential to house a telomerase gene. Next, the experiments from this work will be presented along with results. Finally, future enhancements to the methodology will be discussed before concluding remarks.

2 Machine Learning

Support Vector Machines are a type of supervised machine learning algorithm. Supervised machine learning algorithms are used to approximate non-linear functions for instance classification. These algorithms build models from a group of data instances called training data, and use these models to classify new instances where the class is not known. Each data instance in the training data consists of n features, from an n dimensional feature space S , and a label that tells the algorithm which class the data instance belongs to. These instances describe locations for each class in S , and they are treated as a representation of a non-linear function $f(i)$ where i is an input vector of features. Once the training data is assembled a model is constructed. While constructing the model a portion of the training data is placed into another data set called the validation data. The validation data is withheld from the learner while training it and used to test how effective the model generalizes to instances outside of the training data. There are different schemes for segmenting and utilizing the validation data, this research uses a method called n fold cross validation. N fold cross validation divides

the training data into n data sets and builds $n - 1$ models where one of the n data sets is used as the validation data. The $n - 1$ models are combined to produce a single model that can be used for classification. This technique helps the model generalize better when there are relatively few instances in the training data. Once the final model is built, it can be tested with a separate group of disjoint data instances called production data. These instances are labeled as belonging to a particular class, but this label is withheld from the algorithm during classification to see how accurately the model approximates the targeted non-linear function on data it has never seen.

SVMs have been shown to obtain better results over a wide variety of problems in comparison with other algorithms used in supervised machine learning. This is because they generalize better, due to the nature of how they learn. SVMs learn concepts by separating data distributions into classes of data and treating them as two generalized sets of vectors in a feature space. The SVM will find a separating hyperplane between these two datasets (or concepts) which is the maximum distance from either of the two (Fig. 2). Other machine learning algorithms can find hyperplanes that separate datasets but the power of the SVM comes from the fact that the hyper plane found by the SVM is the one with the greatest distance between either of the two classes. The SVM finds support vectors, which are data instances from either class that are the closest to the opposite class. Once these support vectors are found, geometric operations are applied to find the hyperplane that is equally distant from both sets of support vectors. Finding support vectors and computing the maximum marginal hyperplane is a standard quadratic programming problem [4]. This explanation assumes a linearly separable feature space because that is the easiest way to explain the concept. SVMs can be generalized to support nonlinear features spaces as well as more than two classes of data, but these topics are beyond the scope of this paper.

3 Methodology

Given feature sets representing data instances, SVMs learn concepts and identify instances as belonging to specific classes of data. This project uses SVMs for locating regions within chromosomes that have potential to contain telomerase genes. The classes of data instances in this project are $+$ and $-$ where $+$ is a segment of chromosome that potentially holds a telomerase gene and $-$ is any other region of chromosome. The training data used to construct the classification model for the SVM is used to tell the SVM what telomerase regions look like ($+$), and what they do not ($-$). Since each species has only one telomerase gene region, telomerase gene regions from related species are used to in the training data as $+$ instances. The $-$ instances in the training data were randomly sampled non-telomerase regions from the same group of species. Five times as many $-$ instances were included in the training data as available $+$ instances. The production data

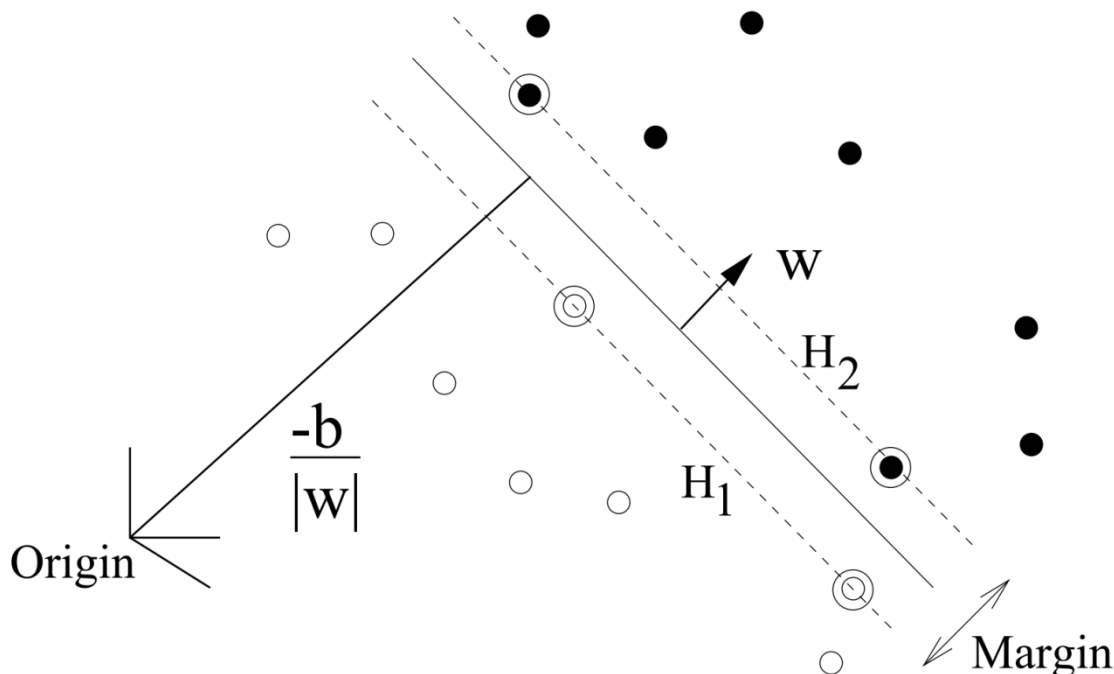


Figure 2. Illustration of what a maximum marginal hyperplane looks like between two set of data instances in a feature space. (Image taken from Christopher J.C. Burges [4])

used to test the correctness of the SVM was an entire chromosome from species X, where X was related to the species in the training data, and the telomerase gene region for X had already been positively identified. This production data allows the correctness of the SVM to be effectively measured by whether the SVM successfully classifies the telomerase region from species X and how much of the rest of the chromosome from species X gets correctly labeled non-telomerase. The data for this research was gathered from two sources. The telomerase gene regions were obtained from <http://telomerase.asu.edu/sequences.html>, and the chromosome in which those regions reside were taken from <ftp://ftp.ncbi.nih.gov/>.

During the construction of data sets for this research it was imperative that the telomerase gene region from the species being used in the production data not be included in the training data. The inclusion of this telomerase region would skew the results for the experiment as the SVM would be trained specifically on one of the instances it is also being evaluated for correctness on. This kind of scenario leads to over fitting on the training data and as a result an SVM generalizes poorly.

To obtain data instances for submission to the SVM, features are computed from segments of DNA. For the + examples used in the training data, the segments of DNA used were the telomerase genes from each of the species involved

in defining the + examples. For the - examples used in the training data, and the examples used in the production data the chromosomes were segmented into regions of length m using a sliding window over the chromosome. This sliding window was started at index $n = 0$ and between segments n was incremented by x . For this project x was set to 75 and m was set to the average length of the telomerase regions used as + instances in the training data. After the necessary chromosomes were segmented and assigned to the training and production data sets features for the segments could be calculated.

There are nine features used in this research to classify instances. These features were either taken from or inspired by Guo et al [5] and Schattner [6]. Schattner's work was of particular use to this research. In Schattner's paper the base composition of sequences are used to determine RNA gene regions. Schattner only uses statistical analysis of these regions to infer their class, but these features work very well for machine learning. The features used by Schattner are (G+C)%, (G-C)%, (A-T)%, and RO(AB). The features used in this research are the following:

Percentage A:

The percentage nucleotides in the DNA sequence that were A.

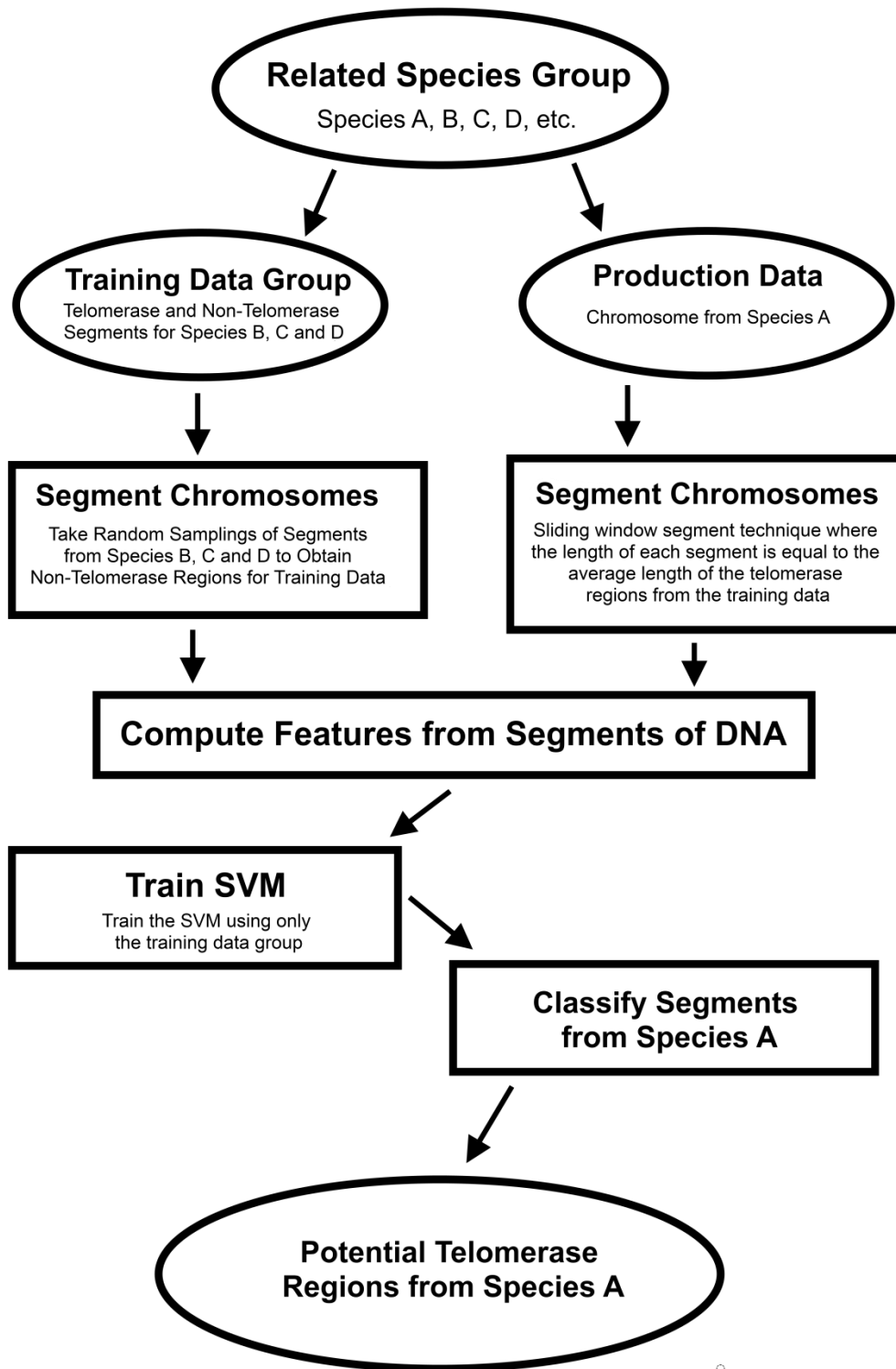


Figure 3. Methodology flow diagram to illustrate the process of obtaining potential telomerase gene regions.

Percentage T:

The percentage nucleotides in the DNA sequence that were T.

Percentage G:

The percentage nucleotides in the DNA sequence that were G.

Percentage C:

The percentage nucleotides in the DNA sequence that were C.

Percentage (X + Y):

The percentage of the nucleotides in the DNA sequence that were either X or Y summed, with this feature each possible combination of nucleotides were computed.

Percentage (X - Y):

The percentage of the nucleotides in the DNA sequence that were X subtracted from the percentage of nucleotides in the sequence that were Y, with this feature each possible combination of nucleotides were computed.

Percentage (X / Y):

The percentage of the nucleotides in the DNA sequence that were X divided by the percentage of nucleotides in the sequence that were Y, with this feature each possible combination of nucleotides were computed.

RO(XY):

The frequency count of XY (FREQ_XY) multiplied by the length of the sequence then divided by the percentage X times the percentage Y.

ex. $(\text{length} * \text{FREQ_XY}) / (\text{Percentage X} * \text{Percentage Y})$

Standard Deviation:

The standard deviation of the percentages of A, T, G, and C.

Once the features are computed for each of the DNA segments the SVM can be trained. Due to the small size of the training data, cross fold validation was used to help prevent over fitting. After the SVM was trained the production data was classified, and then overlapping segments of + classifications were merged together. This results in regions of DNA, of various lengths, that potentially house the telomerase gene. The number of nucleotides in the calculated

regions can be used against the total number of nucleotides in the entire chromosome to compute the percentage of the chromosome classified + or -.

4 Experiment and Results

The experiments for this research were run in two different groups (vertebrates and fungi). A flow diagram outlining the experimental procedure can be seen in Fig. 3. The *training data* for each group consisted of + instances of telomerase genes from as many related species as could be found. For each group three experiments were run. The experiments consisted of removing species X from the *training data* for use as the *production data*. After training the SVM, results were obtained from classification of the entire chromosome containing the telomerase gene region from species X. The results were defined by the recall and precision of the classification of gene regions in the chromosome. The recall was whether the SVM classified the telomerase region in species X correctly, and the precision was how much of the rest of the chromosome was classified correctly as non-telomerase. For this experiment species X had to meet two constraints. First, its telomerase gene region must be known, and second, the rest of the chromosome in which the telomerase region resided must have been sequenced. For vertebrates the three species experimented on were *Mus musculus*, *Rattus norvegicus* and *Equus caballus*, and for fungi the three species were *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae* and *Kluyveromyces lactis*. The results are shown in table 1. The results show the number of potential telomerase regions detected and the percentage of the chromosome those regions accounted for. The percentage of the chromosome the potential telomerase regions account for minus one depicts the amount of the chromosome that is excluded from being a potential telomerase region. This shows how far the SVM narrowed the search for the telomerase gene. In each of the experiments run, the SVM classified the actual telomerase gene correctly. This puts the recall at 100%. Since there is only one telomerase gene within a genome for any species, the percentage of the genome classified as potential telomerase regions can be seen as the false positive rate, within a very small statistical margin of error. The results show significant information gain. However, the results on the vertebrates are significantly better than the results on the fungi. Possible explanations for this could be that the groups of vertebrates used in the *training data* were more closely related. This could make their telomerase genes more alike and provide a better representation for the SVM. Another more likely explanation could be that the results on the vertebrates were better because the SVM had more data to learn from with the vertebrates. The number of known telomerase gene regions in the *training data* for the vertebrates was 22, but only 13 telomerase gene region examples were available for the fungi. A final explanation for the better results on the vertebrates could be that in the vertebrates the telomerase genes were simply more distinct from the rest of the chromosome than they were in the fungi.

Species	# of Regions Classified +	Percentage of Chromosome Classified +
<i>Schizosaccharomyces pombe</i>	680	0.25377804949519833
<i>Saccharomyces cerevisiae</i>	165	0.44209262916606207
<i>Kluyveromyces lactis</i>	196	0.30870646879168767
<i>Mus musculus</i>	960	0.011700037718866478
<i>Rattus norvegicus</i>	1536	0.010321241324534453
<i>Equus caballus</i>	777	0.024617685567403513

Table 1

5 Future Research

Future research should be invested in at least two areas for this work. First, the SVM used in this project utilized default settings in the WEKA machine learning software package (i.e. complexity parameter and a linear kernel). Different settings for these parameters such as an RBF kernel, or different numeric values for the complexity parameter, should be explored to see if the results for the experiment could be improved. Second, new features should be explored to see if they can better detect potential telomerase regions. One such feature could reflect base pairings within the sequence of DNA. Telomerase genes should have a unique and learnable base pairing signature that sets them apart from the rest of the chromosome (i.e. the way a telomerase region folds to create its secondary structure should be distinctive). Another feature that should be looked into would be to isolate the most commonly repeated *l-mer* in the + examples from the training data and provide the number of times the particular subsequence (either exactly or with some accommodation for mutation allowed) appears in the instance. A third feature would be to create a multiple alignment from the + instances in the training data to obtain a median string (or consensus string) used for computing global and local alignment scores for each instance. In telomerase genes from related species there should exist conserved regions, and thus telomerase genes could have a unique scoring signature against this median string.

6 Conclusion

The work presented in this paper provides substantial results showing an SVM can definitively narrow the search for telomerase genes within a genome. A methodology has been outlined that segments a chromosome into DNA sequences that are treated as data instances in a machine learning application. Features are computed from these DNA sequences and the feature vectors are classified by an SVM as either potential telomerase gene regions (+) of not (-). The results from the experiment show significant information gain, however they have potential to be improved through the exploration of new features and parameter refining in the SVM.

7 References

- [1] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michel Schummer, David Haussler, 2000, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, Vol. 10 (2000), 906 – 914
- [2] Simon Tong, Edward Chang, Support Vector Machine Active Learning for Image Retrieval, 2001, ACM International Conference Proceedings, Vol. 1, 107 – 118
- [3] Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik, 2002, Gene Selection for Cancer Classification Using Support Vector Machines, Machine Learning, 1 – 39
- [4] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, vol. 2, no. 2. pp. 121-167, 1998.
- [5] Feng-Biao Guo, Hong-Yu Ou, Chun-Ting Zhang, 2003, ZCURVE: A New System for Recognizing Protein-Coding Genes in Bacterial and Archaeal Genomes, Nucleic Acids Research, 2003, Vol. 31, No.6, 1780 – 1789
- [6] Peter Schattner, 2002, Searching for RNA Genes Using Base Composition Statistics, Nucleic Acids Research, 2002, Vol. 30, No. 9, 2076 – 2082
- [7] Tom M. Mitchell, [1997], Machine Learning, International Edition, MIT Press and The McGraw-Hill Companies, Inc.
- [8] Kim N.W. Clinical implications of telomerase in cancer (1997) European Journal of Cancer Part A, 33 (5), pp. 781-786.
- [9] Toren Finkel, Jan Vijg & Jerry W. Shay. Time, tumours and telomeres. Meeting on Cancer and Aging
- [10] Jiunn-Liang Chen, Carol W. Greider, Telomerase RNA structure and function: implications for dyskeratosis congenita, Trends in Biochemical Sciences, Volume 29, Issue 4, April 2004, Pages 183-192, ISSN 0968-0004, DOI:10.1016/j.tibs.2004.02.003.

SESSION

PROTEIN CLASSIFICATION + STRUCTURE PREDICTION, EVOLUTIONARY, AND COMPUTATIONAL MOLECULAR + STRUCTURAL BIOLOGY

Chair(s)

TBA

Solving Planted Motif Problem Using Modeling Method

S. Desarjau and R. Mukkamala

Computer Science Department, Old Dominion University, Norfolk, Virginia, USA

Abstract - In this paper we describe a new method for solving the Planted Motif Problem that has applications in computational biology. A number of algorithms to solve this problem have been proposed in the past. The largest problem reported solved in the literature is (21, 8). Using the new method we have solved much larger problems, up to a size of (48, 12). The new method is also much faster, and we compare its performance with the best performances reported in the literature.

Keywords: Elmers, heuristics, memory-constrained computing, modeling, motif, planted motif problem

1 Introduction

The Planted Motif Problem (PMP) may be abstractly defined as: "Given a set of n strings, each of length L , over an alphabet Σ , find a string M of length $l < L$ over Σ , such that there is at least one d -neighbor of M in each of the n strings, where a d -neighbor is a string of length l that differs from M in at most d positions and M is the *motif* for the given set of n strings."

PMP has applications in molecular biology. Identifying subtle signals in the transcription-factor binding sites of several genes is the primary application of PMP. Finding such regulatory patterns among DNA sequences aids the study of gene regulatory networks [1]. The alphabet of the Planted Motif Problem is usually $\{A, C, G, T\}$, corresponding to the four nucleotide bases that constitute DNA. In principle, however, the problem can be posed for strings over any finite alphabet.

The Planted Motif Problem can be posed for different values of ' l ' and ' d '. Larger values of ' l ' and ' d ' constitute larger problems, and typically take more time and/or memory to solve. The pair (l, d) is used to express the size of a given problem. Most researchers keep ' n ' and ' L ' fixed at 20 and 600, respectively [2].

Several methods currently exist to solve the PMP. These are: PMS1 [3]; Pattern Branching [4]; WINNOWER [2]; MITRA [5]; Random Projection [6]; Bit-based Multi-core [7]; ExVote [8], Stemming [9], PMSPrune [10], RISOTTO [11], and algorithms for solving the Extended Motif Problem (EMP) [1] [12].

We summarize some of these methods below. All these methods have limitations in the size of the problems that they solve and the running time that they require.

In contrast to these existing algorithms, we approach the motif-finding problem with a method of deriving the motif from clues present in the input strings. If there is a motif of length l present for a given set of input strings, the length l

substrings of the input exhibit certain simple properties. We identify these properties and use them to construct the motif. The method has been used to solve problem sizes much larger than those reported solved in the literature, in times much shorter than the times reported for smaller problems in the literature. The method also holds promise for problems of larger alphabet size than the DNA alphabet size of 4.

The rest of the paper is organized as follows. In section 2, we illustrate the problem and summarize previous work in the area. Section 3 describes 'modeling' and the discovered properties as a set of propositions. Section 4 presents a formal algorithm and analysis of the computational complexity. In Section 5 we provide an overview of the statistical properties of the factors involved in the computation. In section 6, we present the results that show the superiority of the proposed method in terms of ability to solve larger problems with smaller run times. We also state the notable shortcomings of the method. Finally, Section 7 summarizes our contributions and discusses future work.

2 Problem Illustration and Previous work

The Planted Motif Problem has been an area of active research for about the last twenty years, and a number of different approaches have been described.

Before we describe the previous work, let us look at an example to illustrate the problem. Consider a set of 3 DNA sequences, each of length 32. The problem is to find a string (motif) of length 6, for which a 2-neighbor exists in each of the 3 sequences. Here, $n=3$, $L=32$, $l=6$, $d=2$, and $\Sigma=\{A, C, G, T\}$. The size of the problem is $(6, 2)$.

```
0 GTCAGACAGATCGTGTTCATACGACGACTTC
1 CTATGACCAAGGGATTTCTAACCACGGCACTT
2 ATCAGTCCCAGGGTGTTCGCTCGACGTGTT
```

For this problem, there exists a motif – **GGCTCG**. The sub-sequences of length 6 that are in bold face are d -neighbors of the motif. The first substring **AGATCG** differs from the motif at the 1st and 3rd positions. The 2nd substring **GGCACT** differs from the motif at the 4th and 6th positions. The third substring **CGCTCG** differs at only the 1st position. So **GGCTCG** is one of the solutions. There could be more than one motif. But the problem statement calls for finding any one such string.

It is important to note that the motif itself may not literally occur in any of the input sequences. All that is required is that a d -neighbor of the motif occurs in each of the input sequences.

We now summarize previous efforts to solve this problem. WINNOWER [2] takes a *graph-based* approach to solve the problem. It finds cliques in a graph constructed by representing each length l substring of the input as a node. Thus the number of nodes will be $n*(L-1+1)$. WINNOWER is effective up to problem sizes of (18, 5) and requires substantial computational resources (both time and memory).

MITRA [5] is based on a *trie* (or prefix tree) *traversal* algorithm. A *mismatch tree* data structure is used, in which all the possible patterns are segregated into disjoint subsets, with each subset starting with a given prefix. It is effective up to problem sizes of (18, 6), taking 40 minutes and 650 MB of memory.

PatternBranching [4] starts with a random seed string and searches for the length l neighbors of this string in the input. It scores the neighbors with an appropriate scoring function and selects the best scoring neighbor.

The Random Projection Algorithm [6] finds motifs using random projections. For each length l substring of the input, a length k string is constructed as a subsequence of the original substring, sharing k random positions (i.e., it is a random projection). All the length l substrings are hashed using the length k string of any length l substring as the hash value. If a hashed group has at least a threshold number of substrings in it, then it is likely that the motif will have its length k projection equal to the length k projection of this group. The largest problem the method is reported to have solved is (18,6), in about one hour.

The PMS1 algorithm [3] is based on exhaustive enumeration. The Bit-based algorithm [7] is also based on exhaustive enumeration, and is a multicore (i.e. parallel processing) implementation, with modifications to address memory-sharing issues and enhance performance. The largest problem the method is reported to have solved is (21, 8), in 7.8 hours, using 16 CPU cores [6].

3 Proposed Method

Our method is based on a process that we call modeling. If any sequence of length l on the alphabet Σ is an ' l -mer', and if the number of positions at which any two l -mers differ is the 'distance' between them, then given two l -mers l_1 and l_2 that are at a distance of $2d$ from each other, we can construct another l -mer l_m that is at a distance d from both l_1 and l_2 as follows:

- (i) Note the points at which l_1 differs from l_2 (there are $2d$ such points)
- (ii) To obtain l_m , choose any d points in l_1 out of the $2d$ points of difference with l_2 , and replace them with the corresponding letters in l_2 .

For example, consider two l -mers $l_1 = \mathbf{AGATCG}$ and $l_2 = \mathbf{GGCACT}$. They differ in four positions (bold faced). We can obtain l_m by replacing the letters at any two of these four positions in l_1 , say the 1st and 3rd positions, by the corresponding letter in l_2 , thus forming $l_m = \mathbf{GGCTCG}$. This differs from both l_1 and l_2 at two positions.

We define the process of finding l_m by replacing d letters in l_1 with the corresponding letters of l_2 as *modeling* – l_1 is modeled on l_2 , and l_2 is a model for l_1 .

Based on the above definition, we make the following propositions. Note beforehand that for an input having n sequences of length L , there are $L - l + 1$ number of l -mers in each of the n input sequences, and the input sequences are numbered from 0 to $n - 1$.

Proposition 1: If there exists a motif of length l for the given n sequences, then in each sequence, at least one l -mer of length l is a d -neighbor of the motif. In particular, one of the $L - l + 1$ number of l -mers in sequence 0 is a d -neighbor of the motif.

Proposition 2: If there exists a motif of length l for the given n sequences, then the d -neighbor in every sequence from sequences 1 to $n - 1$ is at a distance of at most $2d$ from the d -neighbor in sequence 0.

Proposition 3: If there exists a motif of length l for the given n sequences, then the motif can be found by modeling the d -neighbor in sequence 0, on any d -neighbor in sequences 1 to $n - 1$ that is at a distance *exactly* $2d$ from it.

The logic of Proposition 3 is described as follows:

If two d -neighbors of the motif, dn_0 and dn_1 , are at a distance of exactly $2d$ from each other, then they are identical to each other at exactly $l - 2d$ points. If dn_0 and dn_1 are identical to each other at exactly $l - 2d$ points, the motif consists of the identical $l - 2d$ points. This follows necessarily from the definition of ' d -neighbor' (or, what is the same, from the definition of 'motif').

Therefore, given two d -neighbors of the motif that are at a distance of exactly $2d$ from each other, $l - 2d$ points of the motif are readily identified. The task is to identify the remaining $2d$ points of the motif.

The key is that these points are supplied by dn_0 and dn_1 themselves. The remaining $2d$ points in the motif are the same $2d$ points at which dn_0 and dn_1 differ. At each of the $2d$ points, the symbol in the motif is identical to the symbol at that point in either dn_0 or dn_1 . Further, the motif is identical to dn_0 at d of the $2d$ points, and to dn_1 at the remaining d of the $2d$ points. Again, this follows necessarily from the definition of d -neighbor.

So the task becomes one of constructing the motif by choosing d points from dn_0 out of its $2d$ points of difference with dn_1 , and choosing d points from dn_1 out of its $2d$ points of difference with dn_0 , and inserting the chosen $2d$ points into the corresponding $2d$ points in the motif.

We use modeling to achieve this effect. We take dn_0 and model it on dn_1 at d points out of the $2d$ points of difference. As there are $\binom{2d}{d}$ ways in which d points can be chosen out of $2d$ points, there are $\binom{2d}{d}$ ways in which dn_0 can be modeled on dn_1 . The motif can be found by taking each variant of dn_0 by turns, and testing to see if it is the motif.

As it is not known *a priori* which l -mer in sequence 0 is a d -neighbor of the motif, it is required to choose each l -mer one-by-one for processing. Similarly, as it is not known *a priori* which l -mers in sequences 1 to $n - 1$ are d -neighbors of the motif, all the l -mers in sequences 1 to $n - 1$ that are at a

distance of exactly $2d$ from the chosen l -mer in sequence 0, have to be found and considered as models.

When the d -neighbor of the motif in sequence 0 comes up for processing, all the d -neighbors of the motif in sequences 1 to $n - 1$ that are at a distance of exactly $2d$ from it will be found, during the search for all the l -mers that are at a distance of exactly $2d$ from it (along with other l -mers that happen to satisfy the property). Thereby sooner or later the d -neighbor in sequence 0 will be modeled on a d -neighbor of the motif that is at a distance of exactly $2d$ from it, and the motif will be found.

Accordingly, we construct the following method, consisting of twelve steps numbered 1 thru 12, to find the motif:

Step 1: Take the first l -mer of length l in sequence 0; call this the 'root'.

Step 2: Check whether the root is the motif by finding its distance from all the l -mers in sequences 1 to $n - 1$.

Step 3: If the root is within distance d of at least one l -mer in each of the sequences 1 to $n - 1$, then the root is the required motif. Return the root. Otherwise, continue with the next step.

Step 4: For the root, find all the $2d$ -neighbors in sequences 1 to $n - 1$. Call them the 'candidates'.

Step 5: From the set of candidates, take the first candidate that is at a distance of *exactly* $2d$ from the root. Call it the 'model-candidate'. If no such model-candidate exists, repeat the steps from Step 1, taking as the root the next l -mer in sequence 0.

Step 6: Model the root on the model-candidate. There are $\binom{2d}{d}$ possible combinations for modeling the root on the model-candidate. Take the first of the $\binom{2d}{d}$ possible combinations, and model the root according to it.

Step 7: Check the distance of the modeled root from all the candidates (i.e., all the $2d$ -neighbors of the root.)

Step 8: If the modeled root is within distance d of at least one candidate from each of the input sequences, it is the required motif. Return the modeled root. Otherwise, continue with the next step.

Step 9: Repeat Step 6 by taking the next of the possible $\binom{2d}{d}$ combinations and repeat Step 7. If all the $\binom{2d}{d}$ combinations are exhausted and the motif is not found, repeat the steps from Step 5, by taking as the model-candidate the next candidate that is at a distance of *exactly* $2d$ from the root.

Step 10: If all the candidates found in Step 4 are exhausted and the motif is not found, repeat the steps from Step 1, taking as the root the next l -mer in sequence 0.

Step 11: If all the l -mers in sequence 0 are exhausted and the motif is not found, relocate sequence 0 to the bottom of the input, such that it becomes sequence $n - 1$ and all the other sequences are promoted in the order by one step. In particular, sequence 1 becomes the new sequence 0. Then repeat the entire process from Step 1, with the new sequence 0.

Step 12. If $n - 1$ input sequences have been promoted to sequence 0 and the motif is not found, then stop and return an exception.

Explanation of Step 11: If all the l -mers in sequence 0 are exhausted and the motif is not found (but assumed to exist), it means that either:

- (i) the d -neighbor of the motif in sequence 0 is not at a distance of exactly d from the motif, or
- (ii) no d -neighbor is found in sequences 1 to $n - 1$ that is at a distance of exactly $2d$ from the d -neighbor in sequence 0

In either case, the fundamental requirement of the method, given under Proposition 3, is not met. Hence the method starts over with a different input sequence taken as sequence 0.

It should be noted that the occurrences of condition (i) and condition (ii) have a computable probability, which will be dealt with in Section 6.

4 Algorithms and Complexity Analysis

The algorithm that encapsulates the 12 steps is given below in Algorithm 1. We analyze the computational complexity of the algorithm as follows:

The algorithm halts when it finds the first motif. In the worst case, statement 1 is executed n times. For each execution of statement 1, statement 2 is executed at most $L - l + 1$ times. Statement 3 requires comparing the current root R_{ij} with all possible l -mers in $n - 1$ input strings for determining whether it is the motif. This requires at most $(n - 1) * (L - l + 1)$ l -mer comparisons. Each comparison involves at most l equality checks. If R_{ij} is not the motif, the control comes to statement 5. From here on, we look for a motif using modeling. In statement 5, the set C is constructed. This requires $(n - 1) * (L - l + 1)$ l -mer comparisons. Since the root is a string of size l over an alphabet of size 4, and a candidate is a $2d$ -neighbor of the root, the probability that any l -mer is a candidate is given by the ratio of the total number of $2d$ -neighbors that any l -mer can have, to the total number of l -mers possible. This ratio is:

$$P_C = \frac{\sum_{k=0}^{2d} 3^k \binom{l}{k}}{4^l} \quad (1)$$

The probable number of candidates in C is given by multiplying the probability P_C with the total number of l -mers in the field of search, which is $(n - 1) * (L - l + 1)$:

$$|C| = (n - 1) * (L - l + 1) * P_C \quad (2)$$

Among the candidates in C , those that are at a distance of exactly $2d$ are in the set C_m . These are the model-candidates. By statement 6, the root is modeled on at most $|C_m|$ model-candidates. As $C_m \subseteq C$, $|C_m| \leq |C|$ and therefore the root is modeled on at most $|C|$ model-candidates. In statement 7, the $2d$ points of difference between R_{ij} and one model-candidate are identified. This involves at most l equality checks. In statement 8, there are $\binom{2d}{d}$ possible combinations of d points among the $2d$ points of difference. In statement 9, R_{ij} is modeled according to one combination to get R_{ijm} , which takes at most d operations. For each R_{ijm} , statement 10 is executed to determine whether it is the motif, which involves at most $|C|$ l -mer comparisons. Each comparison involves at most l equality checks.

In summary, the upper-bound on the number of computations is given by:

$$|N| = O(n * (L - l + 1) * (|C|^l * (l + \binom{2d}{d}) * (d + |C|^l))) + (n - 1) * (L - l + 1) * 2^l \quad (3)$$

The values of n and L are usually constant (20 and 600 respectively), and as $L \gg l$ in all practical PMPs, $(L - l + 1) \cong L$. Omitting the constant factors, we have:

$$|N| = O(|C|^l * (l + \binom{2d}{d}) * (d + |C|^l)) \quad (4)$$

Thus the significant factors affecting the running time are the square of the number of candidates per root $|C|^2$, the number of combinations per candidate $\binom{2d}{d}$, and l and d .

Algorithm 1 FindMotif

Input: n, L, l, d

Output: M (motif)

```

1: for i = 0 to n - 1 do
2:   for j = 0 to L - l do
3:     check whether root  $R_{ij}$  (an  $l$ -mer in
       sequence  $i$  starting at position  $j$ ) is the motif
4:     if  $R_{ij}$  is not the motif then
5:       generate  $C$ , the set of all candidates, of which
          $C_m$  is the subset containing the model-candidates
6:       for each model-candidate  $c$  in  $C_m$  do
7:         identify the  $2d$  points of difference
           between  $R_{ij}$  and  $c$ 
8:         for each combination of  $d$  points of
           difference between  $R_{ij}$  and  $c$  do
9:           model  $R_{ij}$  on  $c$  to get  $R_{ijm}$ 
10:        check whether  $R_{ijm}$  is the motif using  $C$ 
11:        if  $R_{ijm}$  is the motif then
           output  $R_{ijm}$  as  $M$ 
           HALT
       end if
     end for
   end for
12:  end for
13:  end for
14: else
15:   output  $R_{ij}$  as  $M$ 
   HALT
16: end if
17: end for
18: end for

```

5 Overview of Statistical Properties

We have performed a detailed statistical analysis of various factors involved in the computation. Owing to space constraints, we discuss here the salient statistical properties revealed by the analysis, omitting the details.

As noted in Section 4, a major contribution to the computational workload of the method comes from the square of the number of candidates per root, $|C|^2$. An increase in this factor increases the computational workload. The value of $|C|^2$ depends on the value of l and d (Equations 1 and 2) such that:

- (i) increasing l keeping d fixed decreases $|C|$, and
- (ii) increasing d keeping l fixed increases $|C|$.

As the computational workload is proportional to $|C|^2$, it is highly sensitive to the ratio l/d . Increasing d keeping l fixed results in massive increase of workload for every step of increment of d . Our analysis shows that, for a broad range of

values of l (from 12 to at least 50), a massive increase of $|C|$ occurs when d is increased from $0.25l$ to $0.25l + 1$, rendering problem sizes in which d is greater than $0.25l$ challenging for this method. Conversely, decreasing d keeping l fixed results in a massive drop in $|C|$ for every step of decrement of d . Problem sizes in which d is lesser than $0.25l$ are solved extremely fast.

Another major contributor to the computational workload, as noted in Section 4, is the number of modeling combinations per model-candidate, given by $\binom{2d}{d}$. This number has a sharply increasing trend for every step of increment in d . Combined with the property that a massive increase of $|C|$ occurs when d is increased from $0.25l$ to $0.25l + 1$, a steep barrier exists at the boundary between those problem sizes in which $d \leq 0.25l$, and those in which $d > 0.25l$ (for all values of l ranging from 12 to at least 50).

Table I presents the values of $|C|$, $|C|^2$ and $\binom{2d}{d}$ for a few selected problem sizes at the $d = 0.25l$ boundary. The notable feature is that as the problem sizes increase, $|C|$ (and $|C|^2$) decrease sharply at every step, and $\binom{2d}{d}$ increases sharply. As the computational load is proportional to $|C|^2$ and $\binom{2d}{d}$, the opposing trends of $|C|^2$ and $\binom{2d}{d}$ mean that the trend of the computational load is essentially U-shaped, with a minima occurring in the mid-range of problem sizes. (The opposing trends of $|C|^2$ and $\binom{2d}{d}$ do not perfectly balance each other as their rates of change are not the same, and the proportions of their contribution to the workload are not the same. Therefore we should not expect a flat trend of the workload.)

TABLE I
NUMBER OF CANDIDATES PER ROOT AND NUMBER
OF MODEL COMBINATIONS PER CANDIDATE FOR
SELECTED PROBLEM SIZES

l	12	16	20	24	28	32	36	40	44	48
d	3	4	5	6	7	8	9	10	11	12
$ C $	609	302	153	79	41	22	11	6	3	2
$ C ^2$	370881	91204	23409	6241	1681	484	121	36	9	4
$\binom{2d}{d}$	20	70	252	924	3432	12870	48620	184756	705432	2704156

Note: All values of d are equal to $0.25l$.

We now turn to the statistical properties of d -neighbors. For modeling to successfully find the motif, the d -neighbor of the motif in sequence 0 has to be at a distance of exactly d from the motif. Those d -neighbors that are at a distance of less than d from the motif are valid d -neighbors, but do not contain enough information to find the motif. As such, the d -neighbor in sequence 0 may or may not be at a distance of exactly d from the motif. The statistics show that the probability of the d -neighbor of the motif in sequence 0 being at a distance of exactly d from the motif is 90% or better, for all problem sizes in which l is in the range of 12 to 50 and d is $d \leq 0.25l$. (Uniform random distribution of d -neighbors is assumed.)

In the 10% of the cases in which the d -neighbor in sequence 0 is at a distance of less than d from the motif, after processing the entire sequence 0 the motif will not be found and method will enter Step 11. Input sequence 1 will become the new sequence 0. The probability that the d -neighbors of

the motif in the first two input sequences are both at a distance less than d from the motif is $\sim 1\%$ (by multiplying the 10% probability of each sequence, as they are mutually independent.) Therefore probability that the d -neighbor of the motif in the new sequence 0 is at a distance of exactly d from the motif is about 99%, and the method can be expected to enter Step 11 for a second time only in 1% of the cases.

The second condition for modeling to successfully find the motif is that at least one d -neighbor in input sequences 1 to $n - 1$ should be at a distance of exactly $2d$ from the d -neighbor in sequence 0. The probability of such a d -neighbor existing has been found to depend on the ratio l / d . If d is increased keeping l fixed, the probability decreases, and if l is increased keeping d fixed, the probability increases. If the probability is too low and therefore such a d -neighbor does not exist, a different input sequence has to be taken as sequence 0. The d -neighbor in the new sequence 0 may be such that there is at least one d -neighbor in input sequences 1 to $n - 1$ that is at a distance of exactly $2d$ from it. That is, Step 11 has to be executed.

For problem sizes that have higher values of d relative to l , the method enters Step 11 more number of times. The number of times that the method enters Step 11 is called the Swap factor (S), and it can be probabilistically calculated for every problem size, from the statistical properties of d -neighbors through the values of l and d . The problem of swapping, however, has been found to become acute only for PMP sizes of (36, 9) and higher (when d is restricted to $0.25l$ or less). Table II shows the calculated values of the Swap factor S for selected problem sizes having l in the range of 36 to 50.

TABLE II
SWAP FACTOR S FOR SELECTED PROBLEM SIZES

$l = 36$					
d	8	9	10	11	12
S	0	1	2	7	25
$l = 40$					
d	9	10	11	12	13
S	0	1	4	11	38
$l = 44$					
d	10	11	12	13	14
S	1	2	5	15	51
$l = 48$					
d	11	12	13	14	15
S	1	3	8	21	67
$l = 50$					
d	11	12	13	14	15
S	1	2	6	15	44

6 Experimental Results

We implemented the modeling method in a single-threaded C++ program and executed it for 11 selected problem sizes on a system with 2.2GHz Intel Core2 Duo Processor T6600, 800 MHz FSB and 4 GB RAM.

Although the algorithm terminates when the first motif is found, in the implementation we processed all the roots so as to observe the processing time for the entire sequence 0. This

is required because the 'correct' root (i.e. the d -neighbor of the motif) in sequence 0 can occur anywhere in the sequence from position 0 to position $L - l$, which means the motif may be found at any stage in the processing of sequence 0. The time taken to find the motif is therefore not a meaningful indicator of performance. The meaningful indicator is the time taken to process the entire sequence 0.

Also in the implementation, 20 trials were conducted for each problem size, using each of the 20 input sequences as sequence 0, by turns. The rotation was done to observe the variation in processing time when different input sequences are taken as sequence 0. (This rotation of input sequences is unrelated to Step 11 of the method, by which if the motif is not found after processing sequence 0, another input sequence is used as sequence 0, till all the 20 input sequences are used up. It has the same effect as Step 11, however, and therefore, Step 11 of the method was omitted in the test runs as redundant.)

The running time, has to be subjected to certain considerations. Firstly, because the d -neighbor of the motif in sequence 0 is at a distance of exactly d from the motif in only 90% of the cases, the extra time taken when the method enters Step 11 in 10% of the cases has to be accounted for. Secondly, when the Swap factor S is ≥ 1 , the method enters Step 11 S times, and processes a new sequence 0 each time. Therefore the running time has to be multiplied by S . (Only problem sizes (36, 9) and above are affected by this, however.) Thirdly, the time taken to process sequence 0 is different when a different input sequence is taken as sequence 0. This is because all the roots are different and exactly the same number of candidates will not be found for the roots (see Equations 1 and 2). As the complexity is proportional to $|C|^2$, the running time is sensitive to fluctuations in $|C|$.

For problem size (36, 9), the lowest time among the 20 trials, to process all the roots in sequence 0 (=565 in number), was 21 seconds. The motif was found in 9 seconds by modeling root # 318 on l -mer # 185 of sequence 4. (This means that the d -neighbor of the motif in sequence 0 was at position 318, and there was a d -neighbor of the motif in sequence 4 at position 119, that was its '2d' neighbor.)

The highest time among the 20 trials was 244 seconds. The motif was found in 129 seconds by modeling root # 185 on l -mer # 318 of sequence 16.

The average time over the 20 trials, for problem size (36, 9), was 117 seconds. The motif was found on 14 of the 20 trials and not found on 6.

We term the average time over the 20 trials as t_{AVG20} , and deem the indicator of the time taken to find the motif in sequence 0 to be $0.5 * t_{AVG20}$. This is the intermediate case, between the two extremes of the 'correct' root occurring at position 0 (in which case it takes ~ 0 time to find the motif), and occurring at position $L - l$ (in which case it takes the full average time of t_{AVG20}).

To account for the extra time taken when the method enters Step 11, in 10% of the cases that the d -neighbor of the motif in sequence 0 is not at a distance of exactly d from the motif, an amortized amount of 10% is added to t_{AVG20} .

The time to process sequence 0, obtained from these two considerations, is:

$$t_{CORR} = 0.5 * t_{AVG20} + 0.1 * t_{AVG20} = 0.6 * t_{AVG20} \quad (5)$$

For problem size (36, 9), t_{CORR} is $0.6 * 117 = 71$ sec. We now consider the extra time taken on account of the Swap factor S . If S swaps are expected, an amount of time equal to $S * t_{AVG20}$ has to be added to t_{CORR} to get the expected time taken to find the motif. Thereby, the expected time taken to find the motif t_{EXP} is: $t_{EXP} = (S + 0.6) * t_{AVG20}$ (6)

Note that the full t_{AVG20} rather than half has to be considered for swap time, because the method always runs through the entire sequence 0 before making a swap.

For problem size (36, 9), 1 swap is expected (see Table II). Therefore, the expected time taken to find the motif t_{EXP} for problem size (36, 9) comes to $(1 + 0.6) * 117 = 187$ sec.

Table III shows the values of the expected time taken to solve problem sizes in the range of $l = 12$ to 50 having $d = 0.25l$. For each problem size, the amount of time added on account of swaps is indicated, as is the 10% correction amount to account for the 'correct' root not occurring in sequence # 0 of the input 10% of the time.

It can be observed from Table III that the best-case performance in the test runs was for problem size (32, 8), with an expected time of 48 seconds, and the worst-case performance was for problem size (48, 12), with an expected time of 6892 seconds, or about 1.9 hours.

The trend in Table III of the expected running time is more or less flat in the range (12, 3) to (24, 6). In the range (28, 7) to (48, 12), there is a clear U-shaped trend with a minima occurring in the mid-range at (32, 8). In this range, the trend is in line with what was expected for the entire range from the statistical analysis in Section 5 (the value for (50, 12) is irrelevant for the trend, as it is an anomalous problem size in the table.) The other notable feature in Table III is the variation over 20 trials, of the range of time taken to process sequence 0. The ratio of the maximum time taken to the minimum time taken increases from about 1 at (24, 6) to about 28 at (44, 11), and then drops to being about 10 for (48, 12) and 5 for (50, 12). The reason for this trend remains to be investigated.

For problem sizes in which d is less than 25% of l , the method is expected to perform much faster than for problem sizes in which d is exactly 25% of l (see Section 5). This has been observed to be the case in practice, and as a ready indicator of the increase in speed for problem sizes in which d is less than 25% of l , the time taken for problem size (50, 12) is included in Table III. This can be compared with the time taken for problem size (48, 12). Although l is larger in the (50, 12) problem, it is solved in less than a third of the time as (48, 12), because d is slightly less than 25% of l in it. The consequence of a slightly smaller d is a significantly reduced computational workload, and also a smaller swap factor S . (It can be observed from Table II that the swap factor decreases with a decrease in d relative to l .) These factors combine to

greatly reduce the time taken to solve the (50, 12) problem relative to the (48, 12) problem. Other problem sizes in which d is $< 0.25l$ have been omitted due to space constraints.

TABLE III
TIME TAKEN BY MODELING METHOD FOR
SELECTED PROBLEM SIZES

(1) Problem size	(2) Time for Seq. 0			(3) t_{CORR} 0.6 x (2c)	(4) Swap factor S	(5) S * t_{AVG20} (4)x(2c)	(6) t_{EXP} (3)+(5)
	Min	Max	t_{AVG20}				
	(a)	(b)	(c)				
(12, 3)	1216	1335	1259	756	0	0	756
(16, 4)	1236	1326	1277	767	0	0	767
(20, 5)	1372	1643	1477	887	0	0	887
(24, 6)	1195	1679	1408	846	0	0	846
(28, 7)	288	519	381	229	0	0	229
(32, 8)	25	150	80	48	0	0	48
(36, 9)	21	244	117	71	1	117	187
(40, 10)	33	522	163	98	1	163	262
(44, 11)	24	666	367	221	2	735	955
(48, 12)	645	6625	1939	1164	3	5818	6982
(50, 12)	389	2126	869	522	2	1738	2260

Note: All times are in seconds.

Min, Max and Average times are from 20 trials.

It should be noted that t_{EXP} reported in Table III is derived from practically observed values, and can vary either way, when working with different input sets generated of the same problem size. A different set of n input sequences would have a different distribution of l -mers, affecting the values of $|C|$ and also possibly the number of actual swaps that happen. However the overall trend over the different problem sizes will be more or less the same.

Further, as with any computer program, t_{EXP} depends heavily on the platform used (including the hardware and the operating system) and also the implementation (for example, using the bitset data structure rather than character or string formats for the input sequences and l -mers results in a speed-up of about 2x, as comparison operations run much faster with the bitset data structure).

Coming to the memory requirements, the method uses very little memory. We have calculated that the worst-case memory requirement is well under 1 MB, which is negligible.

From these facts, it is established that the method is very effective for solving PMPs as large as (48, 12). Thus it solves problems much larger than those reported solved in the literature, in running times much shorter than the times reported for smaller problems in the literature. For comparison, Table IV contains representative samples of the time taken by various other methods as reported in the literature.

7 Summary and Future Work

An efficient method of solving the Planted Motif Problem has been developed that uses a technique called modeling. The method is very fast over a broad range of problem sizes, and

takes up very little memory. Using the method, PMPs having problem sizes up to (48, 12) have been solved, with a single-threaded program executed on a system having one 2.2GHz Intel Core2 Duo Processor T6600, 800 MHz FSB and 4 GB RAM.

The high speed of the method, combined with low memory requirement brings motif-finding problems of the order of (48, 12) within easy reach of ordinary desktop/laptop computers. The program can be run comfortably along with the other applications that are typically found in a desktop environment. (In other words, high-end / dedicated systems are not required.)

In conclusion, we note that modeling is independent of the radix of the alphabet, as it works by one-to-one substitution of characters. The same amount of time is taken to model l -mers over an alphabet of size 20, say, as it takes to model l -mers over an alphabet of size 4. As the method is not restricted to the A, C, G, T alphabet of the Planted Motif Problem, it can have applications in other areas of pattern-finding, which is to be investigated.

TABLE IV
REPRESENTATIVE SAMPLES OF TIME TAKEN BY
VARIOUS OTHER METHODS

A (l, d)	Algorithm			
	Time	Time	Time	Time
	PROJECTION	Styczynski et al.'s	ExVote	
(10,2)	(161.1s)	(8 min)	(0.1 s)	
(11,2)	(12.5 s)	(< 1 min)	(0.7 s)	
(12,3)	(8.7 min)	(10.5 h)	(9.8 s)	
(13,3)	(46.0 s)	(10 min)	(17.4 s)	
(14,4)	(15.4 min)	(> 3 months)	(197.5 s)	
(15,4)	(129.0 s)	(6 h)	(206.1 s)	
(17,5)	(273.2 s)	(3 weeks)	(27 min)	
Source: An Efficient Algorithm for Extended (l, d)-Motif Problem With Unknown Number of Binding Sites, by Leung and Chin [1]				
B (l, d)	Algorithm			
	Stemming	MITRA	PMSPPrune	RISOTTO
(9,2)	0.95s	0.89s	0.99s	1.64s
(11,3)	8.8s	17.9s	10.4s	24.6s
(13,4)	31s	203s	103s	291s
(15,5)	187s	1835s	858s	2974s
(17,6)	1462s	4012s	7743s	29792s
(19,7)	8397s	n/a	81010s	n/a
Source: Efficient Discovery of Common Patterns in Sequences Over Large Alphabets, by Kuksa And Pavlovic [9]				
C (l, d)	Algorithm			
	BitBased			
	16 CPU	8 CPU	4 CPU	
(11,3)	1s	1s	2s	
(13,4)	2s	2s	4s	
(15,5)	15s	24s	47s	
(17,6)	2.8m	5m	9.2m	
(19,7)	35m	63m	112m	
(21,8)	7.8h	-	-	

Source: An Efficient Multicore Implementation of Planted Motif Problem, by Ranjan et al [7]

Note on Table IV: Problems larger than (21,8) have not been reported solved to the best of our knowledge.

8 References

- [1] H.C.M. Leung and F.Y.L. Chin, "An efficient algorithm for the extended (l,d)-motif problem with unknown number of binding sites", *Proc. 5th IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*, 2004.
- [2] P. Pevzner and S.H. Sze, "Combinatorial approaches to finding subtle signals in DNA sequences", *Proc. Eighth International Conference on Intelligent Systems for Molecular Biology*, 2000, pp. 269-278.
- [3] S. Rajasekaran, S. Balla, and C.-H. Huang, "Exact algorithms for planted motif challenge problems", *Proc. Third Asia-Pacific Bioinformatics Conference*, Singapore, 2005.
- [4] A. Price, S. Ramabhadran, and P. A. Pevzner, "Finding subtle motifs by branching from sample strings", *Bioinformatics*, 1 (1), 2003, pp. 1-7.
- [5] E. Eskin and P. Pevzner, "Finding composite regulatory patterns in DNA sequences", *Bioinformatics SI*, 2002, pp. 354-363.
- [6] J. Buhler and M. Tompa, "Finding motifs using random projections", *Proc. Fifth Annual International Conference on Computational Molecular Biology (RECOMB)*, April 2001.
- [7] N. S. Dasari, R. Desh, and M. Zubair, "An efficient multicore implementation of planted motif problem", *Proc. 2010 International Conference on High Performance Computing & Simulation (HPCS 2010)*, France, 2010, pp. 9-15.
- [8] F.Y.L. Chin and H.C.M. Leung, "Voting algorithms for discovering long motifs", in *Proc. 3rd Asia-Pacific Bioinformatics Conference (APBC)*, Singapore, 2005. pp. 261-271.
- [9] P.P. Kuksa and V. Pavlovic, "Efficient discovery of common patterns in sequences over large alphabets", in *DIMACS Technical Report*, 2009.
- [10] J. Davila, S. Balla, and S. Rajasekaran, "Fast and practical algorithms for planted (l, d) motif search", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4), 2007, pp. 544-552.
- [11] N. Pisanti, A. Carvalho, L. Marsan, and M.-F. Sagot, "RISOTTO: Fast extraction of motifs with mismatches", *Proc. Latin American Theoretical Informatics Symposium (LATIN)*, Chile, 2006, pp. 757-768.
- [12] M.P. Styczynski, K.L. Jensen, I. Rigoutsos, and G.N. Stephanopoulos, "An extension and novel solution to the (l,d)-motif challenge problem", *Genome Informatics*, 15, 2004, pp 63-71.

Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1

Xin Wang^{1,4}, Liran Juan^{1,5}, Junjie Lv⁴, Kejun Wang⁴, Jeremy Sanford⁶ and Yunlong Liu^{1,2,3,§}

¹Center for Computational Biology and Bioinformatics, ²Department of Medical and Molecular Genetics,

³Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN 46202, United States

⁴College of Automation, Harbin Engineering University, Harbin, Heilongjiang 150001, China

⁵School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

⁶Department of Molecular, Cellular and Developmental Biology, University of California Santa Cruz, Santa Cruz, California 95064, United States

Abstract - RNA-binding proteins (RBPs) play diverse roles in eukaryotic RNA processing. Despite their pervasive functions in coding and non-coding RNA biogenesis and regulation, elucidating the specificities that define protein-RNA interactions remains a major challenge. Here, we describe a novel model-based approach — *RNAMotifModeler* to identify binding consensus of RBPs by integrating sequence features and RNA secondary structures. Using RNA sequences derived from Cross-linking immunoprecipitation (CLIP) followed by high-throughput sequencing for SRSF1 proteins, we identified a purine-rich octamer 'AGAAGAAG' in a highly single-stranded RNA context, which is consistent with previous knowledge. The successful implementation on SRSF1 CLIP-seq data demonstrates great potential to improve our understanding on the binding specificity of RNA binding proteins.

Keywords: protein-RNA binding, RNA secondary structure, motif, SRSF1, particle swarm optimization

1 Introduction

RNA-binding proteins (RBPs) are implicated in virtually every step of post-transcriptional gene expression including pre-mRNA splicing, RNA editing and polyadenylation [1]. These proteins possess a diverse array of structurally and functionally distinct RNA-binding domains such as RNA recognition motifs (RRM), KH domains, RGG boxes, zinc finger, double-stranded RNA-binding domain, etc [1]. Although the structures of many RNA binding domains have been solved at high resolution, establishing the sequence and RNA-structural determinants to binding specificity remains largely unexplored.

Several methods for elucidating the specificity of protein-RNA interactions enable rapid advances in our understanding of RBP functions. One recent innovation is the Cross-Linking ImmunoPrecipitation (CLIP). CLIP exploits photoreactive residues in RNA and polypeptides to generate covalently linked complexes. Because UV irradiation does not induce

protein-protein cross-links CLIP is thought to be more specific than other IP based assays for protein-RNA interactions. CLIP was successfully applied to identify mRNA targets of the NOVA protein, a neural splicing factor associated with paraneoplastic opsoclonus myoclonus ataxia (POMA) [2-4]. Coupling CLIP with next-generation high-throughput sequencing technology, known as CLIP-seq or HITS-CLIP, provides a cost-efficient method to increase the sensitivity of the assay by surveying the RNA landscape on a more global scale. Several groups have successfully implemented CLIP-seq analysis of NOVA, SRSF1, fox2 and PTB proteins in mammalian systems [2, 5-7]. Both *MEME* and Z-score statistics have been used to reveal consensus binding motifs that are overrepresented in CLIP-Seq data [2, 6]. Although Z-score statistics may be able to find out the overrepresented sequence motifs, it does not consider the degenerated feature of the binding specificities of RBPs. *MEME*-based method is well known to be an excellent tool for cases only regarding sequence specificity [8]. Neither of these approaches can ascertain the roles of RNA secondary structure in establishing the context of the protein-RNA interaction. Hiller et al. extended *MEME* by adding a pre-computing procedure to measure single-strandedness of RNA sequence as *a priori* information to guide the motif search. They demonstrated that their model, *MEMERIS*, is able to identify binding motifs located in single-stranded regions with applications to both artificial and biological data [9]. Recently, Kazan et al. proposed *RNAcontext* for learning both sequence and structural binding preferences of RNA-binding proteins [10].

Here we describe a model-based approach—*RNAMotifModeler* to evaluate protein-RNA interactions using a retained binding affinity ratio, which is considered to be affected by two major factors—sequence degeneracy and RNA secondary structure deviation. *RNAMotifModeler* incorporates predicted unpaired probability of each nucleotide in the protein-RNA binding regions; such probability is derived from RNA secondary prediction algorithms (e.g. *RNAfold* [2]) based on the nucleotide compositions of the neighbouring flanking sequences. This strategy is different

from RNAContext, which uses predicted RNA secondary structures as input such as 'Paired', 'Hairpin Loop', 'Unstructured' or 'Miscellaneous'. Unlike MEMERIS, RNAMotifModeler uses the base-pairing probability for each nucleotide rather than the entire sequence (PU or EF values) [3]. For each binding instance, RNAMotifModeler defines a score that evaluates the consensus binding site within an optimal structural context, and aims at searching for an optimal RNA sequence-structural consensus for an RNA binding protein. These features enhance our ability to calculate and estimate the sequences that yield the highest binding affinity for a specific RBP.

We tested RNAMotifModeler on CLIP-seq data that profile the transcriptome-wide binding pattern of SRSF1, serine/arginine-rich splicing factor 1 [4]. The sequence features of the binding motifs is consistent with the experimentally defined *cis*-acting RNA elements recognized by SRSF1 [5]. Interestingly, the prediction suggests that the second and fifth bases of SRSF1 octamer motif have stronger sequence specificities, but lower p-values of unpaired probabilities, while the third, fourth, sixth and seventh bases are more significantly to be single-stranded, but have less sequence specificities. Therefore, we hypothesize that the sequence and structure specificities are both required and are playing complementary roles during binding site recognition of SRSF1.

2 Results

SRSF1 is an essential splicing factor with multiple roles in post-transcriptional gene expression [6]. SRSF1 is also a potent proto-oncogene and implicated in maintaining genome stability [7]. Moreover, loss of SRSF1 binding sites by mutations linked to genetic diseases can induce aberrant patterns of pre-mRNA splicing [4]. Thus considerable effort has been focused on defining the binding specificity and RNA targets of SRSF1. Here we report a novel model-based approach intended to examine the contributions of structural and sequence elements in RNA fragments co-purified with SRSF1 by CLIP.

2.1 Workflow of RNAMotifModeler

The first step of *RNAMotifModeler* is to do data preparations. In the present study, 904 positive gold standard sequences were selected from commonly targeted regions across three out of four samples in our previous SRSF1 CLIP-seq experiments [4]. The same number of negative sequences were randomly picked from non-SRSF1-targeted regions falling in the same genomic category (exonic, intronic, intergenic, etc) as their positive counterparts. Base pairing probabilities of each nucleotide to its neighbours were subsequently predicted by RNAfold [2] (ViennaRNA package, version 1.8.5) for both positive and negative gold standard sequences.

Our next step, as shown in Fig. 1, is to identify sequence-structural consensus using gold standard sequences and

corresponding base pairing probabilities derived from RNAfold. We took an iterative approach that alternates between: 1) optimization of parameters specifying sequence degeneracy and structural context given a reference motif (the optimal binding sequence), and 2) searching for optimal reference motif given the estimated parameters by evaluation of each motif candidate's contribution to binding affinities of positive gold standard sequences (more details in Methods). The above two steps will be repeated until a convergence when the starting motif candidate makes the most contribution to binding affinities.

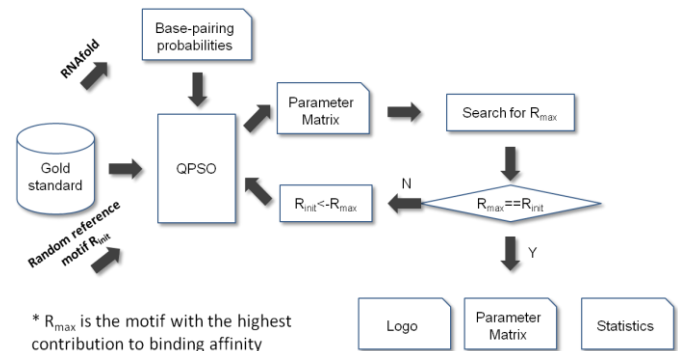


Fig. 1. Workflow of RNAMotifModeler

Finally, RNAMotifModeler outputs the converged reference motif, optimal parameters, statistical evaluation such as the AUC scores. The AUC scores are measured by the area under the ROC (Receiver Operating Characteristic) curves derived from predictions of gold standard sequences being bound by SRSF1 proteins using the predicted parameters. In order to predict binding sites of SRSF1 proteins, we pick the sequence binding affinity yielding the maximal prediction accuracy as a cutoff score. Based on the predicted reference motif and corresponding parameters, positive gold-standard sequences can be scanned to find all potential binding sites with binding affinities higher than the cutoff score. These binding sites can be further used to create a sequence consensus logo and transformed to positional weight matrix, which is much more widely used.

2.2 Convergence of SRSF1 consensus motif searching

We call the converging path from a starting motif candidate to the final consensus motif a *motif searching pathway*. This graph provides a visual demonstration on the pathways through which the reference motifs are determined. To have a global overview of the convergence, motif searching pathways for all motif candidates are organized together to form a *motif searching graph*. In the particular case of hexamer predictions for SRSF1, all 4096 motif candidates converge to a short list of candidates (Fig. 2). All motif candidates converge within three iterations, of which 85.7% converge after the first iteration. AGAAGA, AAGAAG and GAAGAA are top three hexamers with the highest in-degrees, responsible for 99.7% of all motif

candidates (Table 1). The other twelve reference motifs are closely related to these three motifs, only with one or two sequence alterations. It is also noted that nearly an equal number of motif candidates converge to each one of the top three reference motifs. More interestingly, these hexamers share a core of 'AAGA' indicating that they may be adjacent to each other in RNA fragments.

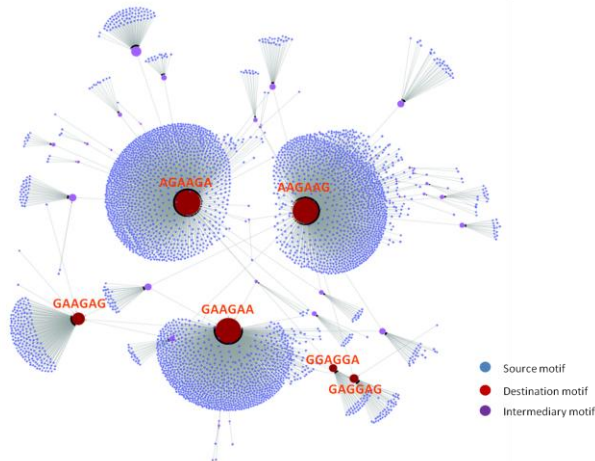


Fig. 2. Motif searching graph. Source, intermediary and destination motifs are denoted by nodes colored in blue, purple and red, respectively. The size of node is proportional to its in-degree. Arrows between nodes indicate converging directions. This figure demonstrates the fast convergence of the vast majority of motif candidates using the Quantum Particle Swarm Optimization algorithm.

Table 1. Converged motifs and corresponding numbers of source motifs

Converged motif	No. of source motifs
AGAAGA	1484
AAGAAG	1375
GAAGAA	1225
others	12

RNAMOifModeler provides an option to predict sequence-structural consensus of different lengths. For short motifs, it is suggested to perform predictions starting from every potential motif candidate and generate a motif searching graph to inspect the global convergence. For longer motifs, however, generating such a graph will be computationally expensive. In this case, we conduct predictions starting from a sufficient number of motif candidates randomly picked from the motif space. The converged motif with the highest prediction power, measured by AUC, is selected as the optimal one.

2.3 Predicted sequence and structural features of SRSF1 binding regions

To better compare RNAMOifModeler predictions with the SRSF1 binding motif reported previously, here we focus on octamer predictions. Consistent with the sequence consensus predicted by MEME [4], the reference motif of SRSF1 identified using RNAMOifModeler is also 'AGAAGAAG'. The optimal parameters associated with the reference motif are displayed in Table 2. The first row listed the reference sequence motif identified while the following four rows include retained binding affinity ratios due to

sequence alterations. The last row in Table 2, however, is constituted by unpaired probabilities for all nucleotides in the motif, indicating the optimal RNA secondary structure of SRSF1 binding regions. We note that every nucleotide of the predicted SRSF1 binding motif has a very high probability to be single-stranded, suggesting that SRSF1 proteins tend to bind on highly unpaired RNA regions.

Table 2. Predicted sequence-structural consensus of SRSF1

	A	G	A	A	G	A	A	G
A	1.00	0.17	1.00	1.00	0.24	1.00	1.00	0.81
G	0.79	1.00	0.65	0.90	1.00	0.84	1.00	1.00
C	0.52	0.32	0.50	0.16	0.35	0.02	0.34	0.63
U	0.75	0.15	0.39	0.63	0.09	0.06	0.73	0.55
UP	0.99	0.96	0.99	0.99	0.98	0.99	0.92	0.83

Based on the predicted optimal parameters, we obtained an AUC of 0.875 (Fig. 3 A) and an maximal accuracy of 0.803 (Fig. 3 B), which are both higher than the MEME-based prediction, of which the AUC is 0.86 and maximal accuracy is 0.78 [4].

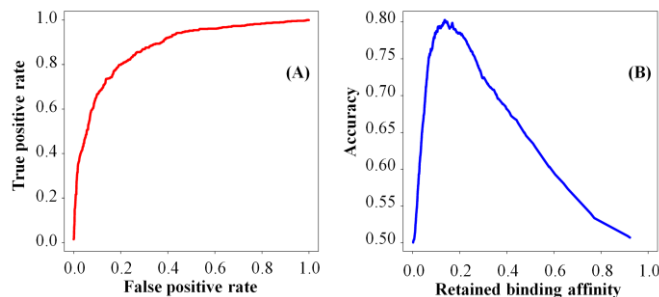


Fig. 3. ROC curve and accuracy curve describing the prediction power of RNAMOifModeler for SRSF1 proteins

To visualize the predicted SRSF1 sequence consensus more straightforwardly, positive gold standard sequence were scanned to search binding sites with binding affinities higher than the threshold 0.138, based on which a sequence logo (Fig. 4) was created by Weblogo [8]. This motif is consistent with the positional weight matrix (PWM) identified by MEME using the same gold standard sequences in our previous study [4], and is similar to the motifs found by other groups [9-11].



Fig. 4. Sequence consensus logo for SRSF1 proteins

2.4 SRSF1-RNA binding regions are significantly single-stranded

To further test the hypothesis that RNA regions bound by SRSF1 proteins are significantly unpaired, we compared

2904 binding sites predicted by RNAMotifModeler with a same number of controls binding sites, randomly selected in the same positive gold standard sequences. P-values were obtained from Wilcoxon rank sum tests on unpaired probabilities of nucleotides between predicted and randomly selected binding sites. All median unpaired probabilities of positive binding sites are significantly higher than controls (Fig. 5B). Wilcoxon tests were also performed on unpaired probabilities of nucleotides between predicted binding sites and random binding sites selected in negative gold standard sequences. For all the eight nucleotides, binding sites in positive gold standard sequences tend to be single-stranded (Fig. 5A).

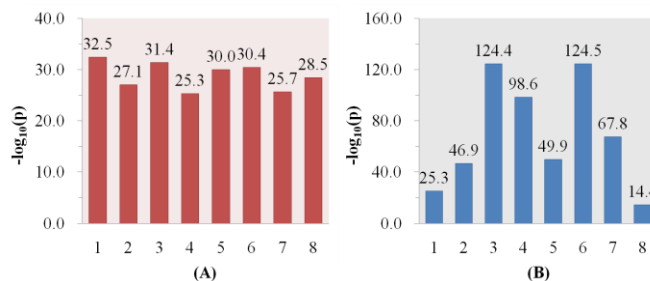


Fig. 5. P-values of nucleotides in the motif suggesting significant single-strandedness. The p-values are derived from Wilcoxon tests, with the alternative hypothesis that (A) predicted binding sites in positive gold standard sequences are more single-stranded than their counterparts in negative gold standard sequences, and (B) binding sites predicted by RNAMotifModeler are more single-stranded than randomly selected binding sites in positive gold standard sequences.

The two groups of Wilcoxon tests demonstrate that binding sites predicted by RNAMotifModeler are not only more single-stranded in positive gold standard sequences than negative controls, but also less structured than by chance within the same CLIP sequences. More interestingly, comparing Fig. 5 B with Fig. 4, we found that the second and fifth nucleotide of SRSF1 motif have much stronger sequence specificities but lower p-values of unpaired probabilities, while the third, fourth, sixth and seventh nucleotide are more significantly single-stranded but have less sequence specificities, suggesting that both the sequence and a lack of secondary structure may play complementary roles in SRSF1-RNA binding.

2.5 Predictions before and after incorporating RNA structure information

RNAMotifModeler can also predict consensus motifs without using structural information. Using the same positive and negative gold-standard sequences, we identified the same reference motif 'AGAAGAAG' and very similar retained binding affinity ratios due to sequence alterations. However, we obtained an optimal AUC of 0.853 and the maximal accuracy of 0.789, suggesting a slightly reduced prediction power when discarding RNA secondary structure information.

Using identified parameter matrix based on only sequences we predicted 2295 binding sites, of which 81% are

commonly identified by incorporating RNA secondary structure information (Fig. 6 A). The unpaired probabilities of the other 437 binding sites are significantly lower than identified binding sites using both sequence and structural information (Fig. 6 B and 6 C). Except the third nucleotide of motif, all of the unpaired probabilities of these binding sites are even lower than background, indicating that binding sites predictions may result in a considerable number of false positives due to ignoring RNA secondary structures. Bringing in RNA secondary structure information, we found 1046 more binding sites. These binding sites may have low sequence specificities, but could be of high structure specificities. Although the AUC increases only by 0.023 after introducing RNA secondary structure information, false positive and false negative binding sites are both significantly reduced.

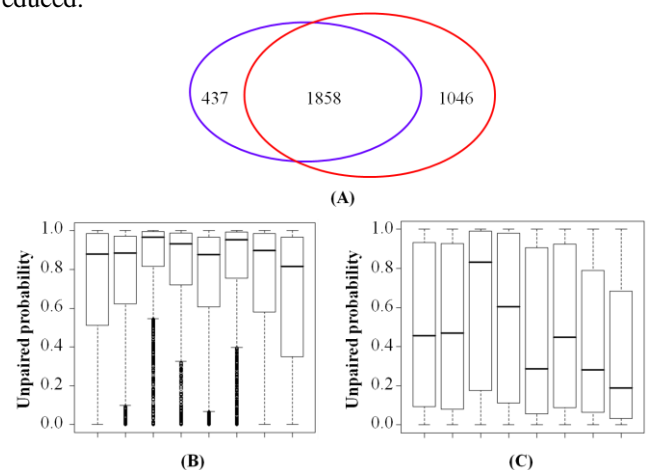


Fig. 6. Comparisons between predicted binding sites before and after incorporating RNA secondary structure information. (A) The number of binding sites predicted by RNAMotifModeler using only sequence information (blue ellipse) and after incorporating structure information (red ellipse); (B) Boxplots of unpaired probabilities of 1858 binding sites both predicted by the two methods; (C) Boxplots of unpaired probabilities of 437 binding sites only predicted without RNA secondary structure information

3 Discussions

In recent years, there is an increasing interest in using high-throughput sequencing technology to study protein-RNA binding specificities, but almost all of currently available bioinformatic approaches used for this purpose do not take into account RNA secondary structures, which have been demonstrated to have critical impact on protein-RNA binding in previous biochemical experiments. Thus, the motivation of our proposed model—RNAMotifModeler is to predict both structural and sequence specificities of protein-RNA binding regions.

RNAMotifModeler incorporates RNA secondary structure using RNAfold derived probabilities of nucleotides being paired with its neighbours. The preference for base-pairing probabilities over RNA secondary structures is due to a couple of concerns: a) It is very difficult to take into account RNA secondary structures directly in many real applications because of multiple RNA folding choices including optimal and sub-optimal structures; b) Unlike

MEMERIS, RNAMotifModeler tries to identify the optimal structural feature that is expected to represent the base pairing probability for each nucleotide in motif. Therefore, we did not use PU or EF scores [3], which are the measurements of single-strandedness of protein-binding regions in MEMERIS. c) The base-pairing probabilities predicted by RNAfold program [2] account for all possible secondary structures.

It is noted from our predictions that almost all unpaired probabilities of bases in the reference motif of SRSF1 predicted by RNAMotifModeler are close to 1, suggesting a very strong preference of SRSF1 to single-stranded RNA context. The statistical significance was further proved by two groups of Wilcoxon tests. These findings are consistent with previous evidences of SRSF1 proteins. It is known that SRSF1 protein contains an arginine-serine rich region (RS domain) and two RNA recognition motifs (RRMs), through which SRSF1 recognizes specific RNA regions [12, 13]. Importantly, RRM is one of the single-stranded RNA-binding domains of proteins [14]. Comparing the sequence consensus and p-values derived from Wilcoxon tests between the unpaired probabilities of predicted binding sites and negative controls, we propose that sequence and structural specificity may be two complementary factors that both facilitate the binding site recognition of SRSF1.

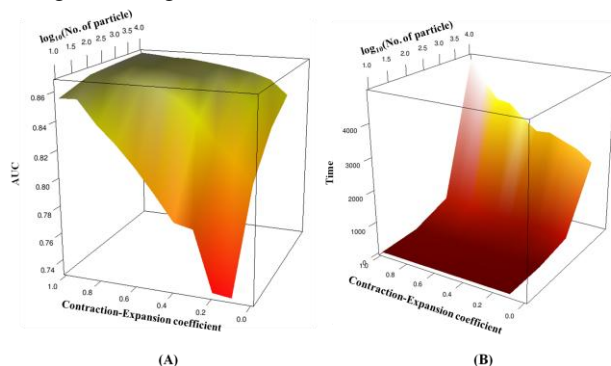


Fig. 7. 3D heatmaps illustrating the effects of the number of particles and the contraction-expansion coefficient in QPSO. (A) The prediction power measure by AUC, and (B) the time consumed are affected by the number of particles and the Contraction-Expansion coefficient, which are two critical parameters of QPSO.

RNAMotifModeler also provides an option to predict only sequence consensus motifs. This can be potentially applied to other fields that only focus on sequence specificities such as prediction of protein-DNA binding motifs. In the specific application to SRSF1, we found that the prediction power in this case is still comparable with MEME-based approach, although the AUC and maximum accuracy were both slightly reduced when RNA secondary structure information was not incorporated. Moreover, only using sequence specificity to predict binding sites could result in many false positives and false negatives.

Two parameters—the number of particles n_p and the contraction-expansion coefficient β of the Quantum Particle Swarm Optimization greatly affect the predicting accuracy of RNAMotifModeler. To estimate and set up these parameters

prior to the optimization procedure, we did a series of hexamer motif searching tests with n_p enumerated from 10 to 10000 and β ranging from 0 to 1 for SRSF1 CLIP-seq data. The AUC scores resulted from optimizations using different combinations of these two parameters are presented in 3D heatmaps (Fig. 7A). We observed a much more rapid decrease in prediction power as β becomes lower when n_p is small. In contrast, when β is sufficiently high, the AUC score is not greatly affected by n_p . Thus, the greater n_p and β are, the higher prediction performance RNAMotifModeler can achieve. However, under the consideration of computational efficiency, we have to consider the time consumed in each test (Fig. 7B). The time consumed is exponential to the increment of the number of particles, and is not actually controlled by β . When n_p is 100 and β equals 1.0, RNAMotifModeler achieved a high AUC score of 0.86 within three minutes. These two parameters are then selected for all other optimizations for the SRSF1 dataset used in this study.

Convergence of optimization algorithms used in predicting protein-DNA or protein-RNA binding sites is a common concern due to a number of parameters needed to fit in model. In this report, we proposed a motif searching pathway and a motif searching graph to inspect whether or not the algorithm of RNAMotifModeler indeed has a good convergence regardless of the randomly initialized motif candidates. In the application to SRSF1 consensus motif, the convergence of randomly initialized motif candidates to final targets turned out to be very fast. Thus, for short motifs, we suggest generate such a motif searching graph in order to have a global overview of all possible converged motifs and their possible relationships.

Despite our successful characterization of the binding features of SRSF1 proteins, our future work will be applying RNAMotifModeler to studying specificities of other RNA binding proteins such as fox2, NOVA and EWS, for which high-throughput sequences are already available.

4 Methods

4.1 Predicting RNA base-pairing probabilities

One of the distinct features of RNAMotifModeler is that the information of secondary structures of the RNA regions bound by SRSF1 proteins is incorporated into the motif identification. For each nucleotide in the RNA fragment, we calculate the base pairing probability using the RNAfold function of the Vienna RNA package (version 1.8.5) [2]. The base pairing probability is used since it integrates likelihood of single-strandedness over multiple possible RNA secondary structures. For the CLIP-seq derived RNA fragments, these probabilities are generated based on the base pairing probability of base i being paired with base j , denoted as $p_{i,j}$. The binding probability of base i with all other neighbouring bases, defined as P_i , is calculated by:

$$P_i = \sum_{j=i+1}^{n_i} P_{ij} + \sum_{j=1}^{i-1} P_{ji} \quad (1)$$

where n_s is the length of sequence s . Similar strategies are also used elsewhere [15, 16].

4.2 Modelling protein-RNA binding affinities

In RNAMotifModeler, the consensus of each binding motif is defined by the following components: 1) the reference motif, a k -base RNA sequence on which the protein preferably binds; 2) retained binding affinity despite of a one-nucleotide deviation from reference motif to the sequence of one binding sites. For each k -base motif, there are $3k$ retained binding affinities that describe all the possible deviations from reference motif. For instance, if the i -th base of the reference motif and a specific binding site is m_i and f_i , respectively, the retained binding affinity is defined as μ_{i,m_i,f_i} ; 3) a vector that denotes the optimal base pairing probability of k bases in the motif $\theta=(\theta_i)$; and 4) the penalty for the deviation from the optimal base pairing probability α . All these parameters will be optimized iteratively. A matching score describing the similarity between an RNA fragment (F) and a reference motif (R) is defined:

$$S_{R,F} = \max_{l=1}^{L-k+1} (S_{R,F,l}) \quad (2)$$

where $S_{R,F,l}$ is the binding affinity for l -th binding site:

$$S_{R,F,l} = \prod_{i=1}^k \left((\mu_{i,m_i,f_i}) (1 - \alpha \cdot |\theta_i - P_{f_i}|) \right) \quad (3)$$

where P_{f_i} represents the pairing probability of the i -th nucleotide in the RNA fragment F , calculated in Eq. (1). This matching score integrates the loss of binding affinity caused by both nucleotide and structure deviances from reference motif. We denote the parameter associated to the reference motif R as $\lambda_R = (\mu, \theta, \alpha)_R$, where μ , θ and α represent the $3k$ retained binding affinities, optimal base pairing probability of k bases, and the penalty for the deviation from the optimal base pairing probability, respectively.

4.3 Identify the optimal reference motif from CLIP-seq data

We adopted an iterative approach to identify the optimal reference motif and its associated parameters, using a Quantum Particle Swarm Optimization algorithm (QPSO) [17]. The iterative strategy includes the selection of reference motif R , and optimization of the parameters associated to the reference motif λ_R . The overall procedure includes the following steps:

1. Motif initiation. Randomly select a motif candidate R_{init} from the motif searching space $\mathbf{M}=\{b_1b_2\dots b_k: b_1, b_2, \dots, b_k \in \{A, G, C, U\}\}$ as the reference motif.

2. Parameter optimization. Optimize parameters associated with the reference motif by maximizing its ability for characterizing the CLIP-seq-derived RNA fragments.

Step 2.1. Parameter initiation. We first create n_p particles in the parameter space by randomly selecting numbers from $U(0, 1)$.

Step 2.2. Particle evaluation. For each particle (parameters), we evaluate its capability for distinguishing the CLIP-seq-derived RNA fragment from background sequences. We plot an ROC (Receiver Operating Characteristic) curve by adjusting the matching score threshold, calculated in Eq. (2). The quality of the parameter will be evaluated based on the AUC (area under the curve) of the ROC plot.

Step 2.3. Particle update. Let $\lambda_i^{selfbest}(t)$ and $\lambda^{globalbest}(t)$ be the best individual particle i and the population of particles has met at the t -th iteration. As part of QPSO, each particle must converge to its local attractor λ_i^{pbest} [17]. Compute $\lambda_i^{pbest}(t)$ and the mean of the best positions of all particles λ_i^{mbest} as follows:

$$\lambda_{i,j}^{pbest}(t) = (\varphi_1 \cdot \lambda_{i,j}^{selfbest}(t) + \varphi_2 \cdot \lambda_j^{globalbest}(t)) / (\varphi_1 + \varphi_2) \quad (4)$$

$$\lambda_j^{mbest}(t) = \sum_{i=1}^{n_p} \lambda_{i,j}^{pbest}(t) / n_p \quad (5)$$

where φ_1 and φ_2 are random variables following $U(0, 1)$;

QPSO employs Monte Carlo method to update parameters:

$$\lambda_{i,j}(t+1) = \begin{cases} \lambda_{i,j}^{pbest}(t) - \beta \cdot |\lambda_j^{mbest}(t) - \lambda_{i,j}(t)| \cdot \ln(1/u), & q \geq 0.5 \\ \lambda_{i,j}^{pbest}(t) + \beta \cdot |\lambda_j^{mbest}(t) - \lambda_{i,j}(t)| \cdot \ln(1/u), & q < 0.5 \end{cases} \quad (6)$$

where β is called contraction-expansion coefficient controlling the convergence speed of QPSO; u and q are random variables which also follow $U(0, 1)$.

Repeat Step 2 and Step 3 until $|\lambda^{globalbest}(t+1) - \lambda^{globalbest}(t)| < \varepsilon$ repeatedly, in which ε is a tolerance used here as the stop criterion;

3. Updating reference motifs. Based on the final parameter vector $\lambda^{globalbest}$, the maximal binding affinity of motif candidate K in positive gold standard sequence F is:

$$a_{K,F} = \text{Max}_{\sigma \in \Omega_{K,F}} a_{K,F,\sigma} \quad (7)$$

where $\Omega_{K,F}$ denotes the set of all binding sites for motif K in sequence F ; $a_{K,F,\sigma}$ is also computed by Eq. (3).

In order to update the reference motif, from each positive fragment in the gold standard binding set, we selected the binding site that contributes to the positive selection (genomic loci with the highest binding affinity score). This potential binding site can be either the same as the reference motif, or different due to degeneracy. The reference motif will be further updated to the binding site that can represent largest amount of positive fragments in the gold standard binding set. Let n_F and n_M be the number positive gold standard sequences and the number of motif candidates, respectively. Let $S_{R_{init},F}$ be the maximal binding affinity computed using optimized parameters for the initial reference motif R_{init} in sequence F . To evaluate contribution

of each motif candidate, we define a motif contribution score matrix $\mathbf{c} = [c_{F,K}]_{F=1,2,\dots,n_S, K=1,2,\dots,n_M}$, in which

$$c_{F,K} = \begin{cases} 0, & a_{K,F} \neq S_{R_{init},F} \\ 1, & a_{K,F} = S_{R_{init},F} \end{cases}, \quad (8)$$

and a motif contribution score vector $\mathbf{v} = [v_K]_{K=1,2,\dots,n_M}$, in which:

$$v_K = \sum_{F=1}^{n_S} c_{F,K}. \quad (9)$$

We denote the motif associated with the maximum score in \mathbf{v} as R_{max} . If $R_{max}=R_{init}$, meaning the initialized reference motif accounts for the most contribution to the retained binding affinities, then we stop the iteration; otherwise, let R_{max} be the next R_{init} , and repeat step 2 and 3 until convergence.

4.4 RBP binding motif logo

RNAMotifModeler provides a parameter matrix consisting of retained binding affinity ratios due to sequence mutations and structure alterations at each base. For the ease of visualization, we provide a method to generate a Positional Weight Matrix (PWM). Once *RNAMotifModeler* reaches a convergence, a set of optimal parameters and reference motif will be acquired, as well as a cutoff score of binding affinity at the peak of the accuracy curve. We trace back subsequently to each positive gold standard sequence to identify binding sites with binding affinities higher than the cutoff score. Finally, using these positive binding sites, we calculate the PWM and create a corresponding logo based on the *Weblogo* tool [8].

ACKNOWLEDGMENT

This work is supported by the grant from National Institutes of Health, R21AA017941 (to YL), R01GM085121 (to JRS), and the Indiana Genomics Initiative of Indiana University (supported in part by the Lilly Endowment, Inc.).

5 References

[1] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," *FEBS letters*, vol. 582, pp. 1977-1986, 2008.

[2] I. L. Hofacker, "RNA secondary structure analysis using the Vienna RNA package," in *Curr Protoc Bioinformatics*, 2008/04/23 ed. vol. Chapter 12, 2004, pp. Unit 12 2.

[3] M. Hiller, R. Pudimat, A. Busch, and R. Backofen, "Using RNA secondary structures to guide sequence motif finding towards single-stranded regions," *Nucleic Acids Research*, vol. 34, pp. e117, 2006.

[4] J. R. Sanford, X. Wang, M. Mort, et al., "Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts," *Genome Res*, vol. 19, pp. 381-94, Mar 2009.

[5] E. Buratti, A. F. Muro, M. Giombi, D. Gherbassi, A. Iaconig, and F. E. Baralle, "RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon," *Molecular and cellular biology*, vol. 24, pp. 1387, 2004.

[6] J. R. Sanford, J. Ellis, and J. F. Caceres, "Multiple roles of arginine/serine-rich splicing factors in RNA processing," *Biochemical Society Transactions*, vol. 33, pp. 443-446, 2005.

[7] R. Karni, E. De Stanchina, S. W. Lowe, R. Sinha, D. Mu, and A. R. Krainer, "The gene encoding the splicing factor SF2/ASF is a proto-oncogene," *Nature Structural & Molecular Biology*, vol. 14, pp. 185-193, 2007.

[8] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res*, vol. 14, pp. 1188-90, Jun 2004.

[9] M. Caputi, G. Casari, S. Guenzi, R. Tagliabue, A. Sidoli, C. A. Melo, and F. E. Baralle, "A novel bipartite splicing enhancer modulates the differential processing of the human fibronectin EDA exon," *Nucleic Acids Res*, vol. 22, pp. 1018-22, Mar 25 1994.

[10] J. Ramchatesingh, A. M. Zahler, K. M. Neugebauer, M. B. Roth, and T. A. Cooper, "A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer," *Mol Cell Biol*, vol. 15, pp. 4898-907, Sep 1995.

[11] R. Tacke and J. L. Manley, "The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities," *EMBO J*, vol. 14, pp. 3540-51, Jul 17 1995.

[12] J. C. K. Ngo, K. Giang, S. Chakrabarti, et al., "A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1," *Molecular cell*, vol. 29, pp. 563-576, 2008.

[13] J. C. Hagopian, C. T. Ma, B. R. Meade, C. P. Albuquerque, J. C. K. Ngo, G. Ghosh, P. A. Jennings, X. D. Fu, and J. A. Adams, "Adaptable molecular interactions guide phosphorylation of the SR protein ASF/SF2 by SRPK1," *Journal of molecular biology*, vol. 382, pp. 894-909, 2008.

[14] S. D. Auweter, F. C. Oberstrass, and F. H. T. Allain, "Sequence-specific binding of single-stranded RNA: is there a code for recognition?," *Nucleic Acids Research*, vol. 34, pp. 4943, 2006.

[15] M. Hiller, R. Pudimat, A. Busch, and R. Backofen, "Using RNA secondary structures to guide sequence motif finding towards single-stranded regions," *Nucleic Acids Res*, vol. 34, pp. e117, 2006.

[16] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, pp. 1105-19, May-Jun 1990.

[17] J. Sun, B. Feng, and W. Xu, "Particle swarm optimization with particles having quantum behavior," in *Evolutionary Computation, 2004. CEC2004. Congress on*, 2004, pp. 325-331.

Correlation of Patristic Distance with Nominal Specimen Collection Date in Influenza A/H1N1 Hemagglutinin-Encoding Segments

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA

Abstract

The influenza hemagglutinins are viral coat glycoproteins that facilitate viral binding to the host cell wall; as a result, the virulence of any strain of flu depends significantly on how well the hemagglutinin of that strain promotes that binding. Characterizing the evolution of the hemagglutinins is thus fundamental to predicting the virulence of the virus. Here, I describe a linear regression of patristic distance in Influenza A/H1N1 hemagglutinin-encoding segments on the nominal specimen-collection date contained in the label field of the hemagglutinin genomic sequence descriptors; the regression predicts an average mutation rate of ~2 bp/year (implying, on average, ~0.1 mutations in the hemagglutinin active site per year).

Keywords: Influenza, H1N1, hemagglutinin

1.0 Introduction

The influenza hemagglutinins are viral coat glycoproteins that bind to sialic acid residues on the glycoproteins exposed at the surface of the epithelial cells of the host respiratory system. As a result, the virulence of any strain of flu depends significantly on how well the hemagglutinin of that strain promotes that binding. Characterizing the evolution of the hemagglutinins is thus fundamental to predicting the virulence of the virus.

The influenza A viruses responsible for the pandemic of 1918 were derived from avian viruses, which typically recognize the cell-wall glycan SAa2,3Gal. The hemagglutinins of early isolates from humans infected in these pandemics seem to have recognized SAa2,6Gal in preference to SAa2,3Gal, suggesting that conversion of the avian hemagglutinin to one that can recognize SAa2,6Gal-terminated polysaccharides on host cells is an important step in the generation of pandemic strains. The principal amino acid substitutions involved in this shift of receptor recognition are residues 226 and 228 in the H2 and H3 hemagglutinins (equivalent to residues 222 and 224 in the H5 hemagglutinin). The introduction of these mutations into the H5 hemagglutinin permitted its binding to an a2,6 glycan, although neither change has been found in the hemagglutinins of H5N1 viruses isolated from humans ([13]). A first-principles theory of hemagglutinin evolution is highly desirable but currently beyond the state of the art. First-principles computational methods such as molecular dynamics could provide insight into relevant drug-site free-energetics, but such methods are often computationally expensive and in the case of the hemagglutinins, would require an initial, realistic specification of the *in situ* environment. Relatively few H1N1 hemagglutinin structures are available at present, and none address the effect of the molecules' environment on their active sites. In contrast, phylogenetic comparisons

of the genomic encoding of the hemagglutinins might, by translational proxy, provide insight; some phylogenetic methods, furthermore, are computationally inexpensive. Over 10000 hemagglutinin-encoding (HA) segments of the viral genomes are available for A/H1N1 ([4]).

2.0 Method

The general method of this study has four steps: downloading H1N1 HA segment descriptors, aligning the descriptors, computing the patristic distances among the

segments, and analyzing the correlation of segment patristic distance with segment collection-date. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment, connected by a 1.5 Mbit/s DSL link to the Internet.

Influenza H1N1 HA segments were downloaded from the Influenza Research Database ([4]) on 13 January 2011. The query/download parameters are shown in Figure 1.

Query parameters:

```
Select Segments: 4 (HA)
Subtype: H1N1
Date Range: 1915 to 2011
Geographic Grouping: All
Host: All
Data to Return: Segment/Nucleotide
```

Advanced Options:

```
Display Fields: Sequence Accession, Date
```

```
Display: sort on (increasing) date
```

Download parameters:

```
Select: All segments
Select Download Type: Segment FASTA
Label Sequence By: Custom -- Accession Number, Date
```

Figure 1. Influenza Research Database ([4]) query/download parameters for the Influenza A/H1N1 HA segment descriptors used in this study.

The file resulting from the previous step was edited in *BioEdit* v7.0.5.3 ([6]) to remove any sequences shorter than 1600 bp or longer than 1800 bp, a range chosen by inspection of the sequence descriptors to include some of the descriptors with the earliest collection dates in the set, while excluding descriptors that were 50% shorter or longer than the average descriptor length

in the set. The *BioEdit* navigation for this filtering was

```
Sequence --> Filter Out
Sequences Containing
Certain Characters -->
Delete them --> are <x
[>x] long (x = 1600
[1800]) -->
File --> Save as (type
```



```
= Fasta, filename =
ten.fasta)
```

If fewer than 10 sequences for a given year were in the resulting file, all sequence descriptors for that year were saved. Else, only the first 10 sequence descriptors in each year were saved. (This helps to reduce time bias in the sample, some of which, due to the scarcity of specimens collected before 1930, is unavoidable). The result was a collection of FASTA-formatted sequence descriptors 1600-1800 bp long. *BioEdit* was then used to save the descriptor Labels of this length- filtered set to a separate file.

The "Label" fields in the FASTA-formatted sequence descriptors obtained from the previous step were edited in *Word 2007* so that each had the form "GenBankAccessionID_yyyy", where yyyy is the year contained in the Label. (In this paper, that year is called the "collection date". It should be noted that such a date is merely part of a free-text field; thus, in principle, that "date" could be, and mean, anything. It is relatively common practice, however, for such a date to represent the

date on which the organism from which the sequence was derived was collected.)

The FASTA-formatted sequences from the previous step were aligned using *MAFFT* v6.847b-win32 ([2]), invoked from a *Vista* Command Prompt window. The parameters for the alignment were

```
Order: input
Output format: clustal
Strategy: FFT-NS-i
          (Standard)
Iterative refinement
          (Maximum of 2 iterations)
All other parameters:
          defaulted
```

The resulting CLUSTAL-formatted ([11]) file was edited in *Notepad* to remove blank lines and lines containing asterisks.

A *PAUP* ([8]) neighbor-joining (NJ, [12]) script was built in *Notepad*, incorporating the descriptor labels and aligned sequences obtained in previous steps. The template for the *PAUP* script is shown in Figure 2.

```
#NEXUS
begin taxa;
    dimensions ntax=389;
    taxlabels
    [descriptor labels go here (not shown)]
;
end;

begin characters;
    dimensions nchar=1794;
    format missing=? gap=- matchchar=. interleave datatype=dna;
    matrix
    [aligned data goes here (not shown)]
;
end;

begin paup;
    [1] log start file=H1N1_HA_nj_patdist.log replace;
    [1] nj;
    [3] savedist file=tenpatdist.txt format=oneColumn;
end;
```

Figure 2. Template of PAUP script used to obtain the patristic distances used in this study.

Patristic distances from a 1918 "reference" segment (AF117241 in [4]), and corresponding label-times expressed as years-since-1918, were extracted using the *get_pats* software ([7]) running under *Cygwin* (in turn running under *Vista*) from the patristic distance file produced by

PAUP. The output of *get_pats* is a comma-separated file. This file was converted to a space-separated file using *Notepad*. A linear regression of patristic distance on time was performed by the *Mathematica* ([5]) script shown in Figure 3 ([9]).

```

patdistimedata = ReadList[ToFilename[{"C:",
  "BIOCAMP2011", "Influenza_H1N1_HA"},
  "tenpatdistime.txt"], {Number, Number}];

model=LinearModelFit[patdistimedata,x,x]

model["BestFit"]

Show[ListPlot[patdistimedata, AxesOrigin -> {0,0},
  AxesLabel -> {"Years After 1918", "Patristic Distance from
  AF117247"}], Plot[model["BestFit"], {x, 0, 100}]]

model["ParameterTable"]

model["RSquared"]

model["AdjustedRSquared"]

```

Figure 3. *Mathematica* script used for linear regression in this study.

3.0 Results

10147 sequences were produced by the Influenza Research Database query/download described in Section 2.0.

The length-filtering and time-debiasing steps in *BioEdit* described in Section 2.0 yielded 389 FASTA-formatted sequences.

The *MAFFT* alignment step described in Section 2.0 yielded CLUSTAL-formatted sequence descriptors with 1794 characters per sequence. 388 patristic-distance/time pairs were produced by the patristic-distance/time extraction (via *get-pats*) from

the patristic distance file produced by *PAUP*.

The linear regression computed by *Mathematica* was

$$\text{patristic_distance_from_AF117241} \\ = 0.0621597 + \\ 0.00128348 * \text{Years_Since_1918}$$

A scatterplot and the best linear fit to that data is shown in Figure 4.

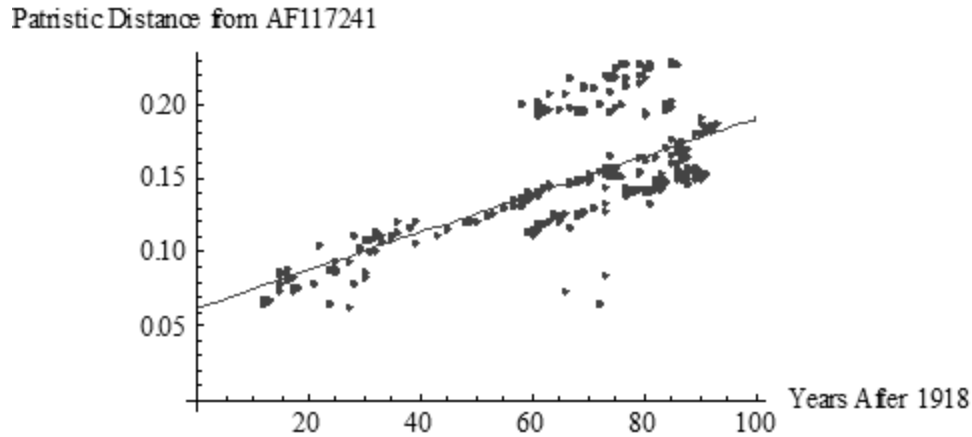


Figure 4. Scatterplot and best linear fit of patristic-distance/time data used in this study.

Some parameter statistics for this regression are:

Parm	Estimate	Standard Error	t Statistic	P-Value
b	0.0621597	0.00437122	14.2202	3.38122×10^{-37}
m	0.00128348	0.0000625213	20.5286	7.74847×10^{-64}

where b is the intercept on the patristic-distance axis, and m is the slope of the regression. The regression coefficient, r^2 , is 0.521937; the adjusted r^2 , 0.520698.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The slope of the regression line suggests that the typical Influenza A/H1N1 HA segment experiences, on average, ~ 0.001 change per year. Since a nominal HA segment has length ~ 1700 bp, we would, based on the regression formula in Section 3.0, expect ($\sim 1700 \text{ bp} \times \sim 0.001$) ~ 2 bp change per year. Such a change would be sufficient to alter at least one amide in the active site of the hemagglutinin encoded by the segment about every 5 years, if we assume the active site is determined by ~ 50 bp and that mutations are uniformly distributed across the molecule. This rate is consistent with

the nominal mutation rate suggested by other considerations ([10]).

In general, we could not expect "collection date" to provide any information about mutation rate. However, if specimens are collected at a rate that is comparable to the mutation rate (as is the case with flu genomic segments), collection dates will tend to exhibit a strong correlation with mutation rates.

2. In contrast to a similar study performed on H1N1 NA segments ([14]), the regression reported in Section 3.0 is relatively small. Inspection of Figure 4 suggests why this is so. Beginning in ~ 1978 , HA segments diverged into three relatively distinct cohorts, two of which were well removed from a linear extrapolation from earlier segments. This sharp change coincides with the beginning of a flu epidemic in swine in the US.

3. The sequence-descriptor sampling protocol described in Section 2.0 is intended to help mitigate time-biasing in the sample by restricting the number of sequence descriptors sampled per year to no more than 10. The results aren't perfect: for some years, [4] contains fewer than 10 (for some years, no) sequence descriptors. Other protocols are of course possible, but the one used in this study is a practical compromise between under-, or over-, sampling any given year, given the data available in [4].

5.0 Acknowledgements

This work benefited from discussions with Town Peterson of the University of Kansas Biodiversity Institute, George Hrabovsky of the Madison Area Science and Technology Institute for Scientific Computing, Tony Pawlicki, and Richard Barrett. For any problems that remain, I am solely responsible.

6.0 References

- [1] Barry JM. *The Great Influenza*. Viking. 2004.
- [2] Katoh K and Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* 9 (1 July 2008), 286-298.
- [3] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.
- [4] Squires B, Macken C, A. García-Sastre A, Godbole S, Noronha J, Hunt V, Chang R, Larsen CN, Klem E, Biersack K, and Scheuermann RH. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. *Nucleic Acids Research* 36 (Database issue), D497-503 (2008).

<http://www.fludb.org/brc/home.do?decorator=influenza>.

- [5] Wolfram Research. *Mathematica Home Edition* v7.0 (2010).
- [6] Hall TA. *BioEdit*: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41 (1999), 95-98.
- [7] Horner JK. *get_pats*, a perl program for extracting patristic distances from a PAUP "one-column" patristic distance file. 2011.
- [8] Swofford D. *Phylogenetic Analysis Using Parsimony (PAUP)* v4.0b10. URL <http://paup.csit.fsu.edu/>. Sinauer Associates. 2004.
- [9] Horner JK. *statpats.nb*, a *Mathematica* notebook for performing linear regression of patristic-distance on time. Available from the author on request.
- [10] Horner JK. An estimate of the mutation rates of the active sites of Influenza A/H5N1 neuraminidases. *Proceedings of the 2010 International Conference on Bioinformatics and Computational Biology*. CSREA Press. 2010. pp. 344-349.
- [11] Higgs DG, Thompson JD, and Gibson TJ. Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* 266 (1996), 383-402.
- [12] Felsenstein J. *Inferring Phylogenies*. Sinauer Associates. 2004.
- [13] Yamada S, Suzuki Y, Suzuki T, Le MQ, Nidom CA, Sakai-Tagawa Y, Muramoto Y, Ito M, Kiso M, Horimoto T, Shinya K, Sawada T, Kiso M, Usui T, Murata T, Lin Y, Hay A, Haire LF. Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type

receptors. *Nature* 444 (16 November 2006), 378-382. doi:10.1038/nature05264.

[14] Horner JK. Correlation of patristic distance with specimen collection date in Influenza A/H1N1 neuraminidase-encoding segments. *Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology*. CSREA Press. 2011. Forthcoming.

Identification of Transcriptional Regulatory Elements by Functional Enrichment Analysis.

Amitava Karmaker^{1*}, Stephen Kwek²

¹University of Wisconsin-Stout, Menomonie, Wisconsin 54751, USA

²Microsoft Corporation, Redmond, Washington 98052, USA

Abstract - Deciphering the complex interaction between transcriptional regulatory (both *trans*- and *cis*-) elements comprehensively and identifying these potential binding sites are fundamental problems in functional genomics. Therefore, determining the transcription factors that regulate a gene in different cell types and the *cis*-regulatory elements they are binding to will help lay the foundation for building gene regulatory networks. While many computational approaches have been developed for lower eukaryotes and prokaryotes, most of them often do not generalize to vertebrates. Here, we use gene ontological evidences to perform functional enrichment analysis among the TFs and genes, and group the functionally related genes to characterize their transcriptional association. We also analyze correlations between TFs and genes using their expression profiles. Thus, we search for putative transcriptional regulatory elements (transcription factor binding sites) along core promoter regions of the grouped genes. The performance of our search is highly satisfactory in term of binding site hit accuracy.

Keywords: Transcriptional Regulatory Elements, Functional Enrichment, Gene Ontology, Gene Expression Profiles, Microarray Analysis.

1 Introduction

With the completion of draft sequencing of genomes of various species (a.k.a. human, mouse, rat, yeast etc.), one of the objectives of functional genomics is to interpret biological significance of the sequences, and to delineate the functional modules along the genomes. Although a large number of genes have been identified, their regulatory mechanism remains mostly unknown at the transcriptional level[1]. To understand the complex interaction of gene regulation comprehensively, we need to identify the regulatory elements in the human genome and comprehend how the genes regulate and interact with each other.

Simply put, the interaction between transcription factor (TF, a.k.a. *trans*-elements) and transcription factor binding sites (TFBS, a.k.a. *cis*-elements) plays a crucial role in controlling gene expression. To modulate transcription and consequently to control the expression of genes, transcription

factor proteins bind to binding sites in the promoter regions and thus either facilitate or inhibit the gene expression. To some extent, the pattern of expression of each gene can be formulated as a function of specific transcription factors, and their binding to the *cis*-elements. So, transcription factors constitute one of the major components in constructing gene regulatory networks. Literally, *trans*-elements can be viewed as “keys” needed to unlock the *cis*-elements which act as “locks”. To comprehend gene transcription mechanism, it is not sufficient to know which keys (*trans*-elements) are needed to lock/unlock a specific gene, but we also need to identify their corresponding locks (*cis*-elements).

Since the human genome sequences are available, quite a number of computational approaches have been developed to discover functional elements in lower prokaryotes by combining genome sequence data and expression profiles[2]. But, due to more degenerate nature and complex interactions of TFs in the multi-cellular mammals (higher eukaryotes), most of the techniques are not able to generalize to mammal genomes. Moreover, these computational techniques are fallible to high false positive prediction rate[3]. In reality, this unusually high false prediction sometimes overwhelms the prospective techniques to deter finding regulatory regions accurately. On the other hand, comparative genome analysis, which is a biologically more relevant approach, provides a powerful way to search for similarities across the species at the sequence level and consequently to assign functional annotations[4]. Besides this, it is assumed that genes with similar functions are most likely to be regulated through the same mechanisms[5]. Thus, we can infer transcriptional sub-networks based on functional enrichment of genes.

In this paper, we propose a systematic technique to identify putative transcriptional regulatory elements in human genome by functional enrichment of genes using ontology. Our hypothesis is inspired by the axiomatic supposition that genes that are in the same functional complex and located in closer cellular proximity are often regulated by the same transcription factors[6]. In fact, two proteins, sharing same molecular function in alike biological process and residing in close physical location, are more likely to interact with each other[7]. Therefore, clustering the genes set using functional enrichment allows us search for *cis*-modules along the

*Corresponding author

promoter regions of the genes more efficiently. Initially, we analyze the correlations among the genes and corresponding TFs using microarray expression data. Besides this, we used the popular gene ontology to come up with the enrichment analysis of the genes. In fact, functional enrichment analysis complements the findings for correlations from expression profiles. To evaluate the efficacy of our approach, we validated our prediction for the transcription factor binding sites from functionally enriched gene clusters by comparing with TRANSFAC[8].

2 Related works

In silico discovery[9] of binding sites is quite effective for prokaryotes, like *Escherichia coli*[10], where genomes are more compact with many genes being regulated by a single operon, is relatively easy to locate. Similar successes have been reported for simple unicellular eukaryotes, like *Saccharomyces cerevisiae*[2]. The main approach for finding *cis*-elements of such simple organisms is to find overrepresented motifs modeled by known background profiles, such as position weighted matrices (PWMs)[11], position specific score matrices (PSSMs)[12], while some use clustering to demarcate *cis*-regulatory modules[13, 14].

For higher multi-cellular eukaryotes, model-based approaches[1, 15] that discover patterns among co-expressed genes with respect to regulating transcription factors have been proposed. The idea behind these techniques involves the proximity of common *cis*-regulatory modules among the co-expressed genes. Among other common model-based (a.k.a. machine learning) techniques, artificial neural networks[16], greedy algorithm[17], Gibbs Sampling[18], Markov chains[19], Expectation Maximization (EM) algorithm[20] are widely used for eukaryotes. However, it has been reported that these model-prediction techniques are susceptible to high false positive prediction rate and majority of predicted TFBS generated with predictive models (*in silico*) have no functional role *in vivo* [21].

Jin et al.[22] analyzed conserved human-mouse orthologous gene pairs to find core promoter elements and Bussemaker et al.[23] addressed the issue of detecting regulatory elements using correlation of expressions. A recent paper by Kim et al.[24] dealt with predicting transcriptional regulatory elements of human promoters using gene expression and promoter analysis data, which compare two pools of genes using z-scores.

3 Methods and materials

3.1 Data preprocessing

We collected publicly available microarray data of normal human tissues[25], which provide us with 26,260 unique genes from 35 different organs. In total, the data set consists of 115 tissue specimens. For each experimental tissue sample, Cy5- and Cy3- labeled samples were co-hybridized to a cDNA microarray containing 39,711 human cDNA's, representing 26,260 different genes [26]. Expression ratios

were globally normalized by mean-centering each gene across all arrays.

3.2 Calculation of correlation co-efficient

If a transcription factor does regulate a gene, according to reported results[15] in the literature, it is expected that they are linearly correlated. However, we observed that very often there seems to be a saturation point where the effect on the expression level of the gene diminishes as the level of transcription factor continues to increase and may reach a plateau or even decrease in some cases. Thus, instead of using simple linear correlation, we measure the correlation using Equation 1 as our regression curve.

$$y' = ye^{\alpha y}, \quad (1)$$

Where α is an exponential constant

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (2)$$

Where, n is the sample size, and x_i and y_i are the sum of \mathbf{X} and \mathbf{Y} from $i = 1$ to n

In Equation (1), y is the original expression level, and it is multiplied by some exponential constant to generate new values. The value of parameter α was set to 0.25. This correlation coefficient is more general than simple linear correlation coefficient. By setting $\alpha = 0.0$, we end up with the simple linear correlation coefficient. We calculated Pearson's Correlation Coefficient (Equation (2)) of all pairs of gene and TF. The correlation coefficients indicate how tightly genes are up-regulated and down-regulated with respect to transcription factors. The values of Pearson's Correlation Coefficient range from -1 to +1. Any value in positive scale indicates increasing correlation, with +1 being perfectly linear correlated and negative values denote the case of a negative correlation. Any value in between in all other cases represents the degree of dependency between the variables (i.e. gene and TF pair).

3.3 Gene Ontology

Genome-wide comparison has revealed that a large fraction of genes encoding the core biological processes and molecular functions are shared by all the eukaryotes, with a few exceptions[27]. In fact, comprehensive knowledge about biological roles of common gene products in diverse species can obviously explain, and often provide strong implication of, its function in the like genomes. However, due to divergent nomenclatures and interpretations of biological elements, it has been difficult for the researchers to talk in common language. To address this issue, the Gene Ontology (GO) Consortium[28] has been formed. Basically, Gene Ontology (GO) provides a great resource for describing gene products by standardizing biological concepts and by

consolidating gene annotation information from heterogeneous data sources in a consistent manner. As a mainstay standard for facilitating annotation of gene products, it has been successfully used in unraveling protein-protein interactions and classifications in genomes, such as *Homo sapiens*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Arabidopsis thaliana*.

Gene Ontology (GO) Consortium has developed a database consisting of standardized, structured, dynamically controlled vocabularies (ontological) to encode various aspects of gene products in organisms[28]. The Gene Ontology (GO) is categorized into three orthogonal entities: (1) molecular function (MF) describes the role of a gene product in molecular level; (2) biological process (BP) outlines the processes (objectives) the gene products partake in; (3) cellular component (CC) refers to the cellular localization of the proteins where they are active. Each GO is represented as a directed acyclic graph (DAG), in which each term is either a child of one or multiple parents ("is-a" relationship) or a constituent instance ("part-of" relationship) of the parent terms. In the graph, the nodes correspond to the GO terms, while edges denote the relationships among the terms. Depending on the depth (level) of a node, we can determine the specificity of the term. The closer to the root a term is, the more general the term is. Conversely, if it is located in the leaf levels, the term is the most specific with respect to that particular ontology.

3.4 Functional enrichment measure

Although semantic similarity based methods are popular in assessing functional similarity among the gene products, there are a number of drawbacks we need to consider. First of all, different methods treat the commonality (a.k.a. generality and specificity) of nearest common ancestors in different ways. Secondly, in the GO graph, the depth of terms does not actually signify the specificity of the corresponding concepts. Different terms in the same rank (depth) are necessarily not equally specific. Finally, as the GO is a continuing project where new vocabularies are constantly added (updated), therefore very often the similarity measures are subject to change.

Regarding all these issues, we attempt to define a similarity metric based on assigned GO terms to a gene product instead of concerning much about frequently changing GO semantic structure. Again, as we are interested in clustering functionally related genes on the basis of their GO terms, our distance measure provides straightforward approach to group them together. The idea behind our metric definition is that the more genes have common (general) GO terms, and the less they have specific GO terms, the more likely they tend to be functionally related. Our distance measure is based on the Czekanowski-Dice formula (see Equation 3).

Let two sets of GO terms of annotated genes $G1$ and $G2$ be $GO_1 = \{go_{11}, go_{12}, go_{13}, \dots, go_{1m}\}$ and

$GO_2 = \{go_{21}, go_{22}, go_{23}, \dots, go_{2n}\}$ in order.

According to our algorithm, the distance measure between $G1$ and $G2$:

$$D(G1, G2) = \frac{|GO_1 - GO_2| + |GO_2 - GO_1|}{|GO_1 \cup GO_2| + |GO_1 \cap GO_2|} \quad (3)$$

The closer the genes are in respect with biological function, the lesser the distance measure ($D(G1, G2)$) is. In our analysis, we label this distance measure as functional enrichment score. This distance formula weighs more on the significance of the common GO terms by giving more emphasis to similarities than to dissimilarities. Thus, if two gene products do not share any GO terms, the distance value would be one (1), the highest possible value, while for two gene products sharing exactly the identical set of GO terms, the distance value is zero (0), which is the lowest possible value.

3.5 Finding *cis*-regulatory elements


To determine the (putative) *cis*-regulatory elements, we identify associated genes with certain TF with correlation co-efficient greater than a threshold (>0.5). Using functional enrichment analysis, we construct cluster of genes that are functionally related to certain transcription factor. After calculating the distance measures (functional enrichment scores) of the respective TF against rest of the genes, we sort them by enrichment score in ascending order (genes with less score at the top). For further analysis, we selected top ten genes from this list, which include genes that are functionally enriched with corresponding TF (enrichment score < 1.0) with moderately high correlation coefficient (~ 0.60).


Transcriptional regulatory elements are found either upstream or downstream of genes, scattered all along thousands of bps in both intergenic and intragenic regions. However, most TFBS predictors tend to focus in the proximal promoter region[3] because the difficulty of TFBS prediction tends to increase with the size of the region of interest. Besides, increasing the region of interest upstream of the transcription start site to more than a few thousand base pairs increases the chances of falsely identifying common repeat elements. This, we focus on the core promoter regions from 1500 bps upstream to 500 bps downstream (-1500 to +500, total 2000 bps) and extracted the nucleotide sequences for the genes as FASTA format.

To ensure that our putative TF binding sites are of high quality, we validated them with TRANSFAC database[15], which is the largest repository for experimentally derived (validated) TFBS. We also performed further corroboration of our putative sites using P-Match[29]-public (which is a TRANSFAC subsidiary) and ConSite[30], which combines pattern matching and weight matrix approaches thus providing higher accuracy of recognition than each of the methods alone. To reduce false-positive validation using P-match, we chose "high quality vertebrate matrices only" as our default option. We obtained the report for all pre-selected

genes, setting cut-off selection for matrices to minimize (1) false-positive, (2) false-negative, and (3) the sum of both error rates. Moreover, ConSite[30] is a user-friendly, web-based tool for finding cis-regulatory elements in genomic sequences using high-quality transcription factor models and cross-species comparison filtering.

Table 1: The list of identified binding sites for *E2F5* and *RELB* TFs. Results were validated using both TRANSFAC and ConSite.

<i>E2F5</i> (TRANSFAC: E2F, ConSite: E2F)				
				
Genes	Correlation Coefficient	Functional enrichment score	Position in sequence (strand)	Consensus sequence
<i>MBD4</i>	0.82118	0.76	942 (+)	TTTGcgc
<i>DCK</i>	0.79218	0.904	1496 (-)	gcgCCAAA
<i>MCM6</i>	0.78034	0.629	1347 (+)	TTTGGcgc
<i>MYBL1</i>	0.76635	0.538	N/A	N/A
<i>DR1</i>	0.76331	0.578	N/A	N/A
<i>LSM6</i>	0.75098	0.739	1755 (-)	ccgCGAAA
<i>EZH2</i>	0.74767	0.583	1533 (+)	TTTGGcgc
<i>PCNA</i>	0.73964	0.769	1442 (-)	gcgGGAAA
<i>HMGB2</i>	0.69681	0.75	336 (+)	TTTGGcgc
<i>NMI</i>	0.61465	0.733	1553 (+)	TTTCGcgg

<i>RELB</i> (TRANSFAC: c-REL, ConSite: c-Rel)				
				
Genes	Correlation Coefficient	Functional enrichment score	Position in sequence (strand)	Consensus sequence
<i>PSMB9</i>	0.91903	0.833	1107 (-)	GGAAAgctcc
<i>COX7B</i>	0.80250	0.76	N/A	N/A
<i>ZFP106</i>	0.76718	0.913	1343 (-)	GGAATcctca
<i>ARHGAP5</i>	0.76682	0.909	1884 (+)	gggtgCTTTC
<i>NFE2L1</i>	0.74318	0.619	641 (-)	GAAACatccc
<i>MAPKAPK3</i>	0.73573	0.904	197 (-)	TGTAGcacc
<i>RYR2</i>	0.72470	0.8	549 (-)	GGAATgctcg
<i>DNAJB6</i>	0.71556	0.809	137 (+)	gggatTTTTC
<i>ARF1</i>	0.71359	0.933	256 (+)	ggggcTTTCC
<i>IRF2</i>	0.70548	0.474	1468 (+)	ggggaTTTCC

4 Results and Discussion

As a case study, we selected *E2F5* and *RELB* for our candidate TF. We screened out genes that are functionally enriched with these TFs. In order to quantify the regulatory elements along these gene sequences, the core promoter regions (see Methods) were fed to P-Match[29] using all three available options for handling false discoveries. Basically, the output with option “minimizing false negative” considers merely minimal number of base pairs match and calls it a hit. Thus it improves its recall numbers (maximize loose-bound relevance at the cost of precision), with a huge list of *cis*-element candidates. We expect the false-positive rate to be extremely high for the predictions to be meaningful. Therefore, we did not discard this option. Among the other options, “minimize false positive” tries to find exact (~100%) PWM match and accounts for the most precise TF hits. The other option “minimize sum of both error rates” seems to take advantage from the best of both worlds (keeping balance on both recall and precision) and evens out high false discovery rates. To ensure better quality of our analysis, we considered only the option “minimize false positive”, which maximizes the precision values without compromising too much with recall values. We summarize the sample results for *E2F5* and *RELB* genes in Table 1. The results for consulting ConSite are furnished as well. The consensus sequences (Logo-plots[31]) for respective TFBS were extracted from TFM-Explorer[11].

Our predictions for *cis*-elements for these two TFs are highly accurate. Out of the ten human genes that are associated with *E2F5* (E2F transcription factor 5), a member of *E2F* TF family, eight genes (80% hit rate) carry the supposed binding sites precisely (negative strands are give as reverse complemented. Comparing the sequence patterns of binding sites, we can say that almost all of them share the consensus sequence ‘**TTTSSCGC**’ where S could be a C or G. Likewise, for the ten human genes functionally correlated with TF *RELB*, we have found nine genes have the consensus sequence for *RELB* binding sites, which achieves a hit rate of ~90%. Here, we found “**TTTCC**” as sense (+), or “**GGAAA**” as anti-sense (-) complementary, to be common motif with a number of out of pattern nucleotides around.

5 Conclusions

In short, we propose a computational method to identify putative transcriptional regulatory elements by analyzing functional enrichment using gene ontology. Although there are a lot of computational techniques for this purpose, it is not possible to extend those from motif finding in lower prokaryotes to that in mammals. These techniques also tend to show higher false discovery rates. We demonstrated that the use of our similarity (distance) metric can group genes based on enrichment score and it strengthens the findings from gene expression profile analysis. In each group genes are functionally related to the corresponding TFs; so searching for functional modules along the promoters of genes is more appropriate for capturing possible regulatory relationship. Finally, we validate our prediction for *cis*-

regulatory motifs in both genomes using TRANSFAC. As a possible further step to confirm the regulatory relationships, the TF-gene pairs and their functional enrichment constructed here may serve as a reference of additional evidence for ChIP-chip results.

6 References

- [1] J. W. Fickett and W. W. Wasserman, "Discovery and modeling of transcriptional regulatory regions," *Curr Opin Biotechnol*, vol. 11, pp. 19-24, Feb 2000.
- [2] M. Tompa, *et al.*, "Assessing computational tools for the discovery of transcription factor binding sites," *Nat Biotechnol*, vol. 23, pp. 137-44, Jan 2005.
- [3] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nat Rev Genet*, vol. 5, pp. 276-87, Apr 2004.
- [4] R. Stevens, *et al.*, "Ontology-based knowledge representation for bioinformatics," *Brief Bioinform*, vol. 1, pp. 398-414, Nov 2000.
- [5] D. J. Allocco, *et al.*, "Quantifying the relationship between co-expression, co-regulation and gene function," *BMC Bioinformatics*, vol. 5, p. 18, Feb 25 2004.
- [6] W. K. Huh, *et al.*, "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, pp. 686-91, Oct 16 2003.
- [7] X. Wu, *et al.*, "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations," *Nucleic Acids Res*, vol. 34, pp. 2137-50, 2006.
- [8] V. Matys, *et al.*, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, pp. 374-8, Jan 1 2003.
- [9] G. Pavesi, *et al.*, "In silico representation and discovery of transcription factor binding sites," *Brief Bioinform*, vol. 5, pp. 217-36, Sep 2004.
- [10] L. A. McCue, *et al.*, "Factors influencing the identification of transcription factor binding sites by cross-species comparison," *Genome Res*, vol. 12, pp. 1523-32, Oct 2002.
- [11] M. Defrance and H. Touzet, "Predicting transcription factor binding sites using local over-representation and comparative genomics," *BMC Bioinformatics*, vol. 7, p. 396, 2006.
- [12] P. E. Boardman, *et al.*, "SiteSeer: Visualisation and analysis of transcription factor binding sites in nucleotide sequences," *Nucleic Acids Res*, vol. 31, pp. 3572-5, Jul 1 2003.
- [13] M. C. Frith, *et al.*, "Cluster-Buster: Finding dense clusters of motifs in DNA sequences," *Nucleic Acids Res*, vol. 31, pp. 3666-8, Jul 1 2003.
- [14] N. Rajewsky, *et al.*, "Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo," *BMC Bioinformatics*, vol. 3, p. 30, Oct 24 2002.

- [15] E. M. Conlon, *et al.*, "Integrating regulatory motif discovery and genome-wide expression analysis," *Proc Natl Acad Sci U S A*, vol. 100, pp. 3339-44, Mar 18 2003.
- [16] C. T. Workman and G. D. Stormo, "ANN-Spec: a method for discovering transcription factor binding sites with improved specificity," *Pac Symp Biocomput*, pp. 467-78, 2000.
- [17] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, pp. 563-77, Jul-Aug 1999.
- [18] M. C. Frith, *et al.*, "Finding functional sequence elements by multiple local alignment," *Nucleic Acids Res*, vol. 32, pp. 189-200, 2004.
- [19] A. V. Favorov, *et al.*, "A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length," *Bioinformatics*, vol. 21, pp. 2240-5, May 15 2005.
- [20] W. Ao, *et al.*, "Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR," *Science*, vol. 305, pp. 1743-6, Sep 17 2004.
- [21] W. B. Alkema, *et al.*, "MSCAN: identification of functional clusters of transcription factor binding sites," *Nucleic Acids Res*, vol. 32, pp. W195-8, Jul 1 2004.
- [22] V. X. Jin, *et al.*, "Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs," *BMC Bioinformatics*, vol. 7, p. 114, 2006.
- [23] H. J. Bussemaker, *et al.*, "Regulatory element detection using correlation with expression," *Nat Genet*, vol. 27, pp. 167-71, Feb 2001.
- [24] S. Y. Kim and Y. Kim, "Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data," *BMC Bioinformatics*, vol. 7, p. 330, 2006.
- [25] R. Shyamsundar, *et al.*, "A DNA microarray survey of gene expression in normal human tissues," *Genome Biol*, vol. 6, p. R22, 2005.
- [26] D. J. Maron, *et al.*, "Gene therapy of metastatic disease: progress and prospects," *Surg Oncol Clin N Am*, vol. 10, pp. 449-60, xi, Apr 2001.
- [27] D. Devos and A. Valencia, "Intrinsic errors in genome annotation," *Trends Genet*, vol. 17, pp. 429-31, Aug 2001.
- [28] M. Ashburner, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, May 2000.
- [29] D. S. Chekmenev, *et al.*, "P-Match: transcription factor binding site search by combining patterns and weight matrices," *Nucleic Acids Res*, vol. 33, pp. W432-7, Jul 1 2005.
- [30] A. Sandelin, *et al.*, "ConSite: web-based prediction of regulatory elements using cross-species comparison," *Nucleic Acids Res*, vol. 32, pp. W249-52, Jul 1 2004.
- [31] J. L. DeRisi, *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680-6, Oct 24 1997.
- [32] T. L. Ferea, *et al.*, "Systematic changes in gene expression patterns following adaptive evolution in yeast," *Proc Natl Acad Sci U S A*, vol. 96, pp. 9721-6, Aug 17 1999.
- [33] N. Ogawa, *et al.*, "New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis," *Mol Biol Cell*, vol. 11, pp. 4309-21, Dec 2000.
- [34] P. T. Spellman, *et al.*, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol Biol Cell*, vol. 9, pp. 3273-97, Dec 1998.
- [35] A. P. Dempster, *et al.*, "Maximum-likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. B39, pp. 1-38, 1977.
- [36] M. Ouyang, *et al.*, "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics*, vol. 20, pp. 917-23, Apr 12 2004.
- [37] E. Frank, *et al.*, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, pp. 2479-81, Oct 12 2004.
- [38] R. Jornsten, *et al.*, "DNA microarray data imputation and significance analysis of differential expression," *Bioinformatics*, vol. 21, pp. 4155-61, Nov 15 2005.
- [39] R. Jornsten, *et al.*, "A meta-data based method for DNA microarray imputation," *BMC Bioinformatics*, vol. 8, p. 109, 2007.

An integrated pipeline for protein classification using specific PSSMs and existing protein annotations

Kyung Dae Ko¹ and Hongfang Liu²

¹Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University

²Department of Health Sciences Research, Mayo Clinic College of Medicine

Abstract - Protein classification has been performed by many protein databases to infer annotations of unknown proteins and therefore enhance the performance of protein annotation. In this study, we implemented an integrated pipeline for protein classification using specific PSSMs and proteins with the same entity name. After clustering sequences on the basis of their evolutionary distances, a target group is selected using Jaccard distance. Finally, each group is represented using specific PSSMs generated from sequences in the target group. Using 76 p53-relative and 155 non-relative sequences to validate the performance of our pipeline, we measured 100% accuracy of protein classification by our pipeline. In addition, we identified 35 homologous proteins of p53 among 86,718 sequences through high-throughput analysis of human proteome.

Keywords: PSSM, protein classification, text mining, Jaccard distance, p53

1 Introduction

There is a high demand of automated protein annotation approaches and methods due to the latest advance in high throughput genomics and proteomics technology. However, automated protein annotation is a very challenging task in computational biology. In general, the first step in annotating a novel protein is to identify homologous proteins related to the protein. If its homologous proteins are well annotated, we can infer the characteristics of the protein from the homologous proteins' annotations.

One of the simplest methods to identify homologous proteins is to measure the similarity between novel and reference sequences [1, 2]. If their identity is high, they can be structural and/or functional homologous. However, for sequences that are distantly related, sequence-sequence comparison algorithms may lose the sensitivity in detecting the homologous relationship [3]. To increase the sensitivity in detecting remote homologues, instead of comparing two proteins directly through pair-wise sequence alignment, the new sequence can be compared with profiles, which contain common information from known protein sequences belonging to the same families. Indeed, after building multiple sequence alignments of related sequences in the same family, a PSSM (Position Specific Scoring Matrice) or HMM (Hidden Markov Model) model is then generated on the basis of the common information from the alignments. Using PSSM

or HMM, sequence-profile comparison methods such as PSI-BLAST (Position specific iterative-BLAST) and SAM (Sequence Alignment and Modeling System) can increase the sensitivity in detecting the distant homologous sequences with low sequence identities [4, 5]. In addition, the sensitivity and specificity of PSSM or HMM tend to depend on sequences used for building multiple sequence alignments. Thus, specific PSSMs generated from functional related sequences can improve the sensitivity of protein classification.

In this study, we implemented an integrated pipeline for protein classification using specific PSSMs and considering proteins with the same name based on the observation that biologists tend to assign related genes or proteins similar names. Sequences are clustered on the basis of their evolutionary distances. After selecting a target group using Jaccard distance, specific PSSMs are generated from sequences in the target group. Finally, each group is represented using specific PSSMs.

In next section, we describe the background information of tools and resources used in the pipeline. We will then introduce our classification pipeline. A case study based on p53 (tumor suppressor protein) is provided in detail.

2 Method and Resources

2.1 Tools and Resources

The tools in this study contain PSSM and RPS-BLAST (Reverse Position specific iterative-BLAST). A PSSM profile is a position-specific scoring matrix with 21 columns and M rows where M is the length of probe. Each row matches a sequence position of the probe [6]. The first 20 columns in each row show the score for searching each of 20 amino acid residues at the specific position of the target sequence. A penalty for insertions or deletions (INDELs) at each position of the target sequence is encoded in the 21st column. When a target sequence is compared with PSSMs, the highest score or scores above a specified threshold are retained as outputs [6]. RPS-BLAST searches homologous sequences in the inverse way of PSI-BLAST [7]. Thus, it reverses the role of a sequence and PSSMs, comparing a query sequence against a library of position-specific scoring matrices (PSSMs).

The resources used in the study include UniProtKB, a comprehensive knowledgebase about protein sequences and functional information, BioThesaurus, a comprehensive collection of gene/protein names collected from over 30 molecular databases for UniProtKB records, and several

gene/protein family classification and functional annotation knowledge bases including PANTHER, PIRSF, and Gene Ontology. The following summarizes them.

UniProtKB provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information [8, 9]. It consists of a manually annotated and reviewed component, Swiss-Prot, and an automatically annotated component, TrEMBL. Proteins with sequence similarities of 50% or 90% were grouped into UniREF50 and UniREF90 clusters [10].

BioThesaurus is a thesaurus aiming to provide a comprehensive collection of protein and gene names for protein records in the UniProtKB. Currently covering six million proteins, the latest version of BioThesaurus consists of over eight million names extracted from multiple molecular biological databases according to the database cross-references in UniProtKB and iProClass [11].

The **PANTHER (Protein ANalysis THrough Evolutionary Relationships Classification System)** is a resource that classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence [12]. Proteins are classified by expert biologists into families and subfamilies of shared function, which are then categorized by GO terms.

The **PIRSF (Protein superfamily classification system)** is a protein classification system based on the domain information of the whole proteins. It provides comprehensive and non-overlapping clustering of UniProtKB sequences into a hierarchical order to reflect their evolutionary relationships [13].

Gene Ontology (GO) presents a structured vocabulary about biological roles of gene and proteins from different species [14]. GO defines three different parts including molecular function, biological process and cellular component. GO terms are organized in directed acyclic graphs (DAG) whose nodes have child-parent relationships [14].

PHYLIP (Phylogeny Inference Package) is a package of programs for inference of phylogenies from sequences. Data types of the package include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters. Methods in the package are to generate distance matrix and consensus trees, and calculate bootstrapping, parsimony, and likelihood [15].

2.2 Method

Figure 1 shows the pipeline that consists of three modules. The first module is to collect sequences from public databases based on names collected in BioThesaurus, then calculate evolutionary distances among sequences, and finally cluster proteins in groups on the basis of their evolutionary distances.

The second module is to characterize clustered groups by measuring the dissimilarity between the groups and reference protein families in PIRSF and PANTHER. After

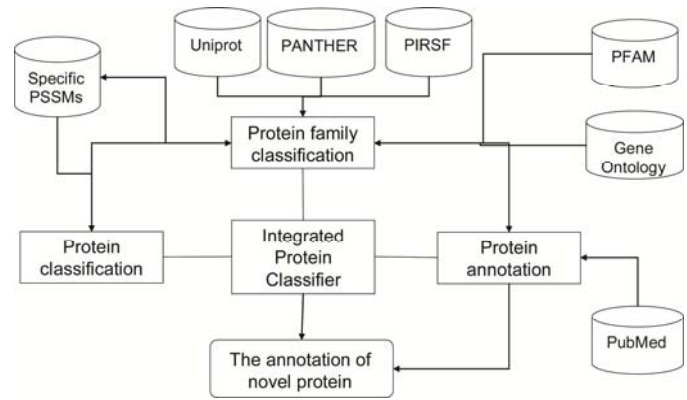


Figure 1: Diagram showing the workflow of the pipeline for protein classification

calculating the relative frequencies of domain architectures and Gene ontology terms which each protein family has, we use weighted Jaccard distance to measure their dissimilarity. Jaccard distance is generally used to measure dissimilarity between sample sets [16], and is calculated by subtracting the Jaccard coefficient from 1 in equation (1) and (2). Then, we give a relative frequency weight to Jaccard distance for reflecting the number of domain architectures or GO terms. Since the sum of the relative frequencies of domain architectures or GO terms is 1 in a protein family, we assume that the probabilities that protein family C_1 and C_2 have the same domain architecture or GO terms are $P(C_1)$ and $P(C_2)$. Then, the weight is defined in equation (3) assuming independency and mutual exclusiveness. We finally define weighted Jaccard distance in equation (4) :

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \quad (1)$$

$$J_d(C_1, C_2) = 1 - J(C_1, C_2) \quad (2)$$

$$E_J = \frac{P(C_1) + P(C_2) - P(C_1) \times P(C_2)}{P(C_1) + P(C_2)} \quad (3)$$

$$J_M = E_J \times J_d \quad (4)$$

The third module is to identify the characteristics of a protein using specific PSSMs. After selecting a target group for the classification, we generate specific PSSMs using sequences in the group. PSSM generally describes the distribution of residues at each position in a conserved pattern such as motif or domain. Thus, if we generate specific PSSMs using sequences in a specific group, the specific PSSMs can allow us to identify proteins whose functional characteristics are similar to the specific group in novel proteins or proteomes.

Based on this assumption, a pipeline first generates PSSMs from sequences which have similar domain architectures and functional GO terms. Second, each query sequence is searched against the specific PSSMs using RPS-

BLAST. If the alignments returned from the search do not satisfy our e-value threshold, they are filtered out. Then, given the alignments to specific PSSMs, a residue score is calculated. For every alignment returned from the RPS-BLAST search of each query against specific PSSMs, each amino acid of a query which is identically or positively (identical, but conserved) aligned is scored with BLOSUM62 score for the aligned pairs. These scores are summed for each amino acid of the query (i.e., residue score). The specific score for a query protein is calculated using equation (5).

$$\frac{1}{n} \sum_{i=1}^n p_i \text{ if } p_i > 0 \quad (5)$$

where n is the length of a protein sequence and p_i is a positional score of i^{th} amino acid of the protein.

3 Results

To validate the performance of our pipeline, we first selected p53 tumor suppressing protein as a key word. Using BioThesaurus, we collected 205 sequences, which have the entity name as “p53”, and 3204 sequences, whose sequence similarities are over 50%, from UniProtKB, based on UniREF50. We calculated evolutionary distances among these sequences using phylib library and clustered them into 38 groups.

Table 1. The weighted Jaccard distances of domain architecture and functional GO term between group1 and protein families in PIRSF.

	PIRSF002089	PIRSF025230	PIRSF031080
Domain architecture	0.5555	0.6545	0.8889
Functional GO term	0.1935	1	1

Calculating the weighted Jaccard distances of these groups against protein families in PIRSF, the weighted Jaccard distance of the biggest group is very close to PIRSF002089 (tumor suppressor p53) in Table 1. Among 111 sequences in the group, we selected 35 reviewed sequences for the generation of specific PSSMs. We then chose 76 sequences as a positive dataset, and 26 RRM (Rna Recognition Motifs) and 127 non-nucleic binding proteins as a negative dataset.

Shown in Figure 2, the specific PSSMs successfully identified the conserved patterns related to p53 in a sequence of testing dataset. X-axis represents the position of amino acid, and Y-axis represents residue score. Since Figure 2 (b) shows only the distribution of residue scores, we filtered residue scores using smoothing filter for the identification of conserved regions. Shown in Figure 2 (c), the conserved regions match the domain regions identified by BLAST. This indicates that our pipeline can predict the conserved regions such as domains or motifs in a protein sequence using specific PSSMs.

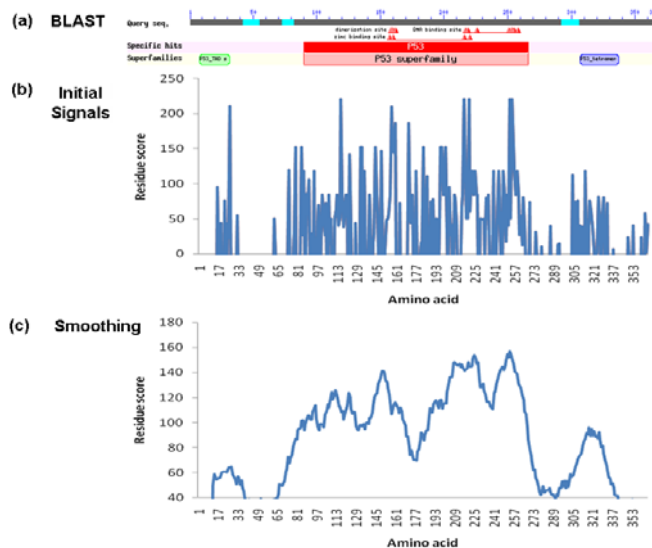


Figure 2: The identification of conserved regions in a protein sequence using specific PSSMs. (a) domain regions predicted by BLAST (b) the distribution of residue scores (c) the conserved regions predicted by the new pipeline using specific PSSMs.

To test the accuracy of the prediction, we selected 76 sequences as a positive dataset and 155 sequences (RRM: 37, non-nucleic binding protein: 127) as a negative dataset. Then, we calculated sensitivity, specificity, and accuracy using equation (5), (6), and (7). The pipeline did not identify any conserved region in proteins not related to p53 proteins, the sensitivity, specificity, and accuracy of the pipeline are 100%.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \quad (6)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (7)$$

Table 2. The sensitivity, specificity, and accuracy of positive and negative datasets about p53 proteins.

TP	FP	TN	FN	Sensitivity (%)	Specificity (%)	Accuracy (%)
76	0	155	0	100	100	100

For further validation, we identified p53 related proteins in human proteome. In fact, after downloading 86,718 proteins from International Protein Index (IPI) site, we did high-throughput analysis of these proteins using specific PSSM for p53 proteins. Among 86,718 sequences, we identified 12 of p53, 13 of p63, and 10 of p73 proteins.

Even though we used specific PSSMs for p53 proteins, our pipeline identified tumor-related proteins including p63 and p73 proteins in the proteomic analysis. In 2009, Dr. Vladimir’s group proved that they are evolutionary close to each other and they have very similar structures [17]. Because of that, our pipeline captured all of p53, p63, and p73 proteins

in human proteome. Therefore, these two experiments suggest that, generating functional specific PSSMs for sequences with similar functional characteristics is able to identify new proteins that have similar characteristics.

4 Conclusions

Many protein databases use homology-based approaches to build protein families and improve their protein annotations. While these protein families provide important resources for biologists to predict structures and functions of novel proteins, it is not clear how well those protein families capture the characteristics of proteins. Generally, we use sequence similarity, domains (or domain architectures), and GO terms to annotate proteins. Since protein families are used to infer protein annotations, proteins from the same family should tend to share similar GO terms and domain architectures. The names of biological entities related to these proteins can also be shared.

Based on the above, we add reliable information related to the characteristics of protein families into a pipeline for protein classification. As we use sequences which are collected on the basis of similar domain architectures and functional GO terms for specific PSSMs, these specific PSSMs allow RPS-BLAST to identify proteins which have similar characteristics in human proteome with high accuracy. Thus, this study suggests that additional information such as the entity name, evolutionary distance, domain architecture, and functional GO terms besides sequence similarity is helpful in improving protein classification. Finally, the integration of different methods in different fields into one pipeline can be cornerstone to implement a unified protein classifier.

5 Acknowledgement

This study was supported by National Science Foundation ABI:0845523 and National Institute of Health R01LM009959A1.

6 References

- [1] Sander, C. and R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 1991. **9**(1): p. 56-68.
- [2] Hilbert, M., G. Bohm, and R. Jaenicke, Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins: Struct. Funct. Genet.*, 1993. **17**: p. 138-151.
- [3] Rychlewski, L., et al., Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci*, 2000. **9**(2): p. 232-241.
- [4] Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 1997. **25**(17): p. 3389-3402.
- [5] Karplus, K., SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res*, 2009. **37**(Web Server issue): p. W492-W497.
- [6] Gribskov, M., A.D. McLachlan, and D. Eisenberg, Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, 1987. **84**(13): p. 4355-4358.
- [7] Marchler-Bauer, A., et al., CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*, 2002. **30**(1): p. 281-283.
- [8] Liu, H., et al., BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 2006. **22**(1): p. 103-105.
- [9] Wu, C.H., et al., The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, 2006. **34**(Database issue): p. D187-D191.
- [10] Suzek, B.E., et al., UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 2007. **23**(10): p. 1282-1288.
- [11] Wu, C.H., et al., The iProClass integrated database for protein functional analysis. *Comput Biol Chem*, 2004. **28**(1): p. 87-96.
- [12] Mi, H., et al., The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*, 2005. **33**(Database issue): p. D284-D288.
- [13] Nikolskaya, A.N., et al., PIRSF family classification system for protein functional and evolutionary analysis. *Evol Bioinform Online*, 2006. **2**: p. 197-209.
- [14] Couto, F.M., M.J. Silva, and P.M. Coutinho, Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering* 2006. **16**: p. 15.
- [15] Retief, J.D., Phylogenetic analysis using PHYLIP. *Methods Mol Biol*, 2000. **132**: p. 243-258.
- [16] Zhou, T., et al., An approach for determining evolutionary distance in network-based phylogenetic analysis. *ISBRA'08 Proceedings of the 4th international conference on Bioinformatics research and applications*, 2008.
- [17] Belyi, V.A. and A.J. Levine, One billion years of p53/p63/p73 evolution. *Proc Natl Acad Sci U S A*, 2009. **106**(42): p. 17609-17610.

Predicting DNA-Binding Sites by Exploring the Distribution of Atom Groups around the Surface

Jing Hu¹ and Changhui Yan²

¹ Department of Mathematics and Computer Science, Franklin & Marshall College, Lancaster, PA, USA

² Department of Computer Science, North Dakota State University, Fargo, North Dakota, USA

Abstract - DNA-binding proteins perform various functions in the cells. Determining the structures of protein-DNA complexes using experimental methods are hindered by many obstacles. Thus, computational methods for predicting DNA-binding sites on protein structures are needed to elucidate the mechanism of protein-DNA interactions. In this study, we divided atoms of amino acid residues into 14 groups and used a vector consisting of the distribution of these atom groups to describe the characteristics of protein surface around an amino acid. We then trained a Random Forest method to predict DNA-binding sites on protein surface. The predictions were then refined using a post-processing procedure based on the clustering of DNA-binding residues on the surface. The method achieved an accuracy of 80.8% when evaluated using 10-fold cross-validation. The results show that the distribution of different types of atoms around the surface provides sufficient structural information for predicting DNA-binding sites on protein structures.

Keywords: Random Forest, DNA-binding, prediction, features

1 Introduction

Structural genomics projects are yielding an increasingly large number of protein structures with unknown function. As a result, computational methods for predicting functional sites on these structures are in urgent demand. There has been significant interest in developing computational methods for identifying amino acid residues that participate in protein-DNA interactions based on combinations of sequence, structure, evolutionary information, and chemical or physical properties. Some methods predict DNA-binding sites using protein sequence-derived information as input [1-3]. Compared to methods that make prediction based on protein structures, these methods have the advantage that they can be applied to proteins whose high-resolution structures are unavailable. However, they also suffer relatively low predicting performance. Thus, methods that can explore structural features to detect DNA-binding sites are also needed. For example, Jones et al. [4] analyzed residue patches on the surface of DNA-binding proteins and used electrostatic potentials of residues to predict DNA-binding sites. Later, they extended that method by including DNA-binding structural motifs [5]. In related studies, Tsuchiya et al. [6]

used a structure-based method to identify protein-DNA binding sites based on electrostatic potentials and surface shape, and Keil et al. [7] trained a neural network classifier to identify patches likely to be DNA-binding sites based on physical and chemical properties of the patches. Neural network classifiers have also been used to identify protein-DNA interface residues based on a combination of sequence and structural information [8, 9]. Many recent studies have also been published [10-13].

Bagley and Altman [14] developed a FEATURE method to investigate the radial distributions of properties around protein sites like binding sites for calcium, the milieu of disulfide bridges, and the serine protease active site. Later, the method was also used to detect zinc-binding sites [15], phosphorylation sites [16], and peptide binding sites [17]. Using a similar approach, in this study, we investigated the distribution of atomic groups around the DNA-binding sites and trained a random forest method to predict DNA-binding sites on protein structures.

2 Materials and methods

2.1 Datasets

139 protein-DNA complexes were extracted from the PDB [18]. All the structures had resolution better than 3.0 Å and R factor less than 0.3. Each protein in this set had at least 40 amino acid residues and the mutual sequence similarity between the proteins in this set was less than 30%.

2.2 Definition of binding-site residues

Binding-site residues were defined based on atom distance [19]. A protein residue was defined to be a DNA-binding residue if the distance from any of its atoms to any atom of the interacting DNA was less than 5 Å. The 139 proteins had 26,862 residues in total and 5,932 of them were DNA-binding residues. A residue was defined to be a surface residue if its relative accessibility is at least 5% as calculated using NACCESS [20].

2.3 Microenvironmental features of DNA-binding sites on protein surface

We calculated the distance from nucleotides to protein surface. The average distance is 6 Å. Thus, for every surface residue, we define a sphere such that the center of the sphere is 6 Å from the protein surface and the line connecting the sphere center and the most exposed atom of the residue was perpendicular to the protein surface. Then we counted the number of different types of atoms from amino acids that fall into the sphere. The atoms were divided into 14 types as described in [17], namely: C3 (aliphatic carbons; sp3), C= (carbonyl carbon; sp2), O= (carbonyl oxygen; sp2), N2H (nitrogen of amides; sp2; also sp2 neutral nitrogen of side chains), Car (aromatic carbon; sp2; general), O2- (negatively charged oxygens (-1/2) in carboxylates; sp2), SH (sulphur in thiols; sp3), OH (hydroxyl group; sp3), NarH (aromatic nitrogen with a hydrogen; sp2), NarH+ (aromatic nitrogen with a hydrogen and a positive charge; sp2), Set (sulphur in thioethers; sp3), C+ (carbon of carbocations; sp2), N3H+ (sp3 nitrogen with a hydrogen and a positive charge), N2H+ (sp2 nitrogen with a hydrogen and a positive charge). Thus, for each surface amino acid residue, a vector of 14 features was obtained. These vectors describe the structural characteristics on the protein surface centering at each surface amino acid. We used these vectors to train a classifier to classify surface residues into DNA-binding and non-DNA-binding classes based on these structural characteristics. Different radius values of the sphere were tested and the best result was achieved when the radius was 20 Å.

2.4 Classifier for predicting DNA-binding residues

We used a Random Forest (RF) method [21] to train a classifier to predict DNA-binding residues. A RF is a method consisting of an ensemble of tree-structured classifiers. It has been applied to solve many bioinformatics problems in recent years. In this study, we used the implementation of RF in WEKA package [22]. Ten fold cross-validations were used to evaluate the performance of the classifier. The proteins in the dataset were randomly split into 10 subsets. In each round of experiments, 9 subsets were used as training set to train a classifier, and the remaining subset was used as test set. This procedure was repeated 10 times with each subset being used as test set once. From a protein in the training set, the feature vectors associated with all binding-site residues were used as positive examples. We noticed that the sphere of a binding-site residue and that of a non-binding surface residue might overlap in space. Thus, to reduce noise in the training set, for the negative examples we only considered the surface residues whose spheres did not overlap with any spheres of binding-site residues. The feature vectors extracted from these residues were used as negative training examples. For a protein from the test set, all surface residues were used as test examples, so that a prediction was made for every surface residue.

2.5 Assessment of prediction performance

Prediction performance was evaluated using sensitivity, precision, accuracy (ACC), and Matthews' correlation coefficient (MCC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (4)$$

where TP was the number of true positives (i.e., residues predicted to be DNA-binding residues that are in fact DNA-binding residues); TN was the number of true negatives (i.e., residues predicted to be non-DNA-binding residues that are in fact non-DNA-binding residues); FN was the number of false negatives (i.e., residues predicted to be non-DNA-binding residues that are in fact DNA-binding residues) and FP was the number of false positives (i.e., residues predicted to be DNA-binding residues that are in fact not interface residues). Sensitivity is a measure of the percentage of DNA-binding residues that are correctly predicted. Specificity is the fraction of non-DNA-binding residues that are correctly predicted. Accuracy is the percentage of overall predictions that are correct. MCC (Matthews correlation coefficient) measures the correlation between predictions and actual class labels, which is in the range of [-1, 1], with 1 denoting perfect predictions and -1 denoting that every example is incorrectly predicted. In a two-class classification, if the numbers of examples of the two classes are not equal, MCC is a better measure than accuracy [23].

3 Results

3.1 Identification of DNA-binding residues by the Random Forest method

A Random Forest (RF) classifier was trained to predict whether a surface residue is DNA-binding residue based on the feature vector associated with its surrounding sphere. 10-fold cross-validation was used to evaluate the performance of the classifier. Table 1 (column 2) shows that the classifier achieved an overall accuracy of 67.3% with a MCC of 0.2, and 57.9% of DNA-binding residues and 69% of non-DNA-binding residues are correctly identified.

3.2 Post-processing of prediction results

A visualization of the DNA-binding sites revealed that DNA-binding residues clustered on protein surface to form a contiguous patch. Thus, the predicted DNA-binding residues were also expected to form a patch on the surface. However, when we analyzed the prediction results by RF, we found that that some predicted DNA-binding residues were isolated on protein surface, and in some cases, the predicted DNA-

binding sites form multiple small patches on the surface. Thus, we designed a post-processing procedure to remove isolated predictions and merge small patches into a large one. For a surface residue that was predicted to be DNA-binding residue, if less than 2 of its neighboring surface residues were predicted to be DNA-binding, then we changed its prediction to non-DNA-binding. For a surface residue that was predicted to be non-DNA-binding, if more than 60% of its neighboring residues were predicted to be DNA binding, then we changed its prediction to DNA binding. After the post-processing (Table 1, column 3), the prediction performance was improved to an overall accuracy of 73.5% with a MCC of 0.26, and 57.2% of DNA-binding residues and 76.0% of non-DNA-binding residues are correctly identified. Compared this with the results without post-processing, we can see that the post-processing improve accuracy, MCC, and precision at only little cost of sensitivity.

3.3 Relaxation of prediction results after post-processing

In this study, the DNA-binding residues were defined based on their distance to the binding DNA using a cutoff chosen in a previous study [19]. However, different cutoff values had been used in many other studies. In our study, the majority of the false positive predictions were very close to the observed DNA-binding residues (either being the direct neighbor of a DNA-binding residue or separated from the DNA-binding sites by only one residue). Some of these false positive predictions could have been counted as true positives if a different cutoff value was used. To account for the uncertainty in the cutoff value, we re-evaluated the performance by relaxing the criterion of true positive as in [9]. With the relaxed criterion when a surface residue was predicted to be a DNA-binding residue, the prediction is considered a true positive prediction if (1) the surface residue was indeed a DNA-binding residue, or (2) it was a direct neighbor (on the protein surface) of a DNA-binding residue. After the relaxation (Table 1, column 4), the prediction had an accuracy of 80.8%, with 0.50 MCC, 71.5% sensitivity and 80.8% precision.

Table 1. Prediction performances of the proposed method

	Random Forest ¹	Post-processing ²	Relaxation ³
Sensitivity (%)	57.9	57.2	71.5
Specificity (%)	69.0	76.0	83.5
ACC (%)	67.3	73.5	80.8
MCC	0.20	0.26	0.50

¹Predictions by the Random Forest method. ²Predictions from the Random Forest method were processed using the post-

processing procedure. ³A relaxed criterion of true positive was used to evaluate the performance.

4 Conclusions

In this study, we used vectors consisting of the distribution of atom groups to describe the characteristics of protein surface and used them to train a RF method to predict DNA-binding residues. A post-processing procedure was used to refine the predictions based on the distribution of DNA-binding residues over the protein surface. After the post-processing, the predicted DNA-binding sites form a contiguous path on the protein surface. The accuracy of the method reached 80.8% based on a relaxed criterion. The results confirmed that the distribution of atom groups on the protein surface provided useful structural information for predicting DNA-binding sites.

5 References

- [1] Yan, C., et al., Identifying amino acid residues involved in protein-DNA interactions from sequence. *BMC Bioinformatics*, 2006. **7**: p. 262.
- [2] Ahmad, S. and A. Sarai, PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, 2005. **6**(1): p. 33.
- [3] Wang, L. and S.J. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucl Acids Res*, 2006. **34**: p. W243-W248.
- [4] Jones, S., et al., Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucl Acids Res*, 2003. **31**(24): p. 7189-7198.
- [5] Shanahan, H.P., et al., Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucl Acids Res*, 2004. **32**(16): p. 4732-4741.
- [6] Tsuchiya, Y., K. Kinoshita, and H. Nakamura, Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, 2004. **55**(4): p. 885-894.
- [7] Keil, M., T. Exner, and J. Brickmann, Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. *J Comput Chem*, 2004. **25**(6): p. 779-789.
- [8] Ahmad, S., M.M. Gromiha, and A. Sarai, Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 2004. **20**(4): p. 477-486.

[9] Tjong, H. and H.-X. Zhou, DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucl. Acids Res.*, 2007. **35**(5): p. 1465-1477.

[10] Xiong, Y., J. Liu, and D.-Q. Wei, An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins: Structure, Function, and Bioinformatics*, 2011. **79**(2): p. 509-517.

[11] Cai, Y., et al., A novel sequence-based method of predicting protein DNA-binding residues, using a machine learning approach. *Molecules and Cells*, 2010. **30**(2): p. 99-105.

[12] Alibés, A., L. Serrano, and A. Nadra, Structure-based DNA-binding prediction and design, in *Methods Mol Biol*. 2010. p. 77-88.

[13] Huang, Y.-F., et al., DNA-binding residues and binding mode prediction with binding-mechanism concerned models. *BMC Genomics*, 2009. **10**(Suppl 3): p. S23.

[14] Bagley, S.C. and R.B. Altman, Characterizing the microenvironment surrounding protein sites. *Protein Science*, 1995. **4**(4): p. 622-635.

[15] Ebert, J.C. and R.B. Altman, Robust recognition of zinc binding sites in proteins. *Protein Science*, 2008. **17**(1): p. 54-56.

[16] Fan, S. and X. Zhang, Characterizing the microenvironment surrounding phosphorylated protein sites. *Genomics Proteomics Bioinformatics*, 2005. **3**(4): p. 213-217.

[17] Petsalaki, E., et al., Accurate Prediction of Peptide Binding Sites on Protein Surfaces. *PLoS Comput Biol*, 2009. **5**(3): p. e1000335.

[18] Berman, H.M., et al., The Protein Data Bank. *Nucl Acids Res*, 2000. **28**(1): p. 235-242.

[19] Ofra, Y. and B. Rost, Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, 2003. **325**(2): p. 377-387.

[20] Hubbard, S.J., NACCESS. 1993, Department of Biochemistry and Molecular Biology, University College, London.

[21] Breiman, L. RF/tools: A class of two-eyed algorithms. in *SIAM workshop*. 2003.

[22] Witten, I.H. and E. Frank, *Data mining: practical machine learning tools and techniques with Java implements*. 1999, San Mateo, CA: Morgan Kaufmann.

[23] Baldi, P., et al., Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 2000. **16**: p. 412-424.

Structural Analysis of Molecular Networks: AMES Mutagenicity

Laurin AJ Mueller and Karl G Kugler and Matthias Dehmer
laurin.mueller@umit.at and karl.kugler@umit.at and matthias.dehmer@umit.at

Institute of Bioinformatics and Translational Research
University for Health Sciences, Medical Informatics and Technology (UMIT)
Eduard Wallnöfer-Zentrum 1
6060 Hall in Tyrol, Austria

Abstract—The characterization of chemical compounds based on their molecular graphs is an important task for identifying properties such as toxicity or mutagenicity. We used different groups of topological descriptors using the AMES mutagenicity data. Instead of optimizing the classification performance, the aim of this study is to perform a structural analysis of the underlying set of molecular graphs to gain better insights of the data set.

The structural analysis identifies two groups of molecular networks. One group contains graphs with linear patterns (outliers), and the other group contains graphs that exhibit patterns of regular graphs (remainders). We show that the set of used topological descriptors chosen for this study cannot capture enough group-specific structural information within the remainders group to achieve the discrimination ability of the outliers group. Finally, this leads us to the conclusion that it is necessary to identify existing or develop new descriptors that capturing specific structural information to achieve better discrimination ability.

Keywords: Topological network descriptors, network biology, drug design, machine learning

I. BACKGROUND

The classification of drug-like compounds by using structural information of their underlying molecular graphs is an important task to identify chemical properties (e.g. toxicity or mutagenicity) [Feng et al., 2003], [Votano et al., 2004]. In general, graph classification is a challenging problem and has been tackled by using different methods [Cook and Holder, 2007], [Dehmer and Mehler, 2007], [Deshpande et al., 2003]. Note that classical work relates to applying methods from exact and inexact graph matching [Cook and Holder, 2007], [Dehmer and Mehler, 2007]. In a more biologically motivated work performed by Li et al., graph kernels to predict gene functions have been utilized [Li et al., 2007]. Chuang et al. used subnetworks to train a classifier for the detection of breast cancer metastasis [Chuang et al., 2007].

For our investigation we use the Ames mutagenicity data set, that is a benchmark set to classify graphs [Hansen et al., 2009]. It consists of 6512 graphs, that represent compounds that are categorized as Ames posi-

tive ($AMES^+$) or negative ($AMES^-$) by the Ames test [Ames et al., 1973]. Hansen et al. [Hansen et al., 2009] used the commercial software tool Dragon [Todeschini et al., 2003] to calculate a large set of molecular network descriptors to classify the Ames mutagenicity data set.

Dehmer et al. [Dehmer et al., 2010] used entropy-based descriptors [Dehmer and Mowshowitz, 2011] for weighted chemical structures to classify the AMES data set. After removing the isomorphic graphs, they showed that it is possible to classify the remaining graphs with a reasonable classification performance, by only using a set of seven descriptors.

For our analysis we modify this set of graphs, as we only consider the structural skeletons of the molecules. We construct a structural skeleton by using unlabeled nodes and unweighted edges. The main contribution of this paper is to identify discriminatory features of the AMES graphs to classify the structures properly. For this, we calculate the descriptors using the freely available R-package QuACN [Mueller et al., 2010b] and selected groups of measures from Dragon [Todeschini et al., 2003] on the resulting set of molecular skeletons.

Note, the classification of Ames mutagenicity by only using structural properties without labels is surely a critical undertaking. The aim of this study is not to increase or optimize the classification performance for this data set but rather to investigate the structural information of molecular networks.

This paper is structured as follows: The Material and Methods section describes the data set of molecular networks that we analyze and gives a brief overview about the used methods. The results section lists the results of the initial classification that motivates the structural analysis. It also contains the results of the structural analysis of the data. In chapter IV we summarize and discuss the results. Section V concludes the paper and provides an outlook on further investigation steps.

II. MATERIAL AND METHODS

The modified AMES Mutagenicity Set for Molecular Networks

The initial data set of Ames mutagenicity [Hansen et al., 2009] was designed to benchmark the

classification performance of different kind of graph classification strategies. It contains 6512 molecular compounds that were categorized positive or negative by the Ames test [Ames et al., 1973] for mutagenicity. Hansen et al. [Hansen et al., 2009] used six different public available data sets and studies to create this benchmark data set. This data set contains $n_+ = 3503$ AMES positive ($AMES^+$) and $n_- = 3009$ AMES negative ($AMES^-$) molecular networks. We used the data set of Dehmer et al. [Dehmer et al., 2010] where isomorphic graphs were removed and modified this set, as we only took the structural skeletons of the molecules. This means that each atom is represented by an unlabeled vertex. Moreover, we represent each kind of bond with an undirected edge. This results in a data set of $n = 3947$ skeletons of molecular networks with $n_+ = 2179$ AMES positive and $n_- = 1768$ AMES negative graphs. This set of molecular networks was used for further analysis.

Topological Network Descriptors

After modifying the AMES data set we calculate different groups of topological descriptors. Topological network descriptors are numerical graph invariants that quantitatively characterize the structure of the underlying network [Emmert-Streib and Dehmer, 2011]. We calculated the entropy-based descriptors available in QuACN [Mueller et al., 2010b] and six groups of descriptors offered by the commercial software tool Dragon [Todeschini et al., 2003]. Table I gives an overview about the calculated descriptors.

Each descriptor in Table I results in a single value that characterizes the structure of the underlying molecular network in a certain way. The calculated descriptors can be treated like features and then be used for machine learning [Mueller et al., 2010a], [Mueller et al., 2011].

Also, we will not describe the descriptors in detail. For a better understanding of the selected measures see corresponding literature (e.g. [Bonchev, 1983], [Dehmer et al., 2010], [Todeschini and Consonni, 2009], [Mowshowitz, 1968]). Dehmer and Mowshowitz [Dehmer et al., 2010] discuss entropy-based descriptors, Todeschini et al. [Todeschini and Consonni, 2009] describes the descriptors implemented in Dragon.

Supervised Machine Learning

To classify the molecular networks between $AMES^+$ and $AMES^-$ we treat every topological descriptor as feature. We use support vector machines (SVM) [Vapnik and Lerner, 1963] with a radial basis function kernel.

To compare the results of the support vector machines we use Random Forest (RF) [Svetnik et al., 2003]. After optimizing the parameters and the classification with the mentioned algorithms we calculate the area under the ROC-curve (AOC), the accuracy and the f-score of the results. For each classification we perform a 10-fold cross validation.

To select the best set of topological network descriptors we use the feature selection algorithm information gain [Quinlan, 1993]. The best features of each group were combined to a so called superindex that is defined as follows [Bonchev et al., 1981], [Dehmer et al., 2010].

Definition 1. Let I_1, \dots, I_j be topological network descriptors. The superindex of these measures is defined as $SI := \{I_1, \dots, I_j\}$.

III. RESULTS

Supervised Machine Learning

The performance of the classification with support vector machines is shown in Table II. The corresponding ROC curves are shown in Fig. 1. It can be seen that the different groups lead to divergent results. The group of vertex degree-based topological descriptors (Dragon 3) achieves the best accuracy with 73.04%. Four groups (Dragon 1, 2, 3, and 5) achieve similar AUCs with about 72%. The groups Dragon 1 and Dragon 3 achieve the best f-scores of about 67%, for details see Table II. It can be summarized that the best classification performances (accuracy, AOC and f-score) is achieved by using the group Dragon 3, containing vertex degree-based topological descriptors.

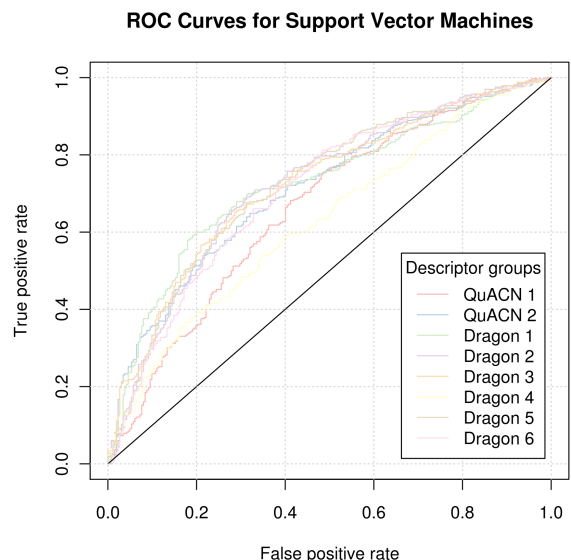


Fig. 1. This figure shows the ROC curves for each descriptor group for the classification using support vector machines.

To evaluate the performance of the support vector machine we use RF to classify the same groups again. Fig. 2 shows the ROC curves for the different groups of descriptors. The results in Table III show that the best performance is achieved by the groups Dragon 1, 2 and 3. The group called Dragon 1 has the highest AUC of 74.89%, Dragon 2 the highest accuracy of 71.55% and Dragon 3 achieved the highest f-score of 67.34%.

This result shows that the different groups of topological network descriptors perform similar using SVM and RF.

TABLE I
OVERVIEW OF THE USED INFORMATION-THEORETIC TOPOLOGICAL DESCRIPTORS.

Group name	Group	Subgroup	No. Descriptors
QuACN 1	Entropy based	Partition based and parametric graph entropy	9
QuACN 2	Polynomial based	-	50
Dragon 1	Walk and path counts	-	46
Dragon 2	Connectivity indices	-	37
Dragon 3	Topological indices	Vertex degree-based	26
Dragon 4	Topological indices	Distance-based indices	13
Dragon 5	Information indices	Basic descriptors	17
Dragon 6	Information indices	Indices of neighborhood symmetry	30

TABLE II
CLASSIFICATION PERFORMANCE FOR EACH GROUP USING SVM

	Precision	Recall	Specificity	Sensitivity	Accuracy	AUC	F-Score
QuACN 1	0.4785	0.6395	0.6486	0.6395	0.6456	0.6625	0.5474
QuACN 2	0.5803	0.6854	0.6971	0.6854	0.6927	0.7120	0.6285
Dragon 1	0.6476	0.7152	0.7344	0.7152	0.7266	0.7216	0.6797
Dragon 2	0.6041	0.7427	0.7210	0.7427	0.7289	0.7220	0.6663
Dragon 3	0.6357	0.7280	0.7320	0.7280	0.7304	0.7223	0.6787
Dragon 4	0.4649	0.6089	0.6357	0.6089	0.6266	0.6265	0.5273
Dragon 5	0.5288	0.7203	0.6855	0.7203	0.6970	0.7243	0.6099
Dragon 6	0.5696	0.6638	0.6868	0.6638	0.6780	0.7088	0.6131

TABLE III
CLASSIFICATION PERFORMANCE FOR EACH GROUP USING RANDOM FOREST

	Precision	Recall	Specificity	Sensitivity	Accuracy	AUC	F-Score
QuACN 1	0.5215	0.5895	0.6450	0.5895	0.6230	0.6281	0.5534
QuACN 2	0.5724	0.6216	0.6740	0.6216	0.6524	0.7102	0.5960
Dragon 1	0.6663	0.6747	0.7319	0.6747	0.7066	0.7489	0.6705
Dragon 2	0.6369	0.7007	0.7256	0.7007	0.7155	0.7406	0.6673
Dragon 3	0.6578	0.6898	0.7324	0.6898	0.7142	0.7363	0.6734
Dragon 4	0.6222	0.6599	0.7070	0.6599	0.6871	0.6022	0.6405
Dragon 5	0.6227	0.6613	0.7077	0.6613	0.6881	0.7223	0.6414
Dragon 6	0.5339	0.5885	0.6483	0.5885	0.6240	0.7166	0.5599

Moreover, the classification with RF achieves a slightly higher performance. However, it can be seen that the groups Dragon 1-3 are qualified best to discriminate between $AMES^+$ and $AMES^-$ for this set of molecular networks.

To study the classification performance we perform a feature selection with information gain for each group and selected the best three descriptors of each group to create a superindex. Classifying by applying the superindex leads to the results shown in Table IV. The ROC curves are shown in Fig. 3.

The performance of SVM and RF are similar but RF performs better with an accuracy of 74.21% and AUC of 76.62 and an f-score of 70.80%.

To evaluate the stability of the results we randomly select 1000 molecular networks and classify them using the superindex and RF. We repeat this procedure 1000 times. This results in a mean f-score of 64% with a standard deviation of 2%. This small standard deviation indicates that the classification performance is stable.

In order to analyze the classification performance we investigate the structural information of the set of the molecular networks. Therefore, we calculate a set of distance-based descriptors [Skorobogatov and Dobrynin, 1988] to explore basic structural properties.

Exemplarily, we use the average path length to outline a prototype of the structural analysis. Fig. 4 shows the average

path length (APL) for all molecular networks. One function represents the graphs that are grouped as $AMES^+$ the other one shows the graphs that are $AMES^-$. The vertical lines represent the mean and the standard deviation (dashed) for each group. Fig. 4 shows, in a descriptive way, that the distribution of the average path length of the two groups ($AMES^+$ and $AMES^-$) is largely overlapping. This can also be observed for the other distance-based descriptors.

We hypothesize that the outliers are more discriminative than the remaining graphs. We define outliers as at least one standard deviation away from the mean of each group (see Fig. 4). Using this criteria we split the molecular networks into two groups (outliers and remainders) with $n_{outliers} = 1102$ and $n_{rest} = 2623$ graphs. We then classifying this two group separately, using the superindex and random forest. This results in an f-score for the outliers of 72.73%. The performance of the classification for the remainders obtained an f-score of 66.63%.

To identify structural information of the different groups we look at single graphs in the two groups ($AMES^+$ and $AMES^-$). Fig. 5 exemplary shows two graphs of each group. Fig. 5(a) and 5(b) show two outliers, and Fig. 5(c) and 5(d) represent two networks of the remainders. It can be seen that the outliers possess linear patterns, in contrast the remainders show regular patterns. A regular graph is a graph where each

TABLE IV
CLASSIFICATION PERFORMANCE OF THE SUPERINDEX

	Precision	Recall	Specificity	Sensitivity	Accuracy	AUC	F-Score
Support vector machine	0.6561	0.7374	0.7439	0.7374	0.7413	0.7367	0.6944
Random forest	0.6980	0.7183	0.7604	0.7183	0.7421	0.7662	0.7080

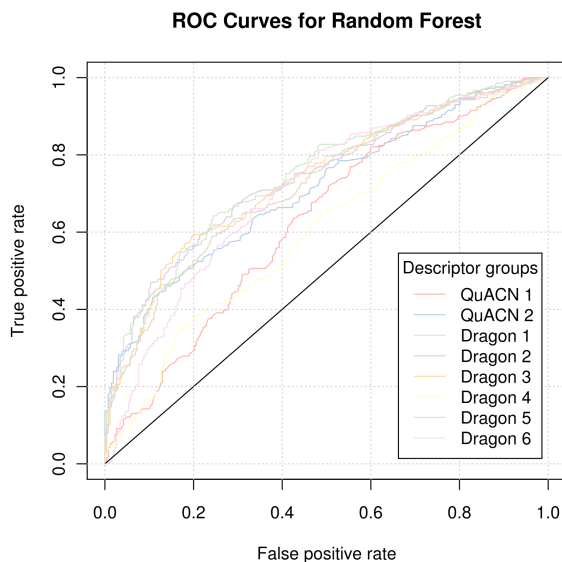


Fig. 2. This figure shows the ROC curves for each descriptor group for the classification using RF.

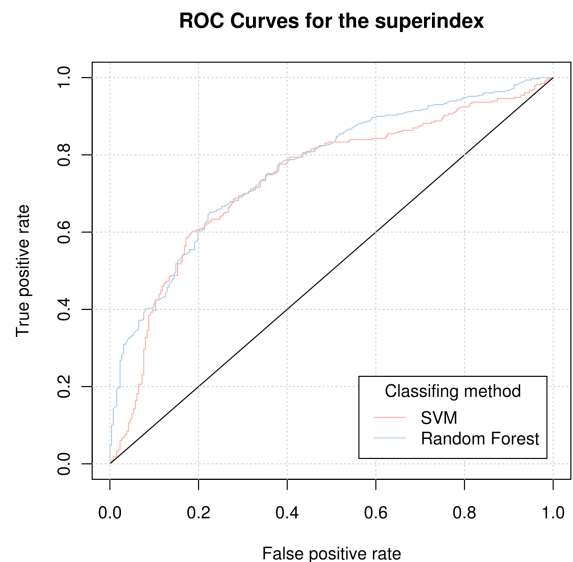


Fig. 3. This figure shows the ROC curves for the classification with SVM and RF using the superindex.

vertex has the same degree. These characteristics can also be observed for other graphs of the corresponding groups.

Repeating this kind of outlier analysis with different distance-based descriptors (i.e. eccentricity or average distance) leads to similar results. In summary we see that the outliers possess linear patterns. This is in contrast to graphs that are close to the mean of the corresponding descriptor, which exhibit rather regularity.

IV. SUMMARY AND DISCUSSION

The AMES mutagenicity set of molecular networks is a benchmark set to evaluate the performance of graph classification algorithms. By only using the underlying skeletons, the classification of this AMES mutagenicity is a difficult and complex endeavor. It becomes even harder, when removing isomorphic graphs, the information of node labels and edge properties. To classify the remaining network skeletons we used different groups of topological network descriptors and constructed a so called superindex by selecting the best features of each group with the feature selection method information gain. We used support vector machines and random forest to perform the classification.

The group of vertex degree-based indices achieved the best results, what indicates that the degree has a high discrimination ability within this set of molecular networks. Different groups of topological network descriptors capture different

structural information, what led to a higher discrimination ability by combining them to a superindex.

Hansen et al. [Hansen et al., 2009] achieved an AUC of 86%. One can see that our classification performance is less than 10% lower. Considering the fact, that Hansen et al. also used groups of descriptors that take information about the atoms (e.g.: atom type, atom weights) and different binding types into account, and we reduced the information in the molecular network by reducing them to their structural skeletons, we achieve fairly acceptable results. Moreover, the removal of the isomorphic graphs can be a reason for the lower discrimination ability. Imagine that if a graph is correctly classified, all isomorphic graphs would also be correctly classified, what would increase the overall performance of the classification. By using molecular skeletons it can happen that two molecules are reduced to the same skeleton and then can be found in the $AMES^+$ and in the $AMES^-$ group. That can also be a possible reason for a lower classification power.

Comparing our results with Dehmer et al. [Dehmer et al., 2010], they achieved 71.4% including label information, shows that our best classification performance by using the superindex and random forest, is only about 0.6% lower. Compared to their results when using unlabeled graphs, the difference is even smaller. Note, that the fact that the results are fairly the same, strengthens our hypothesis that the classification performance cannot be increased dramatically,

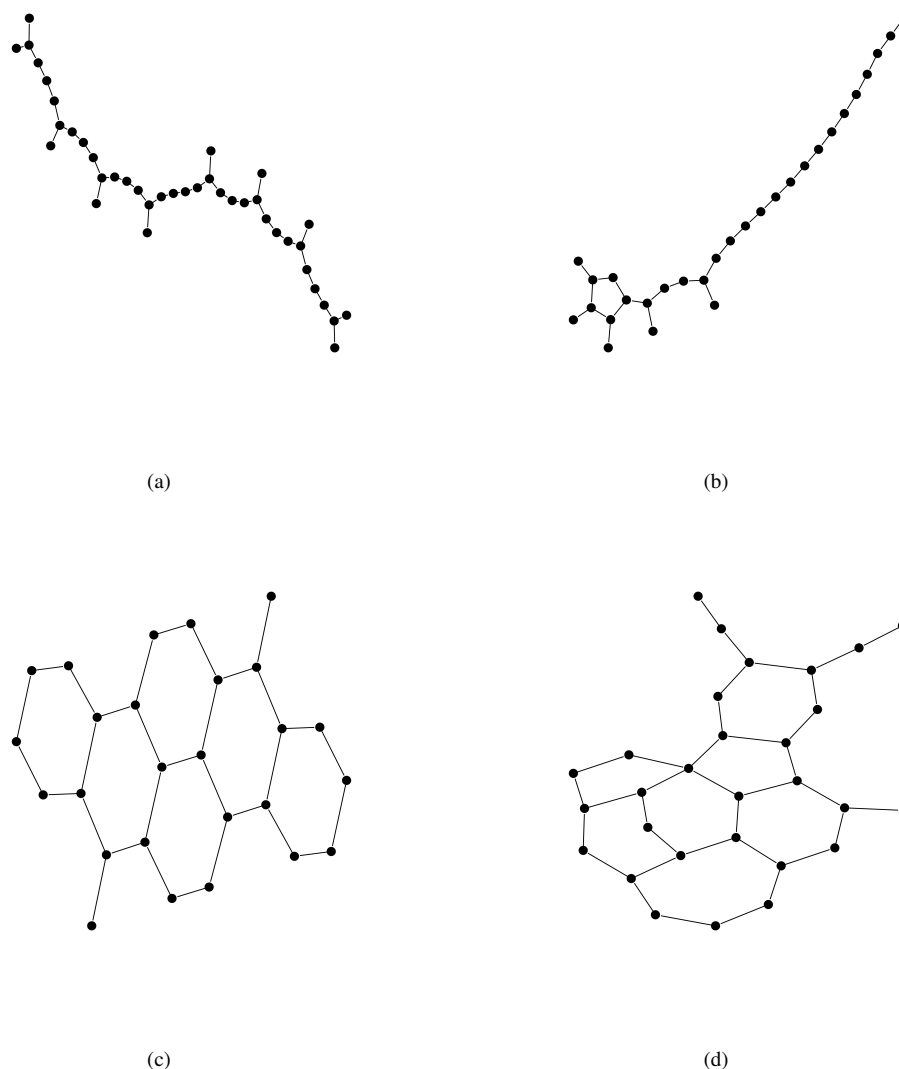


Fig. 5. This figure shows exemplary two graphs of the two groups, split by the value of the average path length. One group ((a) and (b)) represent outliers that are at least one standard deviation away from the mean (see Fig. 4. (c) and (d) represent the groups of the remaining molecular networks.

by only using the this set of molecular descriptors.

In order to increase the classification performance in further studies we analyzed the set of molecular networks structurally. Therefore, we applied a set of distance-based descriptors to them, and analyzed the structure of the outliers. An interesting finding is that the outliers show linear patterns, compared to the remaining graphs that show properties of regular graphs. Moreover, as these regular graphs show more equal vertex degrees than the linear ones, this assumption matches with observation that the group of vertex degree-based descriptors has the highest classification performance of all selected groups of topological network descriptors. An other interesting finding is that the remaining regular graphs contain ring-like structures.

V. CONCLUSION AND OUTLOOK

This study deals with the structural analysis of the AMES mutagenicity data set. It turned out that vertex degree-based descriptors led to a good classification performance. Combining different groups of descriptors to a superindex is promising as it increased the classification performance.

The major challenge of this study was to explore the selected topological network descriptors. They failed to capture enough structural information that would have been needed for achieving a better discrimination ability. The structural analysis showed that there is a set of graphs possessing linear patterns and a set of graphs showing regular characteristics. For future work it is necessary either to identify existing descriptors or develop new descriptors that can better discriminate between these graphs. Therefore, a thorough analysis of

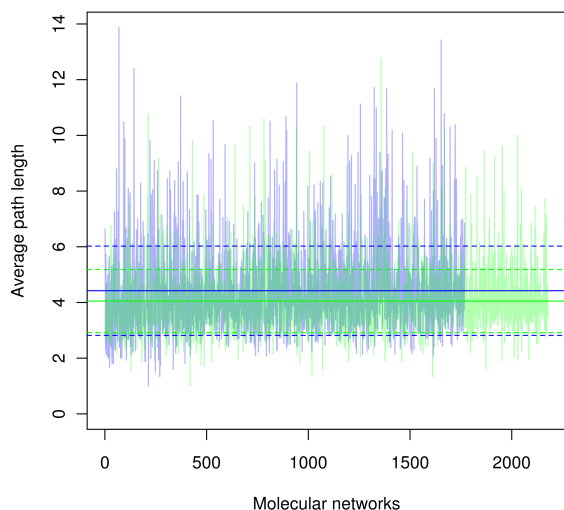


Fig. 4. This figure shows the average path length for each group: $AMES^+$ (green) and $AMES^-$ (blue). The vertical lines represent the mean and the standard deviation (dashed) for each group.

the data set is needed.

Moreover, defining superindices to combine different descriptors, which capture different kinds of structural properties can be a promising strategy. The combination of different superindices could lead to an approach that can capture group-specific combinations of different structural information to distinguish between AMES positive and AMES negative tested chemical compounds.

VI. ACKNOWLEDGEMENT

This work was funded by the Tiroler Wissenschafts Fonds (Project CoNAN - Phase II) and the Tiroler Zukunftsstiftung.

This work was supported by the COMET Center ONCO-TYROL and funded by the Federal Ministry for Transport Innovation and Technology (BMVIT) and the Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth (BMWA/BMWFJ), the Tiroler Zukunftsstiftung (TZS) and the State of Styria represented by the Styrian Business Promotion Agency (SFG).

We are grateful to Kurt Varmuza for calculating the topological network descriptors by using Dragon.

REFERENCES

[Ames et al., 1973] Ames, B. N., Lee, F. D., and Durston, W. E. (1973). An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proceedings of the National Academy of Sciences of the United States of America*, 70(3):782–6.

[Bonchev, 1983] Bonchev, D. (1983). *Information theoretic indices for characterization of chemical structures*. Chemometrics research studies series. Research Studies Press.

[Bonchev et al., 1981] Bonchev, D., Mekenyan, O., and Trinajsitić, N. (1981). Isomer discrimination by topological information approach. *Journal of Computational Chemistry*, 2(2):127–148.

[Chuang et al., 2007] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140.

[Cook and Holder, 2007] Cook, D. and Holder, L. B. (2007). *Mining Graph Data*. Wiley-Interscience.

[Dehmer et al., 2010] Dehmer, M., Barbarini, N., Varmuza, K., and Graber, A. (2010). Novel topological descriptors for analyzing biological networks. *BMC Structural Biology*, 10(1):18.

[Dehmer and Mehler, 2007] Dehmer, M. and Mehler, A. (2007). A new method of measuring similarity for a special class of directed graphs. *Tatra Mountains Mathematical Publications*, 36:39–59.

[Dehmer and Mowshowitz, 2011] Dehmer, M. and Mowshowitz, A. (2011). A history of graph entropy measures. *Information Sciences*, 181(1):57–78.

[Deshpande et al., 2003] Deshpande, M., Kuramochi, M., and Karypis, G. (2003). Automated approaches for classifying structures. In *Proceedings of the 3-rd IEEE International Conference of Data Mining*, pages 35–42.

[Emmert-Streib and Dehmer, 2011] Emmert-Streib, F. and Dehmer, M. (2011). Networks for Systems Biology: Conceptual Connection of Data and Function. *IET Syst Biol*.

[Feng et al., 2003] Feng, J., Lurati, L., Ouyang, H., Robinson, T., Wang, Y., Yuan, S., and Young, S. S. (2003). Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. *J. Chem. Inf. Comput. Sci.*, 43(5):1463–1470.

[Hansen et al., 2009] Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., and Müller, K.-R. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. *Journal of chemical information and modeling*, 49(9):2077–81.

[Li et al., 2007] Li, X., Zhang, Z., Chen, H., and Li, J. (2007). Graph Kernel-Based Learning for Gene Function Prediction from Gene Interaction Network. In *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine*.

[Mowshowitz, 1968] Mowshowitz, A. (1968). Entropy and the complexity of the graphs i: An index of the relative complexity of a graph. *Bull Math Biophys*, 30:175–204.

[Mueller et al., 2010a] Mueller, L. A., Kugler, K. G., Dander, A., Graber, A., and Dehmer, M. (2010a). Network-based approach to classify disease stages of prostate cancer using quantitative network measures. *Conference on Bioinformatics & Computational Biology (BIOCAMP'10), Las Vegas/USA*, I:55–61.

[Mueller et al., 2010b] Mueller, L. A., Kugler, K. G., Dander, A., Graber, A., and Dehmer, M. (2010b). QuACN - An R Package for Analyzing Complex Biological Networks Quantitatively. *Bioinformatics*, submitted.

[Mueller et al., 2011] Mueller, L. A., Kugler, K. G., Netzer, M., Graber, A., and Dehmer, M. (2011). Distinguishing between the three domains of life using topological characteristics of their underlying metabolic networks. *submitted*.

[Quinlan, 1993] Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, CA, USA.

[Skorobogatov and Dobrynin, 1988] Skorobogatov, V. A. and Dobrynin, A. A. (1988). Metrical analysis of graphs. *Commun. Math. Comp. Chem.*, 23:105–155.

[Svetnik et al., 2003] Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., and Feuston, B. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.*, 43(6):1947–1958.

[Todeschini and Consonni, 2009] Todeschini, R. and Consonni, V. (2009). *Molecular descriptors for chemoinformatics*. Vch Pub.

[Todeschini et al., 2003] Todeschini, R., Consonni, V., Mauri, A., and Pavan, M. (2003). Software dragon: Calculation of molecular descriptors, department of environmental sciences. Taletè.

[Vapnik and Lerner, 1963] Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24.

[Votano et al., 2004] Votano, J. R., Parham, M., Hall, L. H., and Kier, L. B. (2004). New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. *Mol Divers*, 8(4):379–391.

A Study of Correlations Between the Definition and Application of the Gene Ontology

Yuji Mo, Catherine Anderson and Stephen D. Scott

Dept. of Computer Science

University of Nebraska

Lincoln, NE 68588-0115

{ymo,anderson,sscott}@cse.unl.edu

Abstract

When using the Gene Ontology (GO), nucleotide and amino acid sequences are annotated by terms in a structured and controlled vocabulary organized into a relational graph. The usage of the vocabulary (GO terms) in the annotation of these sequences may diverge from the relations defined in the ontology. We measure the consistency of the use of GO terms by comparing GO's defined structure to the terms' application. To do this, we first use synthetic data with different characteristics to understand how these characteristics influence the correlation values determined by various similarity measures. Using these results as a baseline, we found that the correlation between GO's definition and its application to real data is relatively low, suggesting that GO annotations might not be applied in a manner consistent with its definition. In contrast, we found a sub-ontology of GO that correlates well with its usage in UniProtKB.

1. Introduction

The Gene Ontology (GO) [1] is a controlled vocabulary describing the domain of gene products, i.e., enzymes and other proteins encoded in DNA. GO is made up of three independent, orthogonal ontologies: (1) the Cellular Component ontology, which describes where a gene product is located at a subcellular level; (2) the Molecular Function ontology, which describes the function a gene product can perform; and (3) the Biological Process ontology, which describes series of events and molecular functions. Each ontology is structured as a directed acyclic graph (DAG). Each node of each DAG is a term with a distinct name and description. The edges of a DAG represent the relations between the connected nodes. The relations are endowed with descriptive logic so that inferences can be made between parent and child nodes. A gene product can be annotated by assigning GO terms to the description of the gene product. This assignment is also referred to as an *association* between a term and a gene product.

GO has become widely accepted in the genomics community as a concise means of annotating gene products for machine translation [2]. However, due to the wide scope of the genomics community, ambiguities in term usage exist.

The GO project is a collaborative effort between groups sharing their vocabularies. Group members participate on a self-interested, best-effort basis to reach consensus on the addition, deletion or editing of terms within the three ontologies. However, individual curators from different communities may interpret the definitions differently, resulting in inconsistent usage, and thus it is necessary to continually refine terms. With the large increase of gene products that are annotated with GO, methods to evaluate semantic similarity based on annotations are critical in evaluating the consistency of usage. This motivates our study, which is to apply measures of *semantic similarity* to estimate the consistency between how GO is defined and how it is used in practice.

The notion of semantic similarity is frequently used in information retrieval, where terms are indexed by similar meaning rather than similar words. This concept was used in early research with natural language processing techniques: associating descriptive language with terms and quantifying this similarity. The ontology terms in GO may be examined by clustering terms together with similar semantics [3] using these techniques.

Earlier work done [4], [5] to determine semantic similarity of terms using the annotation they have been associated with were designed for specific applications: malapropism correction (the correction of outliers in the annotation), assessing functional similarity of gene products [6], predicting protein interaction [7], assessing the influence of electronic annotations [8] and assisting in the annotation of new sequences [4]. In contrast, we use some of the same measures they do, but for the purposes of measuring the consistency of the use of GO.

All three ontologies within GO contain many biologically/biochemically descriptive terms that have not been used (not applied to any annotation). A large number of terms are used only once or not at all. This creates a usage pattern where a large percent of GO terms fall in the tail of the distribution, (called the *long tail phenomenon*). Because of this phenomenon, certain types of similarity measures may be preferable to others in evaluating ontology usage. Thus, one of our results is a test using synthetic data with different characteristics to understand how various similarity measures

measure correlation, and how these measures are influenced by various properties of the data. We then describe how the synthetic data parameters imply properties of real data. Our results show that one measure (called ‘‘Cosine’’) is only useful in recognizing correlations when the gene product usage comes with a long tail and each term is annotated by many moderately concentrated terms in the ontology. Another measure (‘‘Jiang’s’’) is not well suited for unbalanced usage of terms in the ontology. The remaining measures (‘‘Resnik’s,’’ ‘‘Lin’s,’’ and ‘‘Rel’’) are almost independent of the data characteristics that we varied, especially Resnik’s.

Using our results on synthetic data as a baseline, we then sampled partial ontologies from GO and measured correlations between their definitions and their usage. Relative to correlation results found in synthetic data with similar configurations to the real data, we found that the average correlation is low. This might suggest that GO annotations are not applied in a manner consistent with their definition. In contrast, we found that the sub-ontology rooted at the term ‘‘GO:0005275: amine transmembrane transporter activity’’ correlates well with its usage in UniProtKB.

2. Method

2.1 Problem Formalization

An ontology $G = (V, E)$ is a directed acyclic graph (DAG), where each vertex corresponds to a term c_i . There is an edge from c_i to c_j if and only if c_j is explicitly a c_i . Since this ‘‘is_a’’ relation is transitive, c_j is_a c_i if and only if there is a path from c_i to c_j . We consider c_j to be a descendant of c_i if a path from c_i to c_j exists.

According to the gene product annotation guidelines [9], a gene product can be annotated by zero or more nodes of each ontology. Let C_i be the set of terms used to annotate gene product e_i . Similarly, we can define E_j as the set of gene products annotated by term c_j . By definition, $c_j \in C_i \Leftrightarrow e_i \in E_j$. In addition, annotating a gene product with a term implies that the gene product is also annotated by all ancestors of the term. Thus, c_i is a descendant of c_j implies $E_i \subseteq E_j$. The ancestor term inherits all annotations from its descendant, so the root term has all annotations: $E_{root} = \bigcup_i E_i$.

2.2 Similarity Measures

There are many different functions for calculating semantic similarity between terms. We consider the following five measures.

Resnik [10] proposed that the amount of information provided by the common ancestors of the two terms may be used as a measure:

$$Sim_{Resnik}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} -\log P(c_k) , \quad (1)$$

where $S(c_i, c_j)$ is the set of ancestors shared by both c_i and c_j and $P(c_k)$ is the probability that a randomly selected gene product is annotated by term c_k : $P(c_k) = |E_k|/|E_{root}|$.

Lin [11] extended Resnik’s measure by modifying the information content of a term to take both descendants into consideration:

$$Sim_{Lin}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left(\frac{2 \log P(c_k)}{\log P(c_i) + \log P(c_j)} \right) . \quad (2)$$

Generic terms do not have a high relevance for the comparison of different gene products. Andreas’s [5] relevance measure combined both Lin’s and Resnik’s measure by weighting Lin’s similarity measure with $1 - P(c_k)$. For a detailed term c_k , $P(c_k)$ becomes relatively very small and makes $1 - P(c_k)$ close to 1 and negligible:

$$Sim_{Rel}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left(\frac{2(1 - P(c_k)) \log P(c_k)}{\log P(c_i) + \log P(c_j)} \right) . \quad (3)$$

Jiang [12] proposed a similarity measure as the reciprocal of semantic distance:

$$Sim_{Jiang}(c_i, c_j) = \max_{c_k \in S(c_i, c_j)} \left(\frac{1}{-\log P(c_i) - \log P(c_j) + 2 \log P(c_k)} \right) . \quad (4)$$

The Cosine similarity [13] is a measure frequently used in data mining. It is defined as the cosine of the angle between two vectors in a hyperspace. We model each term c_i as a vector $v_i = (v_{i1}, v_{i2}, \dots, v_{im})$, in which $v_{ij} = 1$ if c_i annotates e_j , and 0 otherwise. The measure is then defined as

$$Sim_{cos}(c_i, c_k) = \frac{\langle v_i, v_k \rangle}{\|v_i\| \|v_k\|} , \quad (5)$$

where $\langle v_i, v_k \rangle$ is the dot product of vectors v_i and v_k and $\|v_i\|$ is the length of v_i .

2.3 Evaluation

In order to measure how well an ontology’s usage correlates with its definition, we measure the correlation between how the gene products are annotated with terms (via the similarity measures in Section 2.2) and the terms as they are defined in the ontology. Formally, for each pair of terms (c_i, c_j) , we measure their distance in the ontology DAG. We then sort all term pairs in descending order (greatest distance first) and put them into a sorted list L_{DAG} . We then measure the similarity between each pair of terms via the similarity measures in Section 2.2, sort the term pairs in ascending order (lowest similarity first) and put them into a sorted list $L_{measure}$, where the measure is Resnik’s, Lin’s, Jiang’s, Rel or Cosine. Finally, we measure the correlation between

the two sorted lists L_{DAG} and $L_{measure}$ using Kendall's τ coefficient [14].

The basic τ method requires all values in the ranked lists to be unique, which cannot be guaranteed in our problem setting. Therefore, we make a common modification [15] to the basic method as follows. Let L_1 and L_2 be the two (equal-length) lists that we are comparing. Let $\ell_1^i \in L_1$ be the i th element in L_1 , and $\ell_2^i \in L_2$ be the i th element in L_2 . Similarly define ℓ_1^j and ℓ_2^j for $j \neq i$. Now consider each pair of pairs $((\ell_1^i, \ell_2^i), (\ell_1^j, \ell_2^j))$ for $i \neq j$. We say that this pair is *concordant* if $\ell_1^i > \ell_1^j$ and $\ell_2^i > \ell_2^j$ or $\ell_1^i < \ell_1^j$ and $\ell_2^i < \ell_2^j$. The pair is *discordant* if $\ell_1^i > \ell_1^j$ and $\ell_2^i < \ell_2^j$ or $\ell_1^i < \ell_1^j$ and $\ell_2^i > \ell_2^j$. (Note that all inequalities are strict.) Now let n_c be the number of concordant pairs, and n_d be the number of discordant pairs. Finally, let n_1 be the number of ties among elements of L_1 and n_2 be the number of ties among elements of L_2 . Then the τ coefficient is defined as:

$$\tau(L_1, L_2) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_1)(n_c + n_d + n_2)}}. \quad (6)$$

The τ coefficient ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation).

3. Generating Synthetic Data

Before we apply our correlation technique to real ontological data, we must first determine what τ values we should expect if an ontology's application to annotating gene products in fact does reflect its definition, under each similarity measure of Section 2.2. Thus we generated pairs (e_i, C_i) , where e_i is a synthetic gene product and C_i is its simulated annotation set, i.e. each term $c_j \in C_i$ annotates gene product e_i . The synthetic data has various properties, which we use to characterize the similarity measures.

Let $G = (V, E)$ be the ontology DAG and $m = |V|$. For simplicity, we assume G to be a complete tree of depth d and branching factor k . The synthetic annotation data was generated using the following randomized process on G . For each of the n distinct gene products, we select one term as the first term according to a predetermined initial distribution ω_0 . The annotation data set is then generated using three parameters n , r , and γ as follows.

- 1) Choose a initial distribution $\omega_0 = \{P_0(c_1), P_0(c_2), P_0(c_3), \dots, P_0(c_m)\}$ over terms $C = \{c_1, c_2, c_3, \dots, c_m\}$. We will examine the distribution ω_0 in Section 4.
- 2) Randomly choose a starting term $s_i \in C$ according to ω_0 for each of the n synthesized gene products e_i .
- 3) Let D be the all-pairs shortest path matrix on the ontology DAG G , where D_{ij} is the number of steps needed to reach c_j from c_i . For each s_i , generate a distribution Q_i over C , where the probability for each term decreases exponentially with its distance to s_i , i.e. $Q_i(c_j) = \gamma^{D_{ij}}$.

- 4) Choose r terms from C according to Q_i , and add them to C_i . For each c_j chosen, add all of its ancestors to C_i .

4. Result and Discussion

4.1 Synthetic Data: Parameter Sensitivity Analysis

To observe how the parameters of Section 3 influence correlation, we start by choosing ω_0 to be the uniform distribution. Thus each starting term was chosen uniformly from the ontology DAG. Twenty sets of annotations were generated for each configuration of (n, r, γ) on a complete binary tree of depth 7. We evaluated the mean values of the correlation between L_{DAG} defined in Section 2.3 and the sorted list for each measure, which are $\tau(L_{DAG}, L_{Lin})$, $\tau(L_{DAG}, L_{Resnik})$, $\tau(L_{DAG}, L_{Rel})$, $\tau(L_{DAG}, L_{Jiang})$ and $\tau(L_{DAG}, L_{Cos})$ on various configurations of parameter values.

Figure 1 shows the the average τ for a variable number n of gene products using $r = 15$ and $\gamma = 0.6$. In Figure 1, the average correlation for Cosine increases with increasing n (the number of annotations), while the four other measures are not affected by n . Also, we notice that when $n > 170$, further increase of n will not increase τ for any measure very much.

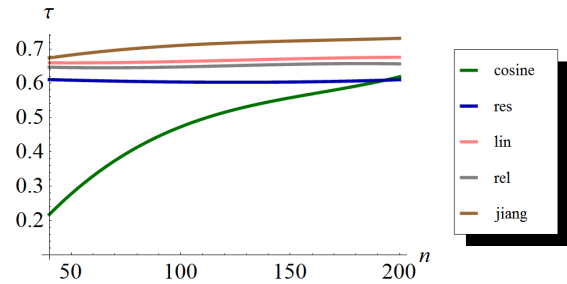


Fig. 1: Average τ of each similarity measure with respect to n the number of distinct gene product when fixing r and γ ($n \in [40, 200]$, $r = 15$, $\gamma = 0.6$).

Figure 2 shows the results for variable γ when $n = 200$ and $r = 8$. For $\gamma < 0.65$, the correlation for Jiang's measure decreases with growing γ . In contrast, τ for Cosine increases with growing γ . Also, the change of γ does not influence the correlation for other three measures. When $\gamma > 0.65$, τ for every measure begins to decrease with increasing γ , especially for Cosine, which decreases dramatically.

In Figure 3, we chose a moderate $\gamma = 0.6$ and sufficiently large $n = 200$ to examine the trend in the values of r . Similar to the results in Figure 1, correlations for Resnik's, Lin's, and Rel change little with increasing r , Jiang's decreases slightly, and the correlation for Cosine increases significantly.

From the three figures, we can see that γ affects τ of all similarity measures, though less so for Lin's, Rel, and

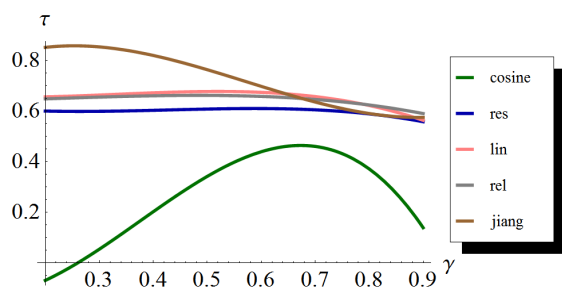


Fig. 2: Average τ of each similarity measure with respect to γ when fixing n and r ($n = 200, r = 8, \gamma \in [0.2, 0.9]$).

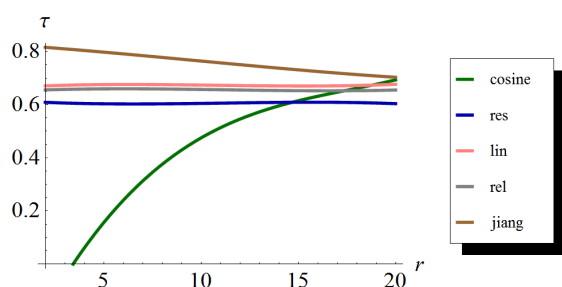


Fig. 3: Average τ of each similarity measure with respect to r the number of terms associated with each gene product when fixing n and γ ($n = 200, r \in [2, 20], \gamma = 0.6$).

Resnik's. A gene product can be associated with a number of distinct terms, and γ defines how sparse the annotation of a gene product is distributed in the ontology. A small γ indicates that the gene product has been annotated by several terms close each other. Results show that Cosine correlates more when $\gamma \approx 0.65$ while the correlation for the other four increases when γ is low.

The parameter r defines the number of terms assigned to a gene product. Higher r indicates that an individual gene product receives more annotations. This parameter affects Cosine significantly: its correlation goes high with increasing r . In contrast, Resnik's, Lin's and Rel show a very slight decrease when r increases, though they are still quite stable.

In contrast to γ and r , the number of gene products n has limited influence on the correlation. Generally, higher τ can be obtained for all measures when more annotations are made. However, as long as there is a sufficient number of annotation records ($n > 170$), further increase brings only a slight increase to the correlation.

From these results we see that Cosine is only suited for evenly annotated data with moderate $\gamma \approx 0.65$ and high r , which means each gene product is annotated by many moderately concentrated terms in the ontology. Jiang's measure is best suited for data with low γ and r , which means each gene product is annotated by very few closely related terms in the ontology. Also, we found that Resnik's,

Lin's and Rel are almost independent of the three parameters.

4.2 Synthetic Data: Geometrically Distributed Number of Annotations

We now modify the synthetic data generation model to be more realistic. When an ontology is used in practice, the terms commonly used often come from a relatively small subset of the entire set of terms. As an example, refer to Figure 4, which shows that in the database UniProtKB/Swiss_Prot, 40% of the gene products are annotated by at most two GO terms, and less than 10% of gene products receive annotation from more than 5 terms. On average, there are five terms used to annotate each gene product. Thus, in our updated model, we let r (the number of terms annotating a gene product) vary among the gene products. Based on Figure 4, we assume the number of terms follows a geometric distribution with parameter p , which is the probability that a randomly selected gene product is annotated by a single term. (So a smaller value of p results in a longer tail.) Figure 4 suggests a value of p between 0.35 and 0.50.

Ten sets of annotations were generated on each configuration of $n = 100, \gamma = 0.3$ and p , whose values ranged from 0.1 to 0.9, on a complete binary tree of depth 7. In Figure 5, we show the average value of τ that resulted from running our experiments for variable values of p . The figure suggests that larger values of p tend to increase the correlation for all measures, except for Cosine (which decreases) and Resnik's (which is the most stable of all). The correlation of Jiang's increases dramatically with p .

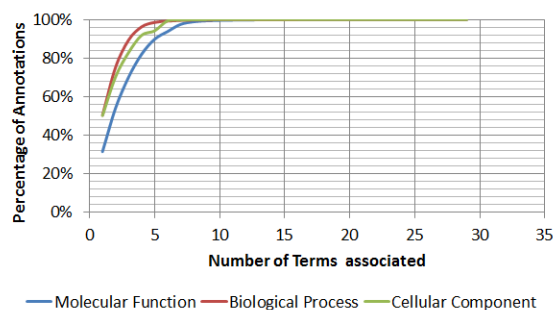


Fig. 4: Percentage of gene products annotated in GO versus number of terms used to annotate them.

The second variation we made over the experiments of Section 4.1 is in the distribution ω_0 . Our results in Section 4.1 used a uniform distribution for initial distribution ω_0 . We now examine the effect of nonuniformity of the ω_0 on the τ correlation coefficient for each similarity measure using skewed ω_0 , where nonuniformity is measured by the

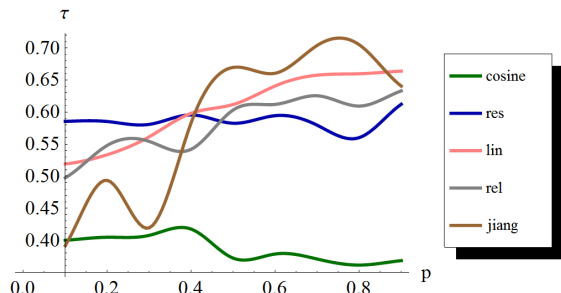


Fig. 5: Average value of τ based on variable number of annotations r geometrically distributed with parameter p ($n = 100$, $\gamma = 0.3$).

normalized entropy H_0 :

$$H_0(\omega_0) = \frac{H(\omega_0)}{H_{max}} = \frac{-\sum_{i=1}^m P(c_i) \log_2 P(c_i)}{\log_2 m}.$$

Two hundred sets of annotations were generated from the configuration $n = 200$, $\gamma = 0.6$ and $r = 2$. In each set, we chose m values at random from $[0, 1]$ according to an exponential distribution with parameter $\lambda \in [0.5, 10]$ and then normalized them to get ω_0 . Figure 6 shows the impact of ω_0 's normalized entropy on τ . We can see that increasing H_0 (making ω_0 more uniform) generally increases the correlation of all five measures, though Resnik's and Lin's are fairly stable. In particular, Cosine and Jiang's increase dramatically with increasing H_0 .

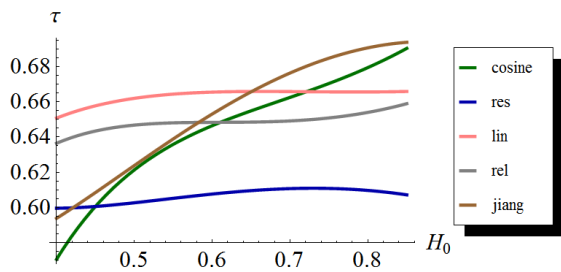


Fig. 6: Average value of τ versus the normalized entropy H_0 of the starting distribution ω_0 ($n = 200$, $\gamma = 0.6$, $r = 5$).

From these results we can see that Cosine and Jiang's are not well suited for skewed data (with a low-entropy ω_0), and Cosine is not well suited for data with a short tail (high p value). Also, unlike Cosine and Jiang's, the correlation values of Resnik's, Lin's and Rel (especially Resnik's) are more stable across many parameter values.

4.3 Real Data: Partial Ontology

We empirically compared Rel, Cosine, Resnik's, Lin's, and Jiang's similarity measures using annotations from UniProtKB [16] with a corresponding sub-ontology from

Table 1: Comparison of τ on "GO:0005275"

Measure	UniProtKB/Prot	UniProtKB
Cos	0.424	0.319
Resnik	0.596	0.576
Lin	0.621	0.602
Rel	0.618	0.630
Jiang	0.441	0.480
Terms	17	25
Genes	895	25105
Annotations	907	25593

GO. We used a subset of 25593 annotations along with the subtree from GO, rooted at the term "GO:0005275: amine transmembrane transporter activity." This annotation set consists of 25105 identified genes and contains 25 unique terms. UniProtKB is comprised of two sections, UniProtKB/Swiss_Prot and UniProtKB/TrEMBL. UniProtKB/Swiss_Prot contains curated annotations while UniProtKB/TrEMBL contains entries with computationally analyzed annotations generated by automatic procedures. These are not reviewed and curated by an author. Thus, UniProtKB/Swiss_Prot may have data of higher quality than UniProtKB/TrEMBL. Note that 98% of the records are electronically annotated. We first computed correlations using only UniProtKB/Swiss_Prot, then using the entire set (UniProtKB).

The electronic annotations in UniProtKB/TrEMBL have many gene products that are each annotated by a single term. Further, the annotation in UniProtKB/TrEMBL contains only a subset of GO terms and is significantly larger than UniProtKB/Swiss_Prot. Thus, in Table 1 we see that Cosine's correlation decreased dramatically while only Rel and Jiang's have slightly improved correlation when switching from UniProtKB/Swiss_Prot to UniProtKB. Since Resnik's, Lin's and Jiang's are almost immune to changes in parameter values (according to Section 4.2), we can use their correlations from our tests on synthetic data as a baseline for our experiments here. The $\tau \approx 0.6$ for these three measures from Table 1 is very close to the baseline suggested by Figures 1–3. This leads us to believe that this partial ontology correlates well to its usage.

4.4 Real Data: Full Ontology

Our experiment on the full ontology was performed on a copy of GO annotations dated April 2010, which consisted of 32651844 annotations of 6729320 gene products using terms from three ontologies (see Table 2). There are 43645 is_a relations defined over the 26664 terms. From the table we see that the three ontologies differ in size. The Biological Process ontology is much larger than the other two. Also, the table shows that more than one third of the terms are defined but have never been used. For Biological Process, almost half are unused.

We studied each of GO's three ontologies by computing

Table 2: Number of terms and relations for each GO ontology. Numbers exclude obsolete terms. “Active” refers to terms that have been used at least once. “Relations” refers to is_a relations.

Ontology	Terms		Relations
	Total	Active	
Cellular Component	2626	1653	3992
Molecular Function	8659	5885	10132
Biological Process	18005	9497	29521

the Kendall τ rank correlation coefficient for every pair of measures in Section 2.2 as well as the ontology DAG distance D . In order to compute τ for m terms, we would need to compute the sorted similarity measure list on all $\binom{m}{2}$ term pairs. Thus the algorithm for computing the Kendall τ rank correlation coefficient in our case has a complexity of $\Theta(m^4 \log(m))$ [17]. Given that the number of terms ranges from 1653 to 9497 (Table 2), it is infeasible to evaluate τ directly. Instead, we estimate τ by uniformly randomly sampling term pairs from the list. In order to do so, each time we sample 1000 term pairs from the list and compute τ_i , and then repeat this sampling process 50 times. We estimate τ as the mean of τ_1, \dots, τ_{50} . Since the standard deviation of τ_1, \dots, τ_{50} between each measure was < 0.01 , we consider the mean to be a good estimate.

Tables 3–5 present the τ values for each pair of similarity measures for each of the three ontologies. The first column of each table shows the correlations between DAG distance and the five measures. Res, Lin, Rel and Jiang each correlate with DAG at about the same values, while Cosine only shows a weak correlation. Also, we noticed that the first four are highly correlated with each other, especially Jiang vs. Lin and Res vs. Rel, which correlate near 0.99. This is unsurprising given the relationships among the definitions of these measures.

Table 3: Estimated τ between similarity measures on Cellular Component.

	DAG	Cos	Jiang	Rel	Lin
Res	0.44	0.25	0.85	0.99	0.83
Lin	0.40	0.45	0.98	0.83	
Rel	0.44	0.25	0.84		
Jiang	0.40	0.43			
Cos	0.23				

Table 4: Estimated τ between similarity measures on Molecular Function.

	DAG	Cos	Jiang	Rel	Lin
Res	0.40	0.20	0.90	0.99	0.89
Lin	0.37	0.33	0.99	0.89	
Rel	0.40	0.20	0.90		
Jiang	0.38	0.32			
Cos	0.19				

Table 5: Estimated τ between similarity measures on Biological Process.

	DAG	Cos	Jiang	Rel	Lin
Res	0.37	0.25	0.96	0.99	0.96
Lin	0.37	0.29	0.99	0.95	
Rel	0.37	0.25	0.96		
Jiang	0.37	0.29			
Cos	0.24				

From Section 4.1, we understand how values for n , r , γ , p , and $H_0(\omega_0)$ for an ontology and its annotations affect correlation values for the similarity measures we use. The values of n , r , and p are directly estimated from the data. However, it is not obvious how to directly estimate γ and $H_0(\omega_0)$ from the data. But if we look at $H_0(\omega)$ (the normalized entropy of the final distribution over the terms), we find that it is generally low. From this we estimate that both $H_0(\omega_0)$ (the normalized entropy of the initial distribution) and γ are generally low in the real data. Specifically, we use $H_0(\omega)$ as an upper bound of $H_0(\omega_0)$. Table 6 shows values of the relevant parameters in GO; γ is omitted and instead is qualitatively estimated as “low”, since Table 6 gives $H_0(\omega)$ as relatively low, ranging from 0.44 to 0.58.

Table 6: Corresponding parameters for each ontology.

Ontology	n	r	p	$H_0(\omega)$
Molecular Function	5860336	2.85	0.35	0.58
Cellular Component	3217382	2.13	0.47	0.44
Biological Process	5127003	1.94	0.52	0.55

Since increasing n beyond a sufficient number (170 in synthetic data) brings only minimal changes in correlation, we expect n will have little effect on correlation values even though it is four orders of magnitude higher than the values used in our synthetic data. The $\tau \approx 0.2$ for Cosine in GO lies in the interval $[0.1, 0.4]$ that is suggested by Figures 3 and 5 for synthetic data of similar characteristics.

Table 6 gives low $H_0(\omega)$ from 0.44 to 0.58, which suggests that both γ and $H_0(\omega_0)$ are low. The $\tau \approx 0.39$ for Jiang’s is low compared to either 0.8 given by low γ in Figure 2, 0.45 given by $p \approx 0.25$ in Figure 5 or 0.6 given by $H_0(\omega_0)$ around 0.4 in Figure 6.

In addition, the average $\tau \in [0.37, 0.44]$ for Resnik’s, Lin’s and Rel are low compared with those from the synthetic data and GO:0005275, where similar configurations show that correlations around 0.6 are possible (and very stable in the case of Resnik’s). All these results suggest that GO’s use correlates less with its definition compared to GO:0005275, though more experimentation should be performed to confirm this.

5. Conclusion

The Gene Ontology (GO) terms are widely used to annotate gene products. However, it is unknown whether

the terms defined in GO are used to label gene products in a manner consistent with their definition. Since there are many ways to measure semantic similarity, we first used various synthetic data models to study several similarity measures to characterize their sensitivity to various properties of the data. We found that Cosine is only suitable for annotation sets that have with long tails (low p values) and in which each term is annotated by many moderately concentrated terms in the ontology. Jiang's measure is not well suited for skewed data (with a low-entropy ω_0) and in which each gene product is annotated by very few closely related terms in the ontology. Also, we found that Resnik's, Lin's and Rel are almost independent of the these parameters, especially Resnik's.

Then we investigated a small sub-ontology and its annotations of data from UniProtKB and found that Rel, Resnik's and Jiang's measures indicate correlations between the DAG and its application relative to what seems to be the best possible based on tests on synthetic data. Thus we conclude that this partial ontology's definition relates well to its usage.

Finally, from our preliminary result on the full GO ontologies, we found that correlation results using the more stable measures (especially Resnik's) seem to indicate that the correlation between GO's use and its definition is low, especially when compared to the correlation between GO:0005275 and UniProtKB. More experimentation should be performed to confirm this.

In addition to a more detailed analysis, future work includes examining other measures that evaluate semantic similarity, and characterizing them based on synthetic data parameters as we did with those of this paper. This might reveal measures that are even less sensitive to the parameter values and might in turn be even more useful for studying real data.

Our synthetic data model was based on complete binary trees that were not similar to the DAGs in GO. Thus it is possible that the trends observed in our synthetic data results might not reflect what we would see in a full ontology. Therefore, in our ongoing work, we randomly selected 100 terms from GO, each with around 100 child terms, yielding 100 subDAGs, each of size approximately 100. We then measured the sensitivity of each similarity measure's τ value to the five parameters by repeating the tests of Section 4.1 on each of the 100 subDAGs. Our preliminary results show that Resnik's measure remained almost invariant to changes in parameter values when the subDAG remains unchanged. However, Resnik's τ value was sensitive to the topology of the subDAG. In our continued research, we will further investigate this, attempting to correlate the similarity measures' τ values to properties of the subDAGs, such as branching factor, depth, diameter, and skewness.

Acknowledgments

The authors thank the anonymous reviewers for their comments. This research was supported by National Science Foundation grant number 0743783.

References

- [1] M. Ashburner, C. A. Ball, and J. A. Blake, "Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25–29, 2000.
- [2] L. Stein, "Genome annotation: From sequence to biology," vol. 2, pp. 493–503, 2001.
- [3] K. Verspoor, K. B. C. D. Dvorkin, and L. Hunter, "Ontology quality assurance through analysis of term transformations," *Bioinformatics*, vol. 25, pp. 77–84, 2009.
- [4] P. W. Lord, "Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation," *Bioinformatics*, vol. 19, pp. 1275–1283, 2003.
- [5] A. Schlicker, F. S. Domingues, J. Rahnenfuhrer, and T. Lengauer, "A new measure for functional similarity of gene products based on Gene Ontology, including indels," *BMC Bioinformatics*, vol. 7, p. 302, 2006.
- [6] M. Mistry and P. Pavlidis, "Gene Ontology term overlap as a measure of gene functional similarity," *BMC Bioinformatics*, vol. 9, p. 327, 2008.
- [7] A. Schlicker, C. Huthmacher, F. Ramirez, T. Lengauer, and M. Albrecht, "Functional evaluation of domain-domain interactions and human protein interaction networks," *Bioinformatics*, vol. 23, no. 7, pp. 859–865, 2007.
- [8] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto, "Metrics for GO based protein semantic similarity: A systematic evaluation," *BMC Bioinformatics*, vol. 9, no. Suppl 5, p. S4, 2008.
- [9] "GO annotation policies and guidelines." [Online]. Available: <http://www.geneontology.org/GO.annotation.shtml>
- [10] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.
- [11] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning*, 1998, pp. 296–304.
- [12] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of ROCLING X*, 1997, p. 9008.
- [13] M. Popescu, J. M. Keller, and J. A. Mitchell, "Fuzzy measures on the Gene Ontology for gene product similarity," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 3, pp. 263–274, 2006.
- [14] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, 1938.
- [15] "Kendall Rank Correlation. Wolfram Mathematica." [Online]. Available: <http://reference.wolfram.com/mathematica/MultivariateStatistics/ref/KendallRankCorrelation.html>
- [16] E. Jain, A. Bairoch, and S. Duvaud, "Infrastructure for the life sciences: Design and implementation of the UniProt website," *BMC Bioinformatics*, vol. 10, p. 136, 2009.
- [17] D. Christensen, "Fast algorithms for the calculation of Kendall's τ ," *Computational Statistics*, vol. 20, pp. 51–62, 2005.

Phylogenetic analysis workflow using the BioExtract Server

Yosr Bouhlal¹, Douglas M Jennewein¹ and Carol Lushbough¹

¹Computer Science Department, University of South Dakota, Vermillion SD, USA

Abstract - Several molecular and genetic tools, browsers and servers are currently available online for biologists to access, analyze and process data. The BioExtract Server represents a powerful web-based data integration application, combining the most common biological databases with the most used algorithms by scientists from all the biological fields. This Server allows researchers to extract data, execute local and web-accessible analytic tools and create customized workflows. Each workflow can be used for different similar queries and offers an easy access to the results of all the executed tools at once. This paper describes a BioExtract workflow providing a simple phylogenetic analysis, as one of the numerous applications that the BioExtract Server offers to biologist researchers.

Keywords: BioExtract Server, workflow, phylogenetic analysis

1 Introduction

The study of genome evolution involves a global comparative approach in which individual genetic events are considered and integrated in their evolutionary context, which in turn may be correlated to the population history, the environment and the different phonemes [1]. Many tools and techniques are currently used to study evolution and infer the evolutionary relationship between species and organisms. These techniques include morphology, anatomy, paleontology, physiology and molecular phylogeny [2].

Phylogeny based analysis provides an ideal framework for performing such investigations, by pinpointing when a genetic event occurred and by identifying the simultaneous occurrence of several events [1]. There are principally five stages in the molecular phylogenetic analysis [2]. The first stage is the acquisition of the sequence which can be performed through many sources including Genbank or HomoloGene gene databases, Rfam for RNA, Pfam for proteins or ICTV for viruses. Once sequences are acquired, a multiple sequence alignment will be performed on homologous sequences. This stage is considered a critical step of phylogenetic analysis subject to many important considerations. The next stages will be the specification of a statistical model of nucleotide or amino acid evolution, the construction of the evolution tree, and finally the interpretation of the generated tree [2]. Among an important number of online tools and servers, the BioExtract Server

(<http://bioextract.org>) is a powerful Web-based data integration application that can be used to help researchers accomplish all these phylogenetic analysis steps. The BioExtract Server was designed to help scientists consolidate, analyze, and serve data from heterogeneous bio-molecular databases [3]. It allows them to query multiple data sources, save query results as searchable data sets, execute local and Web-accessible analytic tools, and create computational customized workflows [3, 4].

We describe here a simple BioExtract Server workflow that can be used for standard phylogenetic analysis starting from a protein sequence query.

2 Methods

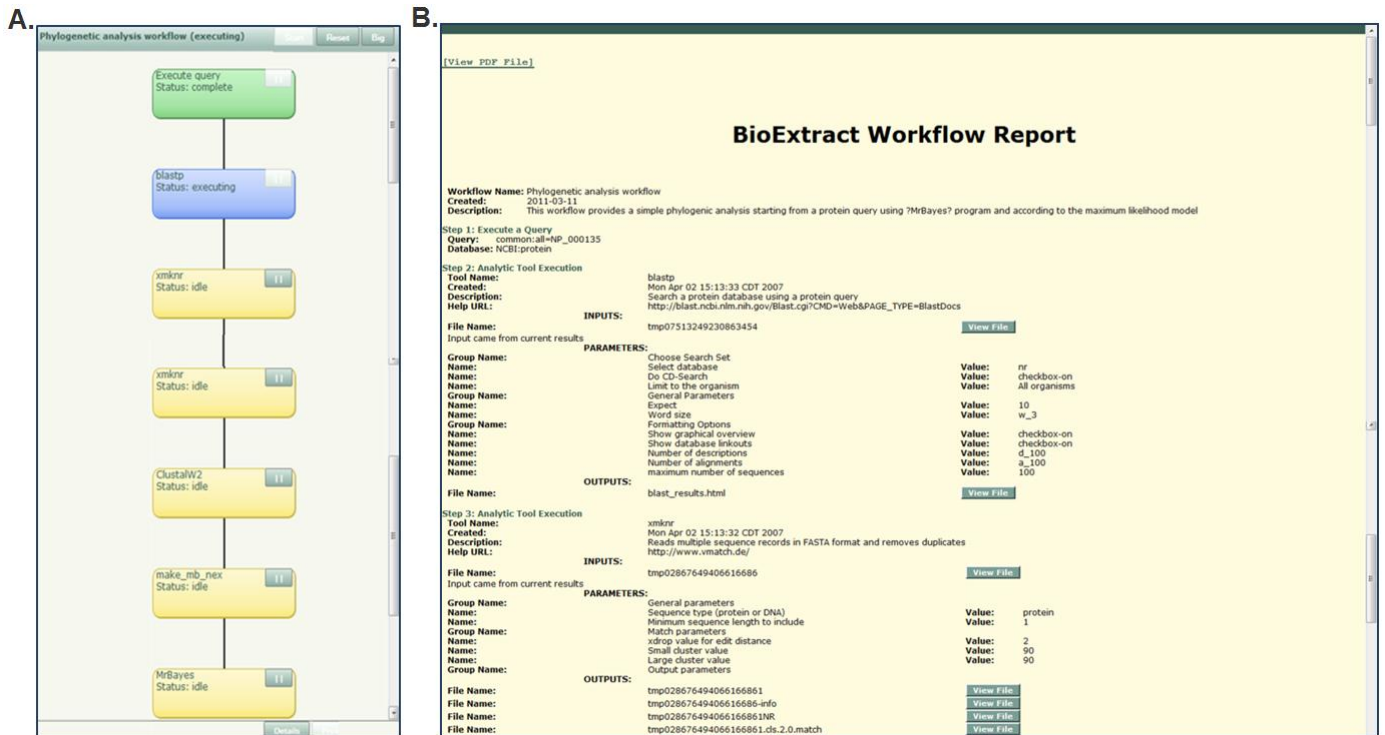
The BioExtract Server was used to create a workflow for comparing and aligning a number of nucleotide sequences to build a phylogenetic evolutionary tree (Figure 1A). The workflow covers five steps using five different molecular online tools (Table 1).

Table1. Tools used in the phylogenetic workflow

Tool	Common Link	Description	Ref.
Blastp	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Search protein database using a protein query	[5]
xmknr	http://vmatch.de	Reads multiple sequence records in FASTA format and removes duplicates	[6]
ClustalW	http://clustal.org	Computes a multiple sequence alignment for Protein or DNA sequences	[7]
make_mb_nex	http://bioextract.org/	Creates a MrBayes nexus file from a clustal alignment file	
MrBayes	http://mrbayes.csit.fsu.edu/	Estimate phylogeny upon Bayesian inference which is based on the probability of a tree conditioned on the observations.	[8]

The query sequence can be selected from the common or specific databases available through the BioExtract Server or simply uploaded from a private source. When executing the workflow, similar sequences will be extracted by the “Blastp” tool. Duplicate sequences will be then removed using “xmknr” tool, a simple shell script utilizing the Vmatch tool. Users can further refine the Blastp results by selecting sequences according to the length or the E score through the “extract page” before running the next step. Once selected, sequences will be aligned using the “ClustalW” multiple sequences alignment program.

Figure 1. (A) BioExtract Server workflow created for the phylogenetic analysis (B) The first three steps of the workflow showed on the general report



In order to perform the phylogenetic analysis for the remaining aligned sequences, we developed a new tool “make_mb_nex” and included it within the BioExtract Server tools page.

This tool will create a nexus file from an alignment file. The user can configure the created nexus file by specifying the appropriate evolutionary model and the MCMC (Markov chain Monte Carlo) algorithm parameters. For this study, Following parameter settings were used: set nst=6 rates=invgamma [according to the General Time Model GTM]; mcmc ngen=1000; samplefreq=10; sump burnin=25 and sumt burnin=25. Finally, the generated nexus file is executed on the “MrBayes” program according to the maximum likelihood model [9] and the evolutionary tree is drawn.

In order to test the feasibility and usefulness of this workflow, the human Frataxin protein sequence (variant 1: NP_000135) was used as the initial query.

3 Results

Human Frataxin protein is a mitochondrial protein encoded by the FXN gene and seems to be implicated in the iron-sulfur clusters. Reduced or modified frataxin causes Freidreich’s ataxia, an autosomal recessive neurodegenerative disorder. Alternative splicing results in multiple transcripts variants [10]. The variant 1 [NP_000135] was used as an input query to run the BioExtract Server phylogenetic analysis workflow.

The execution of the workflow led to the extraction of 100 sequences homologous to the frataxin protein. After duplicates were excluded, multiple sequence alignments are performed for all the sequences. Once poorly aligned sequences are removed, the corresponding phylogenetic

tree is estimated using a Bayesian method based on a general time reversible (GTR) model.

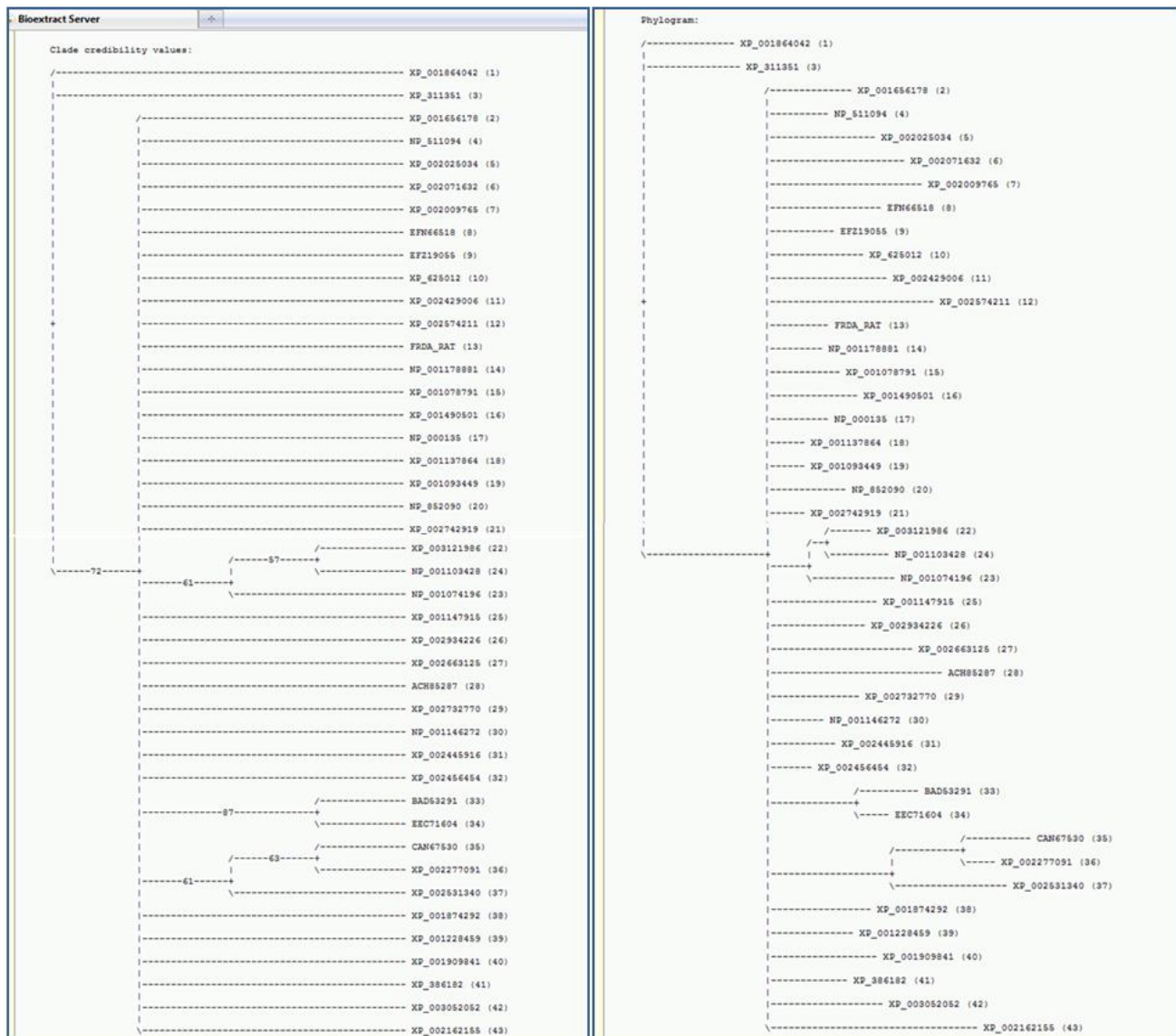
All the results are summarized in a general report (Figure 1B). Each output can be visualized or uploaded by clicking on “View File” corresponding to each tool. As an example of the workflow output, the resulting phylogenetic tree and the credibility values are shown on Figure 2.

This workflow is shared on the “MyExperiment” portal and is accessible through the following link <http://www.myexperiment.org/workflows/1941.html>.

4 Discussion

The BioExtract Server is a Web-based system designed to aid researchers in the analysis of distributed genomic data by providing a platform to facilitate the creation of bioinformatic workflows [4]. The basic operations of the BioExtract server allow researchers via their Web browsers to: specify data sources; flexibly query data sources with a range of relational operators; apply analytic tools; download result sets; and store query results for later reuse. As the researcher works with the system, their “steps” are saved in the background. At any time these steps can be saved as a workflow simply by providing a name and description. Once saved, these workflows can be executed and/or modified [3]. The execution of any created workflow generates the running of all the tools at once, and provides access to all the results via the general workflow report. Consequently, the results are obtained in an extremely reduced time compared to conventional methods. In addition, the results are recorded in the workflow and can be easily retrieved from the server when needed.

The workflow presented in this paper provides a simple phylogenetic analysis starting from a protein query. Users can

Figure 2. Clade credibility values and Phylogram

modify the query by simply changing the accession number on the workflow's query step. Similar workflows can be created to analyse DNA or RNA sequences by modifying the query database on the first step and replacing Blastp with Blastn in the second step of the workflow. The two first steps of the workflow can also be eliminated if the user needs to directly upload his or her own aligned sequences. Several other enhancements can be added to the phylogeny analysis workflow by adding additional phylogenetic tools and packages available through the BioExtract Server such as "PAUP", "dnadist", "propars" and "MrModelist".

This workflow represents one of numerous applications that the BioExtract Server offers to biologist researchers. Containing a large cluster of tools and giving access to numerous databases, the BioExtract Server can be used for genomic and protein annotation, sequence mutation analysis, gene or protein function prediction and many other complex molecular and genetic analyses. Some of these applications are actually shared on the "MyExperiment" website [<http://www.myexperiment.org/>] where they can be easily launched and used.

Various enhancements to the BioExtract Server are under development that, when added, will broaden the spectrum of users by adding more tools and databases which could be used for many additional biological fields.

5 References

- [1] Gouret P, Thompson JD, Pontarotti P. PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinformatics*. 2009 Sep 19; 10:298.
- [2] Jonathan Pevsner. *Bioinformatics and Functional Genomics* (second edition). Wiley-Blackwell 2009: 215-69.
- [3] Lushbough C, Bergman MK, Lawrence CJ, Jennewein D, Brendel V. BioExtract server--an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data. *IEEE/ACM Trans Comput Biol Bioinform*. 2010 Jan-Mar; 7(1):12-24.

- [4] Lushbough CM, Brendel VP. An overview of the BioExtract Server: a distributed, Web-based system for genomic analysis. *Adv Exp Med Biol.* 2010; 680:361-9.
- [5] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.*, 1990; 215(3):403-10.
- [6] Abouelhoda MI, Kurtz S, Ohlebusch E. Replacing suffix trees with enhanced suffix arrays. *J Discrete Algorithms* 2004; 2: 53–86.
- [7] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 1994; 22[22]:4673-80.
- [8] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 2001 Aug; 17(8):754-5.
- [9] Evolutionary trees from DNA sequences: a maximum likelihood approach. Felsenstein J. *J Mol Evol.* 1981; 17(6):368-76.
- [10] Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F et al. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science.* 1996 Mar 8; 271(5254):1423-7.

SESSION

COMPARATIVE SEQUENCE, GENOME ANALYSIS, GENOME ASSEMBLY, AND GENOME SCALE COMPUTATIONAL METHODS

Chair(s)

TBA

A Regression-based Approach for Estimating Recombination Rate from Population Genomic Data

Lan Zhu^a, Feng Feng^b, Carlos D. Bustamante^c

^aDepartment of Statistics, Oklahoma State University, Stillwater, OK 74078

^bDepartment of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705

^cDepartment of Genetics, Stanford University School of Medicine, Stanford, CA 94305

ABSTRACT

Motivation: Recently, much attention has focused on using prediction from population genetic theory to quantify variation in recombination rate along the human genome owing to the promise of association or linkage disequilibrium(LD) mapping to identify genes underlying complex traits. Current state of the art approaches to the problem estimate the local population recombination rate from patterns of LD among common single nucleotide polymorphisms(SNPs) assuming the population is randomly mating and constant in size.

Results: Here we describe an alternative method that can accommodate complex population structure and ascertainment bias. Using multiple linear regression and non-parametric bootstrap re-sampling, our method uses the variances and co-variances of un-phased SNPs at different frequencies to estimate the local recombination rate. We evaluate this new approach via Monte Carlo simulation and compare its performance with three other available methods. Our approach is less biased when the demographic assumptions of the standard neutral model are violated. We also apply our approach to the well-characterized hot spots near the human TAP2 gene and a 206-kb region on human chromosome 1q42.3 near minisatellite MS32. The results are consistent with findings in literatures.

Keywords: Recombination, Regression, Linkage Disequilibrium

Contact: lan.zhu@okstate.edu

1 INTRODUCTION

Understanding how and why recombination rates vary along a genome is a fundamental problem in genomics. From an evolutionary perspective, recombination is a rich source of novel variation and a potent force that can lead to gametic associations among positively selected mutations as well as break up associations among deleterious mutations. Recombination rates also vary dramatically among genomes with some, such as *Drosophila*, showing no clear fine-scale structure while others, such as humans, showing a great deal of local variation where regions of low to moderate recombination are punctuated by short 1-2 kb hotspots of meiotic exchange that can account for 50–80% of all recombination events (McVean *et al.*, 2004; Myers *et al.*, 2005).

When the contributions of recombination and its interaction with selection to the process of evolution are shown essential (Cutter and Choi, 2010; Cutter and Moses, 2011), understanding recombination rate variation is also fundamentally important to the design of efficient methods for association mapping, since the degree of association among markers dictates the density and distribution of markers used for mapping (Noor *et al.*, 2001). Classical methods for estimating recombination rates from natural populations include pedigree studies, sperm typing analysis and methods based on

predictions from population genetics. In humans, the difficulty of obtaining large pedigrees limits the utility of pedigrees to estimation of large-scale (megabase) recombination rates (Kong *et al.*, 2002). Likewise, while sperm typing can provide accurate estimates of the local recombination rate in male gamete production, it is typically only applied to a few individuals and to only short regions of the genome (Greenawalt *et al.*, 2006). Moreover, it is very labor intensive and expensive. These limitations coupled with the increasing availability of genome-wide polymorphism data from humans and other species make estimation of recombination rates via population genetic theory an attractive alternative.

A number of estimators of the population recombination rate ($R = 4N_e r$, where r is the rate of crossing over for the region and N_e is the effective population size) are currently available, including moment-based (Hey and Wakeley, 1997; Hudson, 1985; Hudson, 1987; Wall, 2000), full maximum likelihood estimators (Fearnhead and Donnelly, 2001; Griffiths and Marjoram, 1996; Kuhner, *et al.*, 2000; Nielsen, 2000), and full-likelihood Markov chain Monte Carlo method (Wang and Rannala, 2008). From a statistical perspective, one would prefer to use full-likelihood methods, since these are guaranteed to capture the most amount of information in the data regarding recombination. However since it can take months of computer time to estimate recombination rate for even a modest size region using full-likelihood, there has been considerable effort to develop a litany of approximate likelihood estimators (Crawford *et al.*, 2004; Fearnhead and Donnelly, 2002; Fearnhead, *et al.*, 2004; Fearnhead and Smith, 2005; Haubold, *et al.*, 2010; Hudson, 2001; Jiang *et al.*, 2009; Li and Stephens, 2003; McVean, *et al.*, 2002; McVean, *et al.*, 2004). For example, the composite likelihood methods of Hudson (2001) and McVean *et al.* (2004) use pre-computation of pairwise likelihood for a given sample size to achieve speeds orders of magnitude faster than full-likelihood. Auton and McVean (2007) further constructed a pseudo-likelihood as the product of the likelihood over all pairs of SNPs in the region under consideration. To maintain the computational feasibility, SNPs separated by no more than 50 intermediate SNPs were considered to contribute to the composite likelihood.

These approaches, while quite fast, have several limitations including the need to precompute pairwise likelihood for a novel sample size or demographic model and an apparent lack of power to detect recombination hotspots that do not significantly affect linkage disequilibrium (Jeffreys, *et al.*, 2005). The methods of Fearnhead *et al.* (2004), Li and Stephens (2003), and Fearnhead and Smith (2005) appear to have excellent power to detect hotspots, but are computationally costly (e.g., according to Fearnhead and Smith (2005) it takes their method 10-30 minutes to estimate the recombination rate for a window of six SNPs with sample

size 60 sequences). It is also important to note that the effective population size is confounded within the estimate of the population recombination rate, therefore, population genetic estimators are by definition dependent on assumptions regarding the demographic history of the sample. A limitation of many of these approaches, therefore, is that they are based on the assumption that the population under study is randomly mating and constant in size - an assumption violated by nearly all populations to which the approaches are applied. In theory, population structure and demography can be built into almost any method, but for methods such as composite likelihood that make use of a great deal of pre-computation, this will require months (or years) of computer time for each new model to generate the lookup tables used in estimation.

In this paper, we present a novel statistical method for estimating the population recombination rate via coalescent simulations with recombination coupled with multiple linear regression (MLR) and non-parametric bootstrap. Three advantages of our method are that (1) it can readily accommodate complex demographic history, (2) provide prediction intervals for the estimated recombination rate, and (3) is computationally efficient and applicable to whole-genome data. Furthermore, since the method appears to weight heavily the variance of new mutations in estimating recombination rates, it may be able to detect recent changes in recombination rate that do not leave an explicit LD signal.

Our method is based on a readily discernible statistic of the data: the observed variability in the number of mutations at different frequencies across sub-samples of the data. It is important to note that the idea of using the variance of mutation counts in a sample to estimate recombination rates is not new. About two decades ago, Hudson (1987) introduced an estimator of the population recombination rate based on the variance of pairwise nucleotide differences among sequences in the sample. In 1997, Wakeley proposed an improved version of Hudson's (1987) estimator that has smaller bias and standard error. Our approach is loosely a generalization of Hudson's estimator in that we aim to use the most informative components of the frequency distribution to estimate the local recombination rate. A major advantage of this approach is that it does not require calculation of pair-wise linkage disequilibrium and, thus, does not require phasing of the data. Likewise, while our approach requires some pre-computation to fit the model, it is orders of magnitude less than existing approaches (roughly minutes to hours for our approach compared to days or weeks for composite likelihood). We investigate the accuracy of the approach using Monte Carlo simulations under a wide range of demographic models. We also compare the performance of our method to three commonly used approaches (Hey and Wakeley, 1997; Hudson, 1987; McVean, *et al.*, 2002).

2 METHODS

2.1 Data and Model

Consider a set of n aligned DNA sequences from a population with known demography Q (e.g., population of constant size, bottleneck, island migration, recent population growth, etc.) in which S sites are observed to be variable in the alignment. Let X_i for $i = 1 \dots n - 1$ represent the number of SNPs at frequency i out of n in the sample. For simplicity, here the ancestral state of each SNP is assumed known (i.e., the polarized site-frequency spectrum); a model with unknown ancestral state can be easily derived in the similar way. Across independent realizations of the evolutionary process,

X will vary stochastically so that for each component one has an associated variance V_i . For example, V_1 is the variance in the number of singletons that one would observe if one were to have rerun the evolutionary process and obtained an independent sampling of chromosomes at the same locus. Here we describe how recombination affects the variances and co-variances of the components of the SFS (SFS variances) in a fully predictable way and how by estimating SFS variances, one can predict the recombination rate of a genomic region for a given demographic model. For a given observed data set, however, one only has a single observed vector of frequencies, so we must first define what we mean by variance within components of the site-frequency spectrum.

Here, we consider the variance in X_i under two scenarios: (1) independent realizations of the evolutionary process (i.e., a variance that one can estimate only via simulation) and (2) bootstrap resampling of the sequences (i.e., a variance one can readily estimate via a common statistical readily applicable to the observed data). As we show in the results section, these two scenarios give different, but nearly perfectly correlated variances such that one may estimate the former given an observed value from the later.

First, let us assume that one was able to rerun the evolutionary process Q under the same recombination rate R so as to obtain Q replicate data sets, sampling an independent set of n sequence each time. From population genetic theory we expect variance and co-variance of the X_i s across the Q replicates to be informative about recombination (Fu, 1995; Sawyer and Hartl, 1992; Zhu and Bustamante, 2005).

For example, for a population that evolves according to the standard neutral Wright-Fisher model, Fu (1995) derived that variance and co-variances of the X_i s as a function of the population mutate rate $\theta = 4N_e\mu$ under complete linkage. Specifically, under complete linkage one can write the variance of X_i as $V_i = \text{Var}(X_i) = \theta/i + \sigma_{ii}\theta^2$, where σ_{ii} is a function of i and sample size n . Under complete independence among sites, Ewens (1972) and Sawyer and Hartl (1992) showed that X_i should be Poisson distributed with mean and variance $V_i = \theta/i$. Given these two well-known results, one might posit a monotonic decrease in the variance of V_i with increasing R so that recombination acts simply to decrease the σ_{ii} term above. (These predictions are born out in Figures 1 and 2 as explained below.) The reasoning above immediately suggests a simple and potentially powerful strategy for estimating R .

2.2 Algorithms for Estimating Recombination Rate

2.2.1 Algorithm 1: estimating recombination rate across evolutionary replicates

1. Simulate data by Hudson's ms program (Hudson, 2002) under the demographic model for Q replicates keeping the matrix of site-frequency spectra (SFS) with the Q rows representing the site-frequency spectra for independent replicates (simulations can be carried out conditional on the estimated mutation rate, θ , or on the observed number of segregating sites, S):

$$\begin{pmatrix} x_{1,1} & x_{2,1} & \dots & x_{n-1,1} \\ x_{1,2} & x_{2,2} & \dots & x_{n-1,2} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{1,Q} & x_{2,Q} & \dots & x_{n-1,Q} \end{pmatrix}$$

2. For each pair of columns i and k , calculate the column means \bar{X}_i , column variances V_i , and co-variance V_{ik} across replicates (note $V_{ii} = V_i$ in our notation above). The results of this step will constitute an $n - 1$ dimensional vector of SFS means $\bar{X} = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_{n-1}]$, where $\bar{X}_i = \sum_{j=1}^Q \frac{X_{i,j}}{Q}$ and a variance-covariance matrix with entries: $V_{ik} = \sum_{j=1}^Q \frac{(X_{i,j} - \bar{X}_i)(X_{k,j} - \bar{X}_k)}{Q}$.
3. Repeat above steps across a range of recombination rates (in practice we use $R \in \{1, 5, 10, 20, 50, 100, 200, 400, 1000, 2000\}$) so as to produce a set of predictor variables in the form of the $(n - 1)$ variance

and $\binom{n-1}{2}$ covariance entries of the variance-covariance matrices across levels of R .

4. Natural log-transform both the predictor (V_{ik} for different levels of R) and predicted variables (R).
5. Use stepwise selection or best subset methods to choose the model that is sufficient to explain the relationship among $\log(R)$ and the log of the components of the variance-covariance matrix. Formally, the full model would have $\binom{n-1}{2} + (n-1) + 1$ terms of the form:

$$\log(R_j) = \alpha + \sum_{i=1}^{n-1} \sum_{k=1}^i \beta_{ik} \log(V_{ik,j}) + e_j$$
 where $e_j \sim N(0, \sigma^2)$. In the model above, α is the intercept of the regression, β_{ik} are regression coefficients under the saturated model, and e_j are independent and identically distributed error terms for the residual variance for $j = 1 \dots J$ where J is the number of levels of recombination used to fit the model. In practice, we use stepwise selection and best subset methods to search over the space of models so as to identify the subset of β_{ik} terms that are sufficient to explain the data.
6. Check all assumptions for fitting a linear regression model, including normality, equal variance of residuals, and independence among residuals.

2.2.2 Algorithm 2: estimating recombination rate by bootstrap-based regression (BSTReg) across k -subset replicates A potential problem of applying the above method to real data is that for a given data set, one only has a single observed site-frequency spectrum, X . In order to generate estimates of the variance/covariance matrix across replicates of the evolutionary process we need to use a resampling scheme such as non-parametric bootstrap resampling of the data. Since the estimated variances under the bootstrapping procedure use correlated data, we expect estimates of V_{ik} to be affected. Therefore, we need to modify our MLR fitting procedure as follows:

1. Sample a single data set with n sequences under a demographic model of interest Θ , and label the data set q .
2. Divide the n sequences into k subsets of equal size, calculate the SFS for each subset, then modify the above step so that the mean and variance-covariance matrix are now calculated across the k site-frequency spectra.
3. Repeat this k -subset division sampling for the same n sequences for B bootstrap replicates to obtain B variance-covariance matrices.
4. Let $V_i^{(q)} = \frac{1}{B} \sum_{k=1}^B V_{i,k}$ be the average variance of component i and $Cov_{ij}^{(q)} = \frac{1}{B} \sum_{k=1}^B Cov_{ijk}$ the average covariance across the B replicates of the subsetting approach. (In practice, we use $n = 60$ and $k = 10$. If the data is unphased, resample individuals; if the data is phased, resample phased haplotypes.

Repeat steps 1-4 for Q replicate data sets to obtain the bootstrap estimated variance-covariance matrix V_{bs} for a given model Θ , where $V_{i,bs} = \frac{1}{Q} \sum_{q=1}^Q V_i^{(q)}$ and $Cov_{ij,bs} = \frac{1}{Q} \sum_{q=1}^Q Cov_{ij}^{(q)}$.

3 RESULTS

3.1 Estimating recombination rate when θ is known or S is fixed under evolutionary replication

We first consider the problem of predicting the population recombination rate from polymorphism data arising under a known demographic model. Using standard coalescent algorithms, we simulated 10,000 replicate samples for each of 10 levels of recombination rate $R \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000\}$ under a fixed mutation rate $\theta = 4N_e\mu = 30$ where μ is the regional mutation rate per chromosome. (These parameter values correspond roughly to a $30kb$ region in humans with recombination rate varying from 2.5×10^{-4} cM to $0.25cM$.) or a fixed number

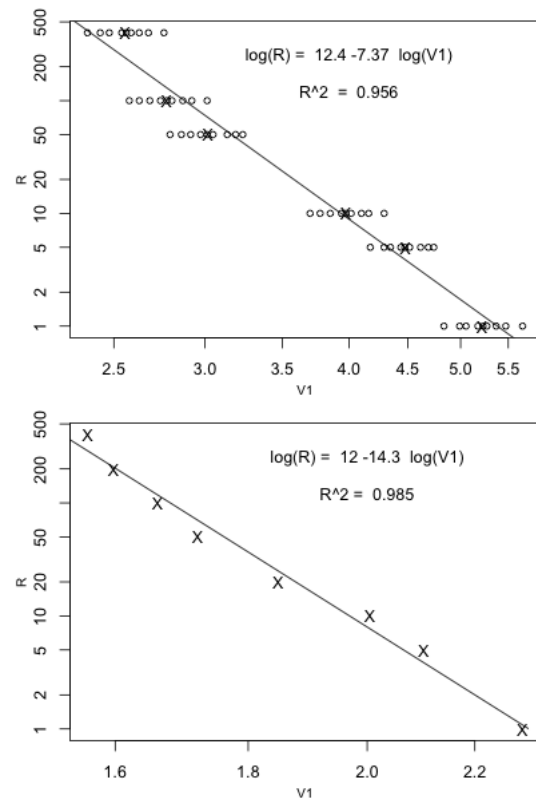


Fig. 1. Linear regression of log transformed recombination rate ($\log R$) and log transformed variance in the number of singletons in the sample. Top: 200 replicates of data sets each with sample size $n = 6$, $S = 10$ were simulated independently under the standard neutral Wright-Fisher model. Each points represents the V_1 quantiles $\{0.025, 0.10, 0.20, 0.40, 0.5, 0.60, 0.80, 0.90, 0.975\}$ corresponding to R in the range of $\{1, 5, 10, 50, 100, 400\}$. Cross signs are the means of $\log V_1$ over 200 replicates; Bottom: Linear model is fitted by $\log R$ on the average of $\log V_{1,bs}$ by k -subset bootstrap resampling over 1000 replicates.

of segregating sites $S = \{10, 20, 30, 50, 100\}$. For a given level of recombination, we calculate the vector of SFS variances $V = \{V_1, V_2, \dots, V_{n-1}\}$ across the Q replicate data sets as explained in the method description above.

When we perform the multiple linear regression of R on all V_{ik} s including all pairwise covariances among SFS components and use both stepwise selection and best subset methods, all terms are dropped except for the variance of singletons (V_1) in the model. Scatter plot of R versus the V_1 across simulated data sets shows a curvilinear relationship suggesting that linear regression of log-transformed data could be used to estimate R from a linear combination of the components in V . Using a step-wise addition rule, we find that $\log(V_1)$ alone is a sufficient predictor variable for the population recombination rate with the best fit linear regression explaining over 95% of the variance in either fixed θ or S scenarios, as shown in Figure 1 (top) when $S = 10$. Diagnostic tests (linearity, constant variance, normality, independence) for validation of the model were performed and none of the tests suggests a violation of the assumptions. (Note: for all regressions performed, diagnostic tests were checked and no violation is found,

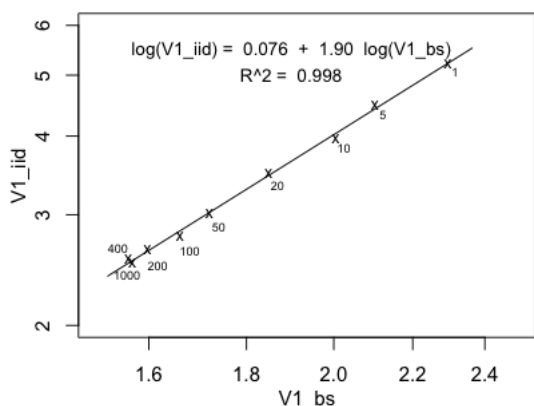


Fig. 2. Relationship between the average of bootstrap estimated variance in the number of singletons (V_{1_bs}) and that from independent sampling (V_{1_iid}). Sample size $n = 6$, $S = 10$ under the standard neutral Wright-Fisher model.

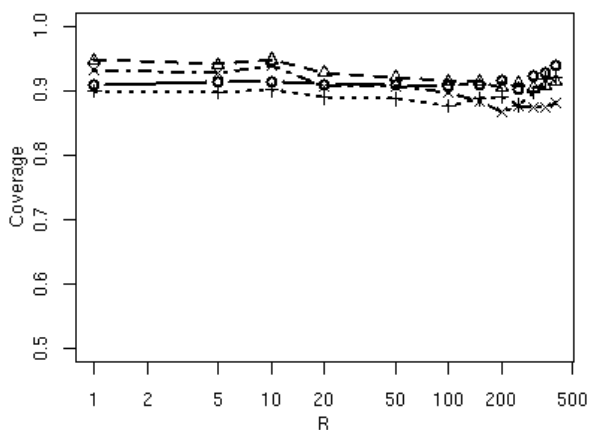


Fig. 3. Coverage of predicted local recombination rate using our bootstrap-based linear regression method with sample size $n = 60$, $S = \{10, 20\}$, $k = 10$. X-axis is plotted in log scale. Linear regression model is fitted in the range of $R = \{1, 5, 10, 20, 50, 100, 200, 1000, 2000\}$ with equation $\log(R) = 13.644 - 16.612 * \log(V_1)$ for $S = 10$ ($R^2 = 0.933$) and $\log(R) = 17.544 - 9.517 * \log(V_1)$ for $S = 20$ ($R^2 = 0.959$). Coverage is defined as the percentage of replicates that have 90% or 95% predicted intervals cover true recombination rate.

results not shown). This simple example shows that for a fixed level of the mutation rate or a fixed number of segregating sites, the transformed recombination rate and the first component of SFS variances are highly correlated. By choosing a fixed number of segregating sites in a genomic region, one can reliably predict the recombination rate for the region using the observed SFS variances across samples.

3.2 Estimating recombination using bootstrap re-sampling and k-subsetting (BSTReg)

For a real data set, however, one only has a single observed SFS vector. To estimate the SFS variances, one therefore needs to couple a re-sampling step such as non-parametric bootstrapping

Table 1. Multiple linear regression output for estimating $\log(R)$ on $\log(V_{1_bs})$ for $Q = 1,000$ replicate simulated data set, each with $n = 60$, $S = 10$, $R \in \{1, 5, 10, 20, 50, 100, 200, 400\}$. Each V_{1_bs} was estimated by K-subset non-parametric bootstrap sampling as described in the method session. Here $K = 10$.

$\log(R) = 11.9959 - 14.2855 * \text{Log}(v_1)$	
RSquare	0.9846
RSquare Adj	0.9820

Parameter Estimates				
Term	Estimate	Std. Error	t value	Prob > t
Intercept	11.9959	0.45210	26.53	1.89e-07
$\log(v_1)$	-14.2855	0.7294	-19.59	1.15e-06

to the MLR procedure. K-subset bootstrap sampling as described in the method session results in a predictive relationship between $\log(R)$ and the average of $\log(V_{1_bs})$ over 1000 replicates as shown in Figure 1 (bottom). The output of the regression is shown in Table 1. In simulations we have also found that non-parametric bootstrap estimates of variances are systematically smaller than the evolutionary variance since the bootstrap procedure only considers variability across samples with the same population history instead of the evolutionary variance across random populations; however, there is a clear linear relationship between these two variance on a log-log scale. Figure 2 shows the near-perfect linear correlation between the average $\log(V_{1_iid})$ and average $\log(V_{1_bs})$ as indicated by the cross-signs. This provides us the flexibility of using i.i.d. samples to estimate the relationship between $\log(R)$ and the SFS variances for a given demographic model. We can then estimate $\log(V_{1_bs})$ from $\log(V_{1_iid})$ greatly speeding up the computation.

3.3 Comparing BSTReg to existing methods

Figure 3 shows the coverage (the percentage of replicates that have 90% or 95% prediction intervals cover true recombination rate) of predicted local recombination rate using our bootstrap based linear regression model with sample size $n = 60$, segregating sites $S = \{10, 20\}$ under the standard neutral Wright-Fisher model. The method performs well with coverage close to or greater than 90% at all level of R in the range of 1 to 400. Moreover, the coverage increases with the number of segregating sites where more information is included in the data. Mean square errors (MSEs) in figure 4 are low and uniform by our new method compared with Hudson's (1987) approach. LDhat results in the lowest MSE. Due to the limit of the maximum R that can be estimated by LDhat software, $R > 100$ are not explored here. We did not include Hey-Wakeley's (1997) performance in MSE comparison because for data with $S = 10$, Hey-Wakeley's (1997) failed to output the estimates and for samples with $S = 20$, the estimators were quite under estimated. This can be seen in figure 5 where we report the ratio of the median estimates over the true parameters for four methods. We can see that Hey-Wakeley's (1997) estimator is uniformly downwardly biased for all levels of the recombination rate in the range of $R \in \{1, 5, 10, 20, 50, 100, 200, 400\}$ while Hudson's (1987) is upwardly biased for $R \leq 50$ and performs better for larger R . Our new BSTReg approach performs almost equally

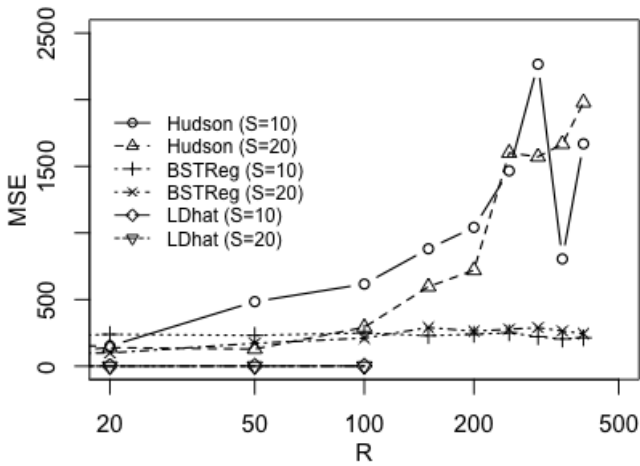


Fig. 4. Comparison of mean square errors (MSE) of predicted local recombination rate over 1000 replicates using Hudson's (1987) method, LDhat (McVean 2004) and our bootstrap based linear regression method. Sample size $n = 60$, segregating sites $S = \{10, 20\}$, $k = 10$. X-axis is plotted in log scale. Linear models are the same as used in the coverage evaluation. Due to the limit of the maximum R that can be estimated by LDhat software, $R > 100$ are not explored.

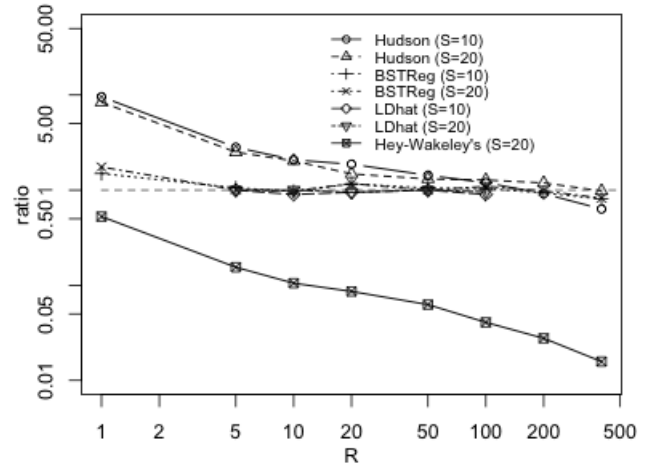


Fig. 5. Performance comparison of Hudson's (1987), Hey-Wakeley's (1997), LDhat (McVean 2004) and our bootstrap-based linear regression method in terms of the ratio of the median of predicted local recombination rates over 1000 replicates to the true recombination rate. Sample size $n = 60$, segregating sites $S = \{10, 20\}$, $k = 10$. Both axes are plotted in log scale. (For $S = 10$, Hey-Wakeley's (1997) method fails to work due to not enough informative segregating sites, results are not included in the figure; results of $R > 100$ from LDhat are not explored as well).

well as LDhat when the population size is constant and without structure: the ratios are around 1 for all levels of R .

One question that arises is: how sensitive are other approaches to the demographic assumptions of the standard neutral model? In figure 6, we report the ratio of median estimates to the true parameter by our BSTReg method and LDhat across a range of recombination rates for 1,000 simulated data sets under two island migration ($4N_e m = 12$) and population growth ($rate = 5.0$). We note that our approach has less bias, presumably since it can incorporate the demographic details explicitly in the estimating equations.

3.4 Application to the TAP2 and MS32 recombination hotspots

We have also used our approach to estimate fine-scale recombination rate variation around two recombination hotspots in the human genome characterized through sperm typing (haplotype sequences were kindly provided by Professor Sir Alec J. Jeffreys). For the TAP2 gene region, a total of 60 sequences with 48 SNPs were included in the analysis. According to Jeffreys *et al.* (2000), 81% of the sperm crossover breakpoints in the data were localized to the 1.4kb region between markers T15 and T30 (depicted as grey box from position 4,017 to 5,417 in Figure 7). We estimated the recombination rate between adjacent pairs of SNPs (as well as associated prediction intervals) using a sliding window approach with 10 SNPs in each window as described in the Methods section. Figure 7 shows the mean and lower bound of the 95% prediction interval of the recombination rate along the TAP2 genomic region before the SNP ascertainment bias correction. As we see from Figure 7, the hot spots regions identified by our approach are completely consistent with the results from both sperm typing and haplotype analysis (Jeffreys, *et al.*, 2000). That is we detect a strong signal of dramatically active recombinational exchange in the regions between markers $T16(4180)$ and $T18(4553)$,

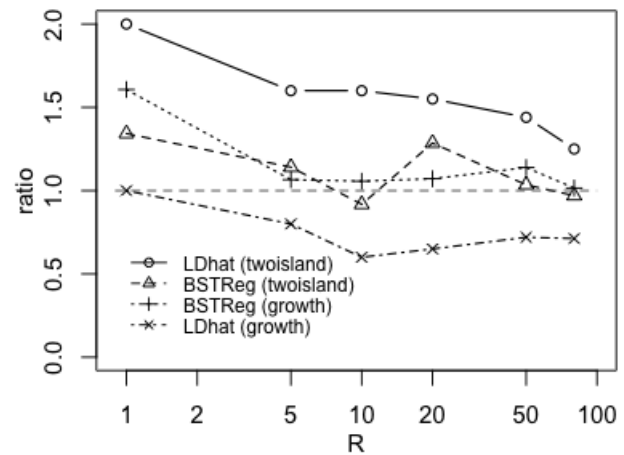


Fig. 6. Median estimates over the true recombination rate ratio over 1000 replicates by LDhat (McVean 2004) and our bootstrap-based linear regression methods under two island migration model ($4N_e m = 12$) and population exponentially growing model (growth rate $G = 5.0$). Sample size $n = 60$, segregating sites $S = 10$. X-axis is in log-scale.

$T23(4917)$ and $T24(4934)$, and $T27(5188)$ and $T30(5417)$. After the ascertainment bias correction, the same hot spots regions are identified (result not shown); but without correcting the ascertainment bias will result in more conservative estimation. In this case, the ascertainment bias increased the variance of singletons about 1.55 fold.

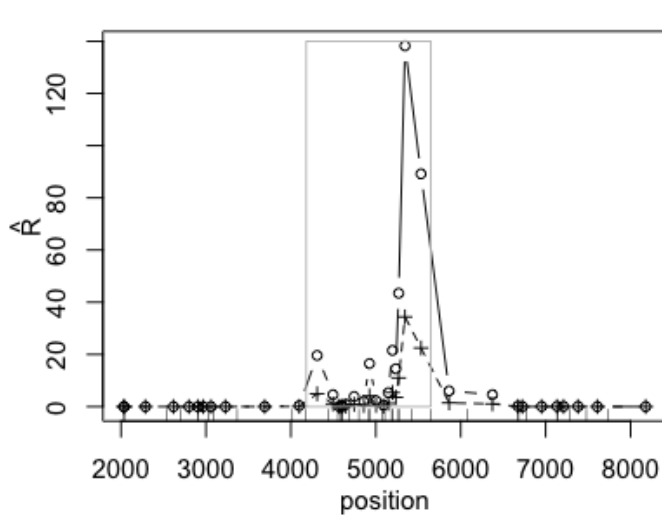


Fig. 7. The mean and lower bound of 95% prediction interval of recombination rate along the TAP2 region. The regression model in Table 1 is used for the prediction. SNPs marker positions are consistent with those in Jeffreys *et al.* (2000). Region in the grey box is the location where sperm crossover breakpoints were highly clustered (Jeffreys *et al.*, 2000).

We have also applied this approach to a 206 kb region on human chromosome 1q42.3 which contains several well-characterized autosomal crossover hotspots around the highly variable minisatellite *MS32* (Jeffreys, *et al.*, 1998). Due to the complexity of the SNPs identification in this data set, we only estimated the recombination rate without correcting the ascertainment bias. For this analysis, 80 individuals with 214 SNPs were included (we again use a $w = 10$ SNP window). Figure 8(top) shows the mean ratio of predicted recombination rate to the estimated background rate (the estimated background rates along the region which are the average rates of the local predicted rates exclude the putative hotspot regions are shown in figure 8 bottom) as well as the location of predicted hotspots by several approaches as reported in figure 1b of Jeffreys *et al.* (Jeffreys, *et al.*, 2005). The black rectangles in our figure 8(top) show the location of recombination hotspots as estimated by sperm typing (figure 1b, Jeffreys *et al.*, 2005). As demonstrated in Jeffreys *et al.* (2005), the approximate likelihood method of Fearnhead *et al.* (2004) (white triangles) and the PAC likelihood method of Li and Stephens (2003) (grey triangles) do an excellent job of identifying the location of the hotspots in the region as evidence by the strong concordance with hotspots estimated from sperm typing. Both of these approaches are very computationally intensive and require hours to run on the data set, and are thus not currently viable options for genome-wide estimation of recombination rate variation. We note that our approach (which takes about 70 seconds by a Power Mac G5 with 2.5GHz CPU speed and 4GB memory to run on the same region) shows clear signatures of recombination rate variation near the six putative hotspots (*NID1*, *NID2*, and *NID3* in and near the *NID* gene, as well as *MS32*, *MSTM1* and *MSTM2*).

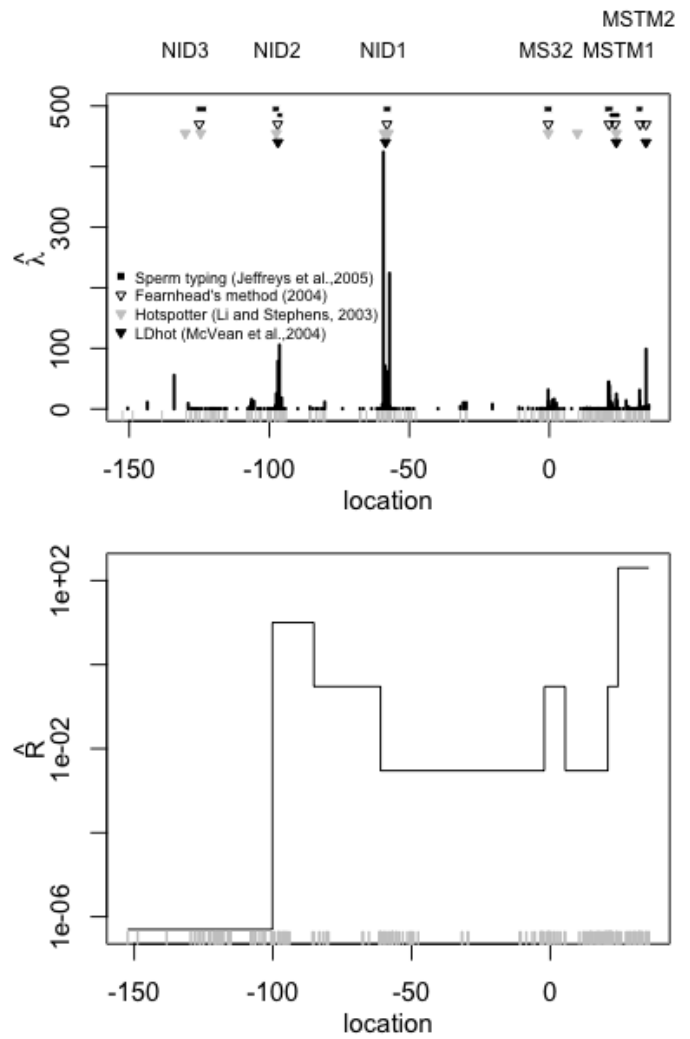


Fig. 8. Top: Ratio of recombination rate estimates to the background values in the 206 kb interval surrounding minisatellite *MS32* on chromosome 1q42.3. Putative hotspots identified by sperm typing (Jeffreys *et al.*, 2005), Fearnheads method (2004), Hotspotter (Li and Stephens, 2003), and LDhot (McVean *et al.*, 2004), respectively, are also shown as reported in Jeffreys *et al.* (2005), Figure 1b. Bottom: Estimated background recombination rate along the region. Data from: <http://www.le.ac.uk/ge/ajj/MS32/MS32%20genotypes%20file.html>.

4 DISCUSSION

We proposed a new bootstrap-based linear regression approach to estimate the population recombination rate. While the algorithm we have presented is fast, flexible, and scalable to the whole genome level, a few caveats must be raised. In order to make inference, we must still presuppose some demographic model for the data. Our preliminary results confirm the predictions of population genetic theory that recombination rate estimates will be sensitive to the demographic model used in the MLR fitting step. This sensitivity is not likely unique to our approach and probably holds for the majority of algorithms currently in use. At the same time, it also appears that our approach is robust to demography for the problem

of detecting recombination rate variation. Secondly, our method can currently only deal with uniform ascertainment schemes. When ascertainment differs dramatically among SNPs in the same region, however, this may likely cause problems for any method aiming to discover variation in recombination rate.

It is important to note that the choice of the window size on the regression region may affect rate estimation. Windows significantly overlap when we move one SNP site step by step. If the window size is too large, rate estimates are upwardly or downwardly affected by adjacent SNPs, especially when the window ranges from no or low recombination rate region to a hot spots region. From our experience with this model, we suggest that a window size between 10 to 20 SNPs appears to be an optimal trade-off between signal of recombination rate variation and noise due to stochastic variation of individual SNPs.

Lastly, we have assumed (as all other methods) that the SNPs in our sample are evolving neutrally. Since natural selection is known to affect both the patterns of linkage disequilibrium as well as the site-frequency spectrum in a region, our method is likely sensitive to this assumption. For example, a region that has experienced a recent selective sweep is expected to have low levels of nucleotide variation as well as a skew towards rare alleles. If the variance of singletons in the region is also reduced, then one may overestimate the recombination rate. One possible way to distinguish these two factors is to test explicitly for evidence of a selective sweep in the region (which is expected to leave a characteristic spatial pattern of reduced variation around the target of selection). For regions that show strong evidence of a sweep other approaches such as direct sperm typing may be necessary for accurate estimation of recombination rate variation.

ACKNOWLEDGEMENT

We thank Charles Aquadro, Andrew Clark, Martin Wells, and Scott Williamson for advice and help on earlier drafts of this paper. This work was funded by the National Science Foundation grant 0516310 to CDB.

REFERENCES

- [1]Auton, A. and McVean, G. (2007) Recombination rate estimation in the presence of hotspots. *Genome Res.*, **17**, 1219-1227.
- [2]Crawford, D.C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A. and Stephens, M. (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.*, **36**, 700-706.
- [3]Cutter, A.D., Choi, J.Y. (2010) Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Res.*, **20**, 11031111.
- [4]Cutter, A.D. and Moses, A.M. (2011) Polymorphism, divergence, and the role of recombination in *saccharomyces cerevisiae* genome evolution. *Mol. Biol. Evol.*, **28**, 1745-1754.
- [5]Ewens, W. (1972) The sampling Theory of selectively natural alleles, *Theor. Pop. Biol.*, **3**, 87-112.
- [6]Fearnhead, P. and Donnelly, P. (2001) Estimating recombination rates from population genetic data, *Genetics*, **159**, 1299-1318.
- [7]Fearnhead, P. and Donnelly, P. (2002) Approximate likelihood methods for estimating local recombination rates, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **64**, 657-680.
- [8]Fearnhead, P., Harding, R.M., Schneider, J.A., Myers, S. and Donnelly, P. (2004) Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots, *Genetics*, **167**, 2067-2081.
- [9]Fearnhead, P. and Smith, N.G.C. (2005) A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes, *Am. J. Hum. Genet.*, **77**, 781-794.
- [10]Fu, Y.X. (1995) Statistical properties of segregating sites, *Theor. Pop. Biol.*, **48**, 172-197.
- [11]Greenawalt, D.M., Cui, X., Wu, Y., Lin, Y., Wang, H.Y., Luo, M., Tereshchenko, I.V., Hu, G., Li, J.Y., Chu, Y., et al. (2006) Strong correlation between meiotic crossovers and haplotype structure in a 2.5-Mb region on the long arm of chromosome 21. *Genome Res.*, **16**, 208-214.
- [12]Griffiths, R.C. and Marjoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination, *J. Comput. Biol.*, **3**, 479-502.
- [13]Hey, J. and Wakeley, J. (1997) A coalescent estimator of the population recombination rate, *Genetics*, **145**, 833-846.
- [14]Haubold, B., Pfaffelhuber, P., and Lynch, M. (2010) mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology*, **19**, 277-284.
- [15]Hudson, R.R. (1985) The sampling distribution of linkage disequilibrium under an infinite allele model without selection, *Genetics*, **109**, 611-631.
- [16]Hudson, R.R. (1987) Estimating the recombination parameter of a finite population model without selection, *Genet. Res.*, **50**, 245-250.
- [17]Hudson, R.R. (2001) Two-locus sampling distributions and their application, *Genetics*, **159**, 1805-1817.
- [18]Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation, *Bioinformatics*, **18**, 337-338.
- [19]Jeffreys, A.J., Murray, J. and Neumann, R. (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot, *Mol. Cel.*, **2**, 267-273.
- [20]Jeffreys, A.J., Neumann, R., Panayi, M., Myers, S. and Donnelly, P. (2005) Human recombination hot spots hidden in regions of strong marker association, *Nat. Genet.*, **37**, 601-606.
- [21]Jeffreys, A.J., Ritchie, A. and Neumann, R. (2000) High-resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot, *Hum. Mol. Genet.*, **9**, 725-733.
- [22]Jiang, R., Tavare, S., and Majoram, P. (2009) Population genetic inference from resequencing data. *Genetics*, **181**, 187-197.
- [23]Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R. and Stefansson, K. (2002) A high-resolution recombination map of the human genome, *Nat. Genet.*, **31**, 241-247.
- [24]Kuhner, M.K., Beerli, P., Yamato, J. and Felsenstein, J. (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters, *Genetics*, **156**, 439-447.
- [25]Li, N. and Stephens, M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, *Genetics*, **165**, 2213-2233.
- [26]McVean, G., Awadalla, P. and Fearnhead, P. (2002) A coalescent-based method for detecting and estimating recombination from gene sequences, *Genetics*, **160**, 1231-1241.
- [27]McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004) The fine-scale structure of recombination rate variation in the human genome, *Science*, **304**, 581-584.
- [28]Noor, M.A.F., Cunningham, A.L., and Larkin, J.C. (2001) Consequences of recombination rate variation on quantitative trait locus mapping studies: simulations based on the *drosophila melanogaster* genome. *Genetics*, **159**, 581-588.
- [29]Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005) A fine-scale map of recombination rates and hotspots across the human genome, *Science*, **310**, 321-324.
- [30]Nielsen, R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms, *Genetics*, **154**, 931-942.
- [31]Sawyer, S.A. and Hartl, D.L. (1992) Population genetics of polymorphism and divergence, *Genetics*, **132**, 1161-1176.
- [32]Wall, J.D. (2000) A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.*, **17**, 156-163.
- [33]Wang, Y. and Rannala, B. (2008) Bayesian inference of fine-scale recombination rates using population genomic data. *Phil. Trans. R. Soc. B*, **363**, 3921-3930.
- [34]Zhu, L. and Bustamante, C.D. (2005) A composite-likelihood approach for directing selection from DNA sequence data, *Genetics*, **170**, 1411-1421.

Evaluation of the suitability of a Zipfian gap model for pairwise sequence alignment

Ramu Chenna¹ and Toby Gibson²

¹Biotechnology Center, Dresden University of Technology, Tatzberg 47-51, 01307 Dresden, Germany

²European Molecular Biological Laboratory, 1 Meyerhofstrasse, Postfach 10.2209, Heidelberg, Germany

Abstract—*Insertions and deletions occur during evolution of biological sequences resulting in gaps in sequence alignments. The quality of an alignment depends on the placement of the gaps. Reliable pairwise as well as multiple sequence alignments are useful in inferring protein protein interaction sites through residue conservation[23], [24]. It has been reported that the Zipfian distribution best approximates the observed gap-lengths in the sequence alignments. The probability of a gap of length N decreases, inversely related to length, as a function of N^{-c} for some suitable c . We have analysed four different gap scoring models: affine, log, power and the new Zipf that is based on Zipfian distribution. When tested on pairwise alignments from the BALiBASE benchmark suite, the widely used affine gaps were outperformed by the three other models. Log, Power and Zipf gap models performed comparably well.*

Keywords: protein sequence, sequence alignment, gap penalty, parameter reduction, zipfian distribution, riemann zeta function

1. Introduction

Aligning a new protein sequence to a known sequence is an essential and first step to study the structural and functional information of the new protein molecule. Pairwise alignments are done through a method called dynamic programming, first applied to biological sequences by Needleman and Wunsch [13]. Historically local sequence alignments are calculated using the algorithm Smith-Waterman [19] and global alignments are calculated by Needleman-Wunsch [13], each having their own advantages. For example, local alignments are more suitable for identifying protein domains irrespective of their domain shuffling and global alignment is necessary when you want full length alignments.

Many variants of sequence alignment algorithms are used for searching sequence databases e.g. SSEARCH [16] as well as methods that use word search for example FASTA [15], BLAST [2], PSI-BLAST [3] etc. Alignment scores are used to rank sequences and provide statistics for the likelihood of homology with the query. Therefore alignment quality directly influences the signal-to-noise, hence the sensitivity of database searches.

Gaps are common in alignments of biological sequences. They occur more frequently between distantly related se-

quences. Gaps in pairwise or multiple sequence alignments represent insertion or deletion (indels) events in the evolution of biological sequences.

The quality of an alignment is obtained in part through scoring aligned pairs of residues. The indels are scored by pairing a residue in one sequence with a gap in another sequence. The placement of gaps influences the quality score and hence the quality of an alignment.

Thus placement of gaps is critical in sequence alignment and they have been studied extensively [19], [18], [11], [1], [12].

A number of different gap scoring models have been discussed. However three parameters, gap open, gap extension and length of the gap are common to most of the gap models.

It has recently been proposed that observed gap lengths obey a Zipfian distribution and that this could be used to derive an appropriate gap penalty model, although this was not tested [6]. Since the Zipfian equation is so simple, we were interested in evaluating its performance for pairwise alignment. Here we report the performance of Zipfian gap penalties using the BALiBASE testbed and compare it to other concave gap models.

2. Methods

2.1 Gap Models

Gaps in sequence alignment represent the insertions and deletions that occurred in the history of the protein family of sequences [4]. Placing gaps in the right place is essential to the quality of an alignment. We have studied four different gap models by modifying the Monotone pairwise alignment package [12]. The quality of alignment for different gap models are assessed with the BALiBASE benchmark database [20].

2.2 Affine gap model

The affine gap model is the most widely used gap scoring scheme in alignment algorithms.

$$\text{gapcost} = m \cdot x + c, \quad m < c \quad (1)$$

where c is the cost for opening a gap and m is the cost of extending a gap and x is length of the gap. The default values in Monotone are for gap open $c = 9$ and gap extension

$m = 3$. However, $c = 10$ and $m = 1$ are commonly used for protein database searches.

The affine gap model is an extension of a linear gap model of the form $gapcost = m \cdot x$. In the affine model the condition $m < c$ is set to allow long insertions and deletions to be penalised less to overcome the deficiency of the linear gap model where short and long gaps are treated as equally likely. It has been shown by [6] that for gaps observed in aligned protein sequences, the affine gap is a poor approximation. The affine gap model was used to study the distribution of indel lengths [17] and they suggested a quadruple affine gap model as an alternative to plain affine gap model. This would be expensive to calculate. The linear gap model equation is in fact a straight line equation.

2.3 Log gap model

The log gap model [12] is of the form

$$gapcost = c + m \cdot \log(x) \quad (2)$$

where c is the gap open penalty, m is the gap extension penalty and x is length of gap. The default values in Monotone are for gap open $c = 9$ and gap extension $m = 3$.

2.4 Power gap model

The Power gap model [12] is of the form

$$gapcost = c + m \cdot x^d \quad \text{where } d > 0 \quad (3)$$

where c is the gap open penalty, m is the gap extension penalty, d is gap power and x is the length of the gap. The power law is convex only for $0 < d \leq 1$.

The default values in Monotone are for gap open $c = 9$ and gap extension $m = 3$ and power $d = 0.5$.

2.5 The new Zipf gap model based on Zipfian distribution

Chang and Benner have studied the gap length distribution in a set of pairwise alignments and suggested that the Zipfian distribution can be used as a best approximation for scoring the gaps in an alignment [6]. Their detailed study shows that the number of gaps say n of length N decreases according to the expression

$$n = c_1 N^{-c_2} \quad (4)$$

where c_1 and c_2 are parameters empirically selected to fit the data.

Benner also suggested that this function is independent of the length of the gap and the extent of divergence. One caveat is that they used a dataset with just one gap per pairwise alignment. Even if the Zipfian holds for multiple gaps, the derived parameters may not. We have further tested their suggestion by incorporating the Zipfian gap scoring model into the Monotone [12] pairwise alignment package.

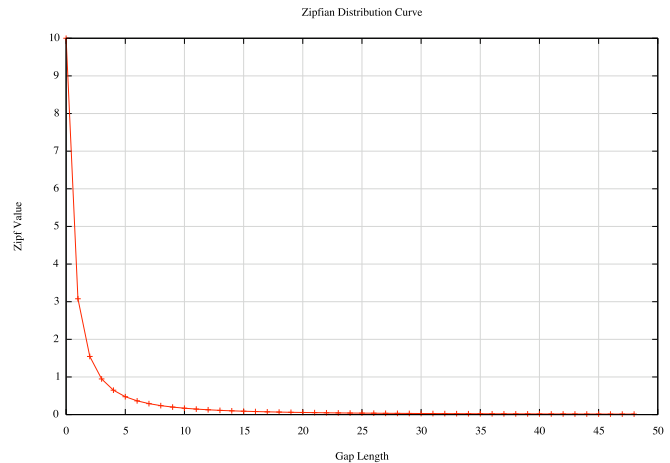


Figure 1: Gap penalty scores generated by the Zipfian curve for $10 * N^{-1.7}$ where N is the gap length. Starting value of $c_1 = 10$ is typical for pairwise alignment with the Blosom62 matrix.

Figure 1 shows the gap penalty scores generated by the Zipfian curve for $10 * N^{-1.7}$ where N is the gap length. A starting value of 10 is typical for pairwise alignment with Blosom62 [10]. The value of the curve at position N is added to the gap extension cost at position $N-1$. Since the curve converges, we cannot use the equation 1 as it is. Therefore we take the cumulative sum over the entire given gap length.

Define the cumulative sum

$$gapcost(n) = P = \sum_{N=1}^n \frac{c_1}{N^{c_2}} \quad (5)$$

Here the cumulative sum P can be used as the gap cost for inserting a gap of N (or n) symbols. This gap cost function is monotonically increasing where $gapcost(n) > gapcost(n-1)$ for all n . In other words it is a non-decreasing, concave gap function.

The equation 5 is in fact the partial sum of the infinite series of the famous Riemann Zeta function of the form

$$\zeta(p) = \sum_{n=1}^{\infty} n^{-p} \quad (6)$$

As a special case when $p = 1$, the $\zeta(p)$ becomes the logarithm function which is an advantage that one could mimic different gap models inside the alignment algorithm by changing the exponent p of Riemann Zeta function [21].

Riemann Zeta function is extensively studied in number theory and has number of interesting properties. When one considers indels as infinite series of evolutionary events then it would be interesting to study these events in the light of Riemann Zeta function.

The Figure 1 shows the plot of equation 4. The curve is asymptotic to X-axis.

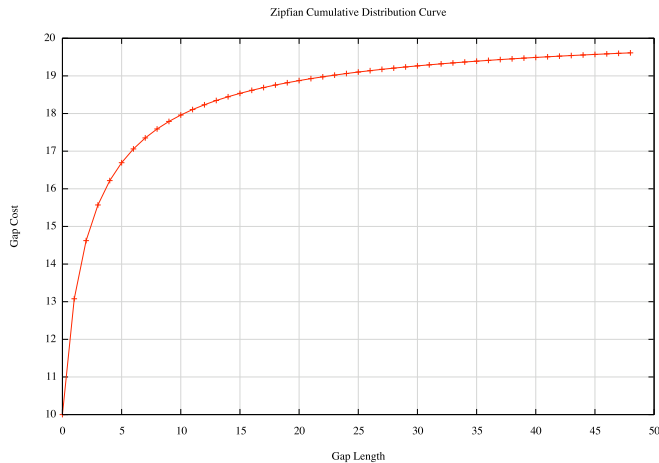


Figure 2: Shows the cumulative distribution plot of equation-5 showing the gap score for different gap lengths. The value of the curve at position n is added to the gap extension cost at position $n-1$. For gaps of more than 20 residues, the extension cost becomes very small. This curve diverges as the length of the gap increases. The costs of opening and then extending small gaps are high. However the curve becomes asymptotic and the costs of further extending long gaps become very small.

Figure 2 shows the Cumulative sum for the Zipfian values in Figure 1 (See equation 5). The costs of opening and then extending small gaps are high. However the curve becomes asymptotic and the costs of extending long gaps become very small.

3. Results

3.1 The Scoring Method

We used the BALiBASE Version 2.0 benchmark alignment database [20]. BALiBASE is designed for evaluating multiple sequence alignment algorithms. Alignments in BALiBASE were derived from visually inspected structure alignments. Therefore they are not biased toward any sequence alignment method.

BALiBASE consists of mainly five different reference alignment sets. Reference 1 consists of equidistant sequences, Reference 2 consists of related families with divergent, orphan sequences, Reference 3 consists of families of related sequences, Reference 4 consists of N and C terminal extension sequences, Reference 5 consists of internal insertions. Refer to the BALiBASE [20] paper for more details.

We have modified Richard Mott's software package Monotone [12] to allow Zipfian values to be computed and used for gap scoring. The Monotone package is elegantly designed and it was easy to incorporate the new Zipfian gap model. Monotone also comes with the affine, log and power gap models.

Monotone reads two sequences from two different files. So it was necessary to split the sequences from the BALiBASE sequence files. First the sequences from reference files were separated into single files and the multiple sequence alignment files were also separated with all the possible pairwise combinations intact. See table 1 for details.

Table 1: Pairwise alignments available in BALiBASE-2

Set	Number of Files	Sequences	Pairwise alignments
Ref1	82	367	652
Ref2	23	412	3544
Ref3	12	266	2865
Ref4	12	107	504
Ref5	12	112	570

Each file in each Balibase reference set consists of different numbers of sequences. Note that the total pairwise $\frac{n(n-1)}{2}$ comparisons is based on the number of sequences in each file.

A script was written to generate all the possible pairwise alignment commands to run Monotone for different gap models. The program BaliScore was used to assess the quality of the alignment with reference to the test alignment. BaliScore gives SP, Sum of Pairs score and CS, Column Score. The SP score determines the extent to which the test program, in this case Monotone, succeeded in aligning the sequences. The CS score is designed to see whether the test program can align all of the sequences correctly in a multiple alignment (that is not relevant here).

We plotted the overall SP scores for all the different gap models using Blosum62 [10] and Gonnet PAM250 [5] matrices by varying the exponent of the Zipfian gap model by 0.1 increments over a range from 1.0 to 2.0.

3.2 Comparison of four penalty schemes

Figures 3 and 4 show the comparison of the four gap penalty models: affine, log, power and Zipfian. The X-axis shows the gap open penalty varied from 1 to 25 and the Y-axis shows the overall percentage BaliScore [20] for all five different reference sets (See Table.1). The range of baliscore is from 0, the lowest, to 1, the highest for each alignments. We used two different popular comparison matrices namely Blosum62 [10] and Gonnet PAM250 [5].

Examining the affine scores in Figures 3 and 4, the best scores are achieved with gap opening in the range 7.0 - 9.0 for both matrices, however the overall score is higher for PAM250, indicative of better alignments. Better quality alignments for the PAM250 matrix are in accordance with previous matrix comparisons (Vogt et al., 1995). Optimal gap opening penalties observed here are slightly lower than the typical values used as defaults in sequence alignment.

In both figures, the peak BALiBASE scores for affine gap are below the peak scores of the other gap functions. It is also clear that the affine gap penalty does not tolerate the higher penalty values as well as the other models. (This is not a major consideration provided that gap penalties are

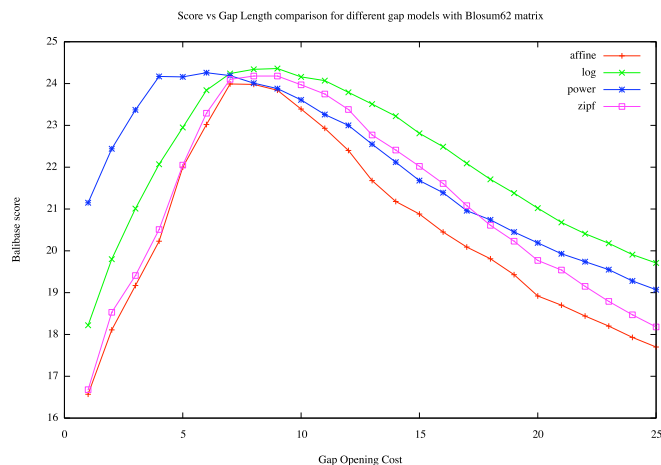


Figure 3: Comparison of 4 gap penalty schemes tested by pairwise BALiBASE scores for the Blossum62 exchange matrix. The obtained BALiBASE score (Y axis) is reported for the cost of opening a gap varied over a range of 2 to 25 (X axis). For all gap penalties, affine gaps always perform worse than the other functions. Peak performances of the log, power and Zipf functions are very close with log slightly ahead. See text for the equations describing the gap functions.

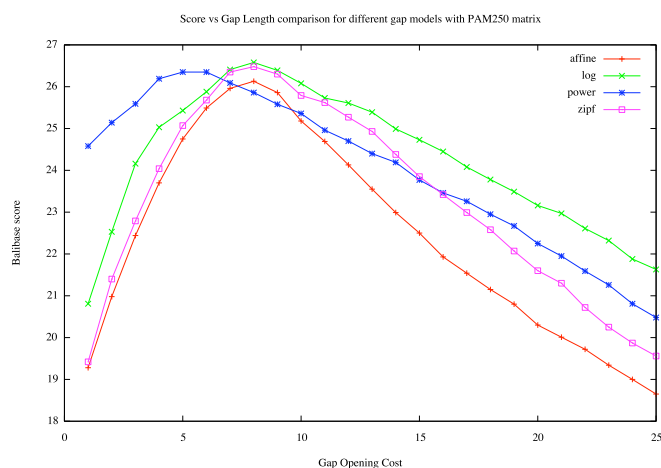


Figure 4: Comparison of 4 gap penalty schemes tested by pairwise BALiBASE scores for the Gonnet PAM250 exchange matrix, suitable for aligning highly divergent proteins. The obtained BALiBASE score (Y axis) is reported for the cost of opening a gap varied over a range of 2 to 25 (X axis). Affine gaps again perform worse than the other functions. Note that the better performance of the smooth models is more apparent with the more sensitive PAM250 matrix than with the Blossum62. Peak performances of the log, power and Zipf functions are very close again with log slightly ahead. See text for the equations describing the gap functions.

being set close to optimal performance). The peak difference is smaller than we expected, especially for Blossum62. The poorer performance of affine becomes clearer using the more sensitive PAM250 matrix. This may imply that alignment with better residue exchange parameterisation benefits more from the improved gap penalty models.

The log, power and Zipfian models outperform affine for both Blossum62 and PAM250. However, the peak performances of the three smooth models are all very close for both tested matrices. Based solely on performance in our tests, we would not be able to choose between the three models. Note that for log and power we ran the tests using the default Monotone gap extension value of 3.0. This value has already been well optimised for the smooth gap models supplied in the Monotone package. However, we observed very poor performance for affine with the Monotone defaults (data not shown) - a gap penalty of 3.0 is much higher than usually recommended for protein alignment. Therefore, in accordance with standard practice, we have kept the affine gap opening and gap extension penalties in the ratio of 10 to 1 for the tests.

From these figures, it is clear that the default gap opening penalty value 9.0 for Monotone could be set to higher.

3.3 Effect of varying the Zipfian exponent

The exponent of the equation 5 c_2 has been varied in the range of 1.0 to 2.0 keeping c_1 at a constant 1.0. The Figure 5 and Figure 6 are graphs showing the overall percentage BaliScore score distribution for the variations of c_2 . The gap opening penalties are computed using the equation 5. With the exponent 1.7 the highest score 24.18 is achieved when gap opening penalty is 8 or 9 for Blossum62 matrix whereas for PAM250 matrix with the same exponent the highest score is 26.48 when the gap open penalty is 8. For exponent 1.8 the highest score 26.49 with gap open 8.0 for PAM250, and with Blossum62 it is 24.15 with gap open 9.0.

The results are in good agreement with the observed value of exponent $c_2 = 1.8$ [6]. Though the higher exponent tolerates large gap penalties they do not get higher score comparatively. In a progressive multiple sequence alignment scenario the exponent in the Zipfian could be used to adjust the gap openings dynamically for different divergence.

4. Discussion

The Zipf law has been used to study phenomena in many areas e.g. linguistic, audio signals [8] and also recently to study the human transcriptome [14]. The Zipf law suggests that the frequency of occurrence of a word is inversely proportional to its rank.

Chang and Benner showed that the gap-lengths can be approximated by the Zipfian distribution with the probability of a gap of length N decreasing as a function of the gap length [6].

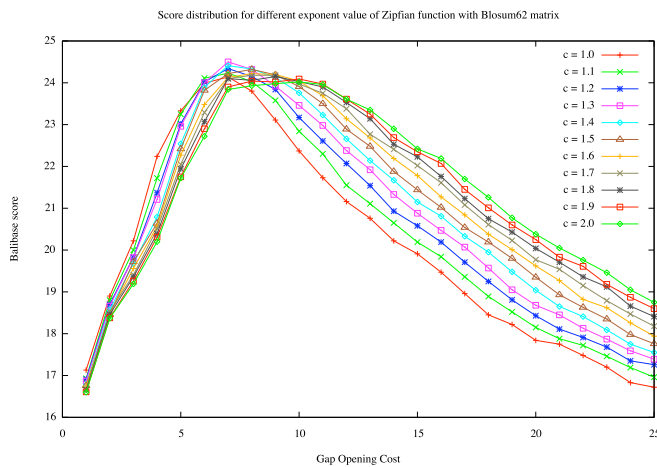


Figure 5: BALiBASE score distributions for different exponent values of the Zipfian function and the Blosum62 matrix. Varying the exponent yields a small difference for peak performance with $N^{-1.4}$ marginally best. However, this value performs relatively less well when gap penalties are set too high. Values closer to the Benner exponent of $N^{-1.8}$ have a broader peak and are more tolerant of higher gap penalties: These values may be appropriate when sequence similarity varies widely and gap penalty values are necessarily imprecise.

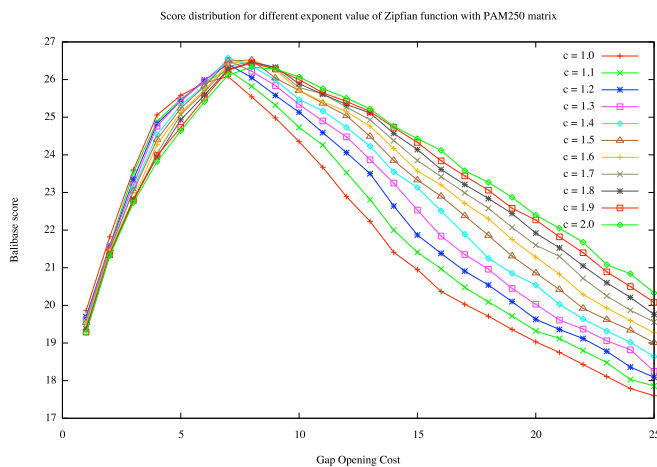


Figure 6: BALiBASE score distributions for different exponent values of the Zipfian function and the Gonnet PAM250 matrix. Varying the exponent yields even smaller differences than for Blosum62 for peak performance with $N^{-1.4}$ marginally best. Again, this value performs relatively less well when gap penalties are set too high. Values closer to the Benner exponent of $N^{-1.8}$ have a broader peak and are more tolerant of higher gap penalties: These values may be appropriate when sequence similarity varies widely and gap penalty values are necessarily imprecise.

Other authors have proposed concave gap functions. For long gaps, Gotoh used a piecewise linear gap-weighting function that approximated a smooth concave function [9]. Miller and Myers [11] and Waterman [22] also used concave weighting functions to compare sequences. Mott has shown that by using monotonic gap penalties, the chances of detecting a similarity containing a long gap is greater over affine gap penalties. We have seen that in Monotone the default value of gap opening 9 and gap extension 3 for affine gaps performed very poorly. The situation improved when we modified the affine gap extension as one tenth of gap opening penalty, which is typical for protein sequence alignment, but still the affine gap penalty was worse.

Despite the extensive literature on concave penalty functions, all widely used alignment and database search software continue to use affine gaps. One reason may be increased computational costs. Myers and Miller found affine gaps to be three times faster than other concave functions. Performance should be less important with recent computer hardware. Furthermore precalculation and array lookup can reduce the time penalty for any gap scheme that is more complicated than affine.

BALiBASE is the most widely used alignment benchmarking suite. Using BALiBASE we have now shown that the non-affine gap penalties are better suited for pairwise sequence alignments. Although it does not outperform the other smooth gap models, the Zipfian model shows promise as the simplest of the models tried, with the lowest parameter space. From the Figure 5 and Figure 6 it is clear that the higher exponent tolerates longer gaps.

We would like to suggest that in a progressive multiple alignment environment where the highly homologous sequences are aligned first, the gap opening in equation 4 could be adjusted automatically to fit to the extent of divergence of the sequence or profile that are already aligned with the new sequences or profile. This is quite logical because a fixed gap opening cannot perform well for merging group of sequences with varied degree of divergence.

BALiBASE covers a range of alignment test cases including long gaps. Though BALiBASE benchmark alignments are designed for testing multiple sequence alignment programs, BALiBASE can also be adopted for use with pairwise alignments. The pairwise test with local alignment approximates database search properties. Thus Zipfian gap model should be suitable for use in sequence database searches.

The Zipfian gap model might also be useful for nucleic acid alignment, since genomic sequence alignment needs to handle very large indels. For example, insertion of a Line-1 element creates a gap of more than 8000 bases and affine gaps are completely unsuitable for dealing with such long gaps.

In future work, we hope to examine the performance of Zipfian penalties for the progressive alignment algorithm of Clustalw [7]

Acknowledgement

We would like to thank Richard Mott for providing the Monotone software package, and Julie Thompson for the BALiBASE benchmark suit and helpful discussions. I also thank Des Higgins, Mark Larkin, Ian Wallace, Lars Juhl Jensen and unknown referees for their careful reading and critical comments.

References

- [1] Altschul, S. F. (1989) Gap costs for multiple sequence alignment. *J Theor Biol*, **138** (3), 297–309.
- [2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol*, **215** (3), 403–410.
- [3] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25** (17), 3389–3402.
- [4] Benner, S., Cohen, M. & Gonnet, G. (1993) Empirical and Structural Models for insertion and deletions in the divergent evolution of proteins. *J. Mol. Biol*, **229**, 1065–1082.
- [5] Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng*, **7** (11), 1323–1332.
- [6] Chang, M. S. S. & Benner, S. A. (2004) Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *J Mol Biol*, **341** (2), 617–631.
- [7] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, **31** (13), 3497–3500.
- [8] Delland, E., Makris, P. & N, V. (2004) Zipf analysis of audio signals. *Fractals*, **1** (12), 73–85.
- [9] Gotoh, O. (1990) Optimal sequence alignment allowing for long gaps. *Bull Math Biol*, **52** (3), 359–373.
- [10] Henikoff, S. & Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89** (22), 10915–10919.
- [11] Miller, W. & Myers, E. W. (1988) Sequence comparison with concave weighting functions. *Bull Math Biol*, **50** (2), 97–120.
- [12] Mott, R. (1999) Local sequence alignments with monotonic gap penalties. *Bioinformatics*, **15** (6), 455–462.
- [13] Needleman, S. B. & Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48** (3), 443–453.
- [14] Ogasawara, O., Kawamoto, S. & Okubo, K. (2003) Zipf's law and human transcriptomes: an explanation with an evolutionary model. *C R Biol*, **326** (10-11), 1097–1101.
- [15] Pearson, W. R. & Lipman, D. J. (1988a) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85** (8), 2444–2448.
- [16] Pearson, W. R. & Lipman, D. J. (1988b) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85** (8), 2444–2448.
- [17] Qian, B. & Goldstein, R. A. (2001) Distribution of Indel lengths. *Proteins*, **45**, 102–104.
- [18] Sankoff, D. & Kruskal, J. (1983) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA.
- [19] Smith, T. F. & Waterman, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147** (1), 195–197.
- [20] Thompson, J. D., Plewniak, F. & Poch, O. (1999) BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15** (1), 87–88.
- [21] Titchmarsh, T. (1951) *The Theory of Riemann Zeta Function*. Oxford Univ Press Oxford.
- [22] Waterman, M. S. (1984) Efficient sequence alignment algorithms. *J Theor Biol*, **108** (3), 333–337.
- [23] Panjkovich, A. and Aloy, P. (2010) Predicting protein-protein interaction specificity through integration of three dimensional structural information and the evolutionary records of protein domains. *Molecular Biosystems* **6**(4), 741-749.
- [24] Want, B. and Wong, HS. (2006) Inferring protein-protein interacting residue conservation and evolutionary information. *Protein and Peptide Letters* **13**(10), 999-1005.

A probabilistic approach for characterizing the marking system of multiplex sequencing in ABI SOLiD platform

F. Lobato¹, P. Machado¹, A. Gonçalves², Â. Ribeiro-dos-Santos², D. Alencar², S. Darret², Á. Santana¹

¹Technological Institute, Federal University of Pará, Belém, Pará, Brazil

²Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil

Abstract—*High-Throughput Sequencers such as Illumina and ABI SOLiD, generate large quantities of data, typically above 10 Gigabytes of text files. These platforms enable multiplex sequencing, that is, the sequencing of multiple samples in a single run, through a marking system. This requires a computational process for separation the data generated, which contains the mixture of all samples in a single output. It is necessary that the quality of the marking system is evaluated to ensure the reliability of this separation. This work proposes measures to characterize the marking system obtained from SOLiD sequencing. In fact, measures presented are proven to be sufficient to describe the sequencing and hence guide the process of filtering the data and the analysis of the sequencing protocol.*

Keywords: Statistical Analysis, Multiplex Sequencing, SOLiD, Barcode System.

1. Introduction

The volume of data generated by new DNA sequencers increased substantially in recent years. Platforms such as Illumina and ABI SOLiD, High-Throughput Sequencers, can generate millions of small reads sequences from different samples in a single multiplex run.

The SOLiD platform has its own peculiarities when compared to other sequencers, particularly, the data representation, which is coded in "two base color encoding", also called colorspace. This represents the transition between two nucleotides of a characteristic color obtained by the detection of fluorochromes: FAM; Cy3; TXR; or Cy5 [1]. What, in turn, implies the need to add a step for converting the data, in order to obtain the DNA sequence of nucleotide bases.

Additionally, the SOLiD sequencing supports up to 256 multiplex samples by means of, among other items, the marking system called barcode [2]. These have a characteristic central to the process of discernment between the samples, the orthogonality, meaning that a barcode of the standard library has no correlation with each other. However, even with all the security surrounding the library of markers, there were failures of common error, known as erroneous color calls, and in the quality aspects associated with the transitions of nucleotide bases [3].

These failures should be identified and, if possible, mitigated, given the importance of accuracy in the recovery of

sequences per marker. However, each run held in the SOLiD platform has unique characteristics for the marking system, which ratifies the need to study methods to assess the quality of sequencing protocols.

It is important to consider the impact of a high degree of reliability for the sequencing data due to the fact that failures in barcode systems can cause a shuffle in the sequences of interest. And it implies in a waste of computer processing in genome analysis and eventual errors in results.

In this context, the lack of literature, studying measures to characterize the sequencing as the marking system, motivated this work, in which a statistical analysis is developed in order to identify summary measures to characterize the SOLiD sequencing as the marking system.

Through computational tests, it was defined four measures: median, mode, variance and sequences to barcodes ratio. The results obtained allowed demonstrating the differences in the marking system for each run analyzed. In fact, the previously mentioned measures are sufficient to describe the sequencing and hence, guide the process of filtering data and the analysis of the sequencing protocol.

This paper is organized as follows: section two presents the related work to the analysis made, section three describes the materials and methods used in the development of this work, which is presented in section four. The results obtained are discussed in section five and finally, section six presents the conclusions.

2. Correlated Works

Most studies that involve the filtering of errors or quality assessment of data generated by the SOLiD platform are based on heuristics. Taking for example the work of [4], sequences that show some transition with quality below a predetermined threshold are retained by the filter. In this same study, the default values adopted for filtering independent errors is a Quality Value (QV) ≤ 10 , while the errors of polymorphism is QV ≥ 25 .

Other works like [5] treat the errors of substitutions, insertions and deletions. However, this treatment does not take into account the quality value, because they filter the data in advance using a system based on heuristics.

It should be noted that such heuristics are useful for the analysis of large sequences. Heuristic-based algorithms usually have lower complexity and require less processing

time compared with analytical algorithms. On the other hand, analytical algorithms improve reliability by generating accurate results based on analysis, something required for barcodes.

The differential matter of this study is the adoption of a statistical analysis in order to assess the quality of barcodes, that allow the characterization of sequencing protocols to guide the development of a custom filter to the marking system used in the SOLiD platform.

3. Materials and Methods

The biological material used in the studies was derived from cancer patients, extracting two samples each, with the written consent for study approved by the Committee of Ethics in from the Federal University of Pará, protocol number 14052004 / HUIBB.

The sequencing process is preceded by some essential procedures: sample collection after the DNA is fragmented, as the sequencer can only read fragments from 35 to 50 base; the linkage to known sequence fragments with adapters, named P1 and P2, as shown in Figure 1.

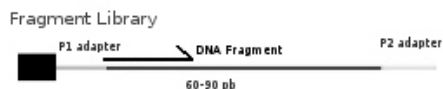


Fig. 1: Schematic drawing of sample preparation.

The adapter has a P1 sequence complementary to the metal beads; and P2 has complemented by a polystyrene coated bead, a material which floats in water; this way, the fragments that do not affect P1 and P2, will be discarded. The templates relating to the selected beads have their 3' end modified, so they can join covalently to the blade. In the end, they are deposited on the blades and taken for sequencing, as shown in Figure 2.

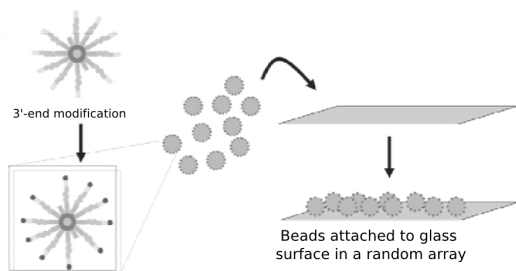


Fig. 2: Modification of the P2 adapter to allow the grip to glass surface.

The markers in each sample are inserted into the adapter P2. The SOLiD Small RNA Expression Kit (Ambion Inc., U.S.), was used for the preparation of fragments. All miRNAs were attached to the library with a specific extension of primers, in this case, the barcode system.

The data generated from three runs consists by the barcodes, samples sequences and quality values files. Together they amounted over 175 Gigabytes (Gb), which about 24 Gb correspond to information from barcodes. These data were analyzed in order to obtain sufficient statistics from the following summary measures: mean, mode and variance, in addition to viewing the probability distribution to facilitate evaluation of multiplex sequencing.

4. Statistical Analysis of Multiplex Sequencing

For this analysis, it is necessary to know the probability of the marking sequence, however, the Applied BioSystems does not provide the mapping function between the value of quality and the associated probability.

In [6] we found a function that adapted for the range of quality values generated by the SOLiD platform and approximate the quality-probability mapping, as follows:

$$P(Q) = 1 - 10^{-(Q+1)/10} \quad (1)$$

The value $P(Q)$ represents the probabilistic degree of confidence of a given transition, as evaluated by their quality value, represented in Equation 1 to $(Q + 1)$. This aspect was used for normalization, for Q is comprised in the range from -1 to 35, so the value of -1 would indicate a negative outlook.

To calculate the confidence level of a given sequence (θ), multiply the probabilities of all existing transitions. The result represents the probability that the sequence obtained is, in fact, present in the sample.

$$P(\theta) = \prod P(Q) \quad (2)$$

To optimize the calculation of summary measures and plotting the probability distribution, the results obtained by applying Equation 2 are stored in a data structure, the map [7]. This structure is composed by two fields, one for storing the key and another for the value; in this problem, the likelihood is the access key that points to the number of occurrences, so the probability distribution is easily manipulated. Other relevant information is the relationship between the total amount of marking sequences obtained and those which corresponded to one of the ten barcodes presented in the standard library used in the experiments.

5. Results

The statistics contained in Table 1 have revealed that C1, with an average of 81.23% has a higher confidence than C2 and C3 in the quality of markers and, consequently, the recovery of sequences from samples. The variance of 5.5% was the lowest compared to other runs, indicating a low dispersion of data in relation to the expected value.

It is observed that 30.88% was the most frequent likelihood in the runs C1 and C3. Moreover, C2 had higher

Table 1: Information about Sequencing Analyzed.

Data	First Run(C1)	Second Run(C2)	Third Run(C3)
Number of sequences	142,453,565	19,523,621	27,634,981
Mean	0.8123	0.4715	0.5917
Mode	0.3088	0.3726	0.3088
Variance	0.0553	0.0687	0.0806
Sequence to Barcode ratio	73,42%	1,45%	44,66%

value, of 37.26%, in the value of mode. The proportion the Sequence to Barcode ratio presented in C1 is 73.42%. C3 showed a drop to 44.66%; and in C2, the ratio is extremely low, 1.45%, indicating prior sequencing problems.

However, between C2 and C3, the latter showed the best performance, with an average of 59.17% and 8% of variance. The graphical analysis of these two runs, obtained by comparing Figures 4 and 5, evidences this difference, especially regarding the density of sequences with low confidence, presented in Figure 4 ; even though the mode of this run is higher, it does not reflect the general scenario.

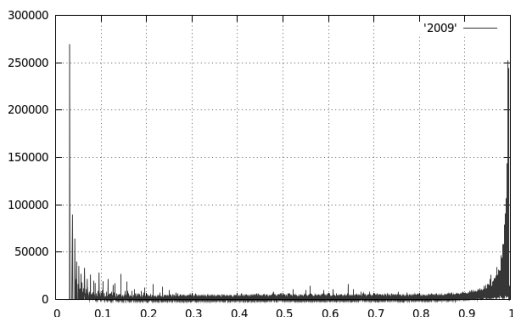


Fig. 3: Probability distribution of first run.

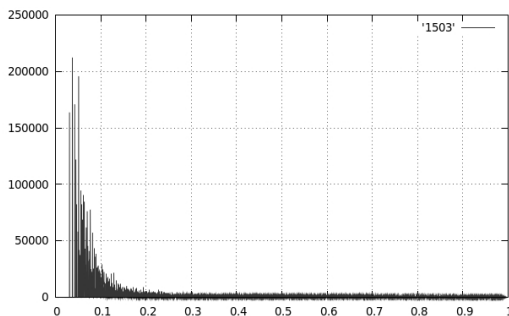


Fig. 4: Probability distribution of second run.

Such statements were pertinent to the sequencing protocol analysis and verification of possible disposal of the data generated in the case of low confidence. For example, the values for runs C1 and C3 are statistically significant, detecting at least 45% of the sequences marked, thus proving trustworthy for the genomic analysis.

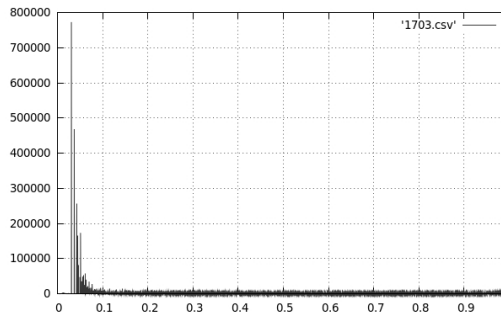


Fig. 5: Probability distribution of third run.

However, the data qualities of C2 are extremely low, with only 2% of recognized barcodes that were actually used in the sequencing; possibly indicating a shuffling of the samples. A scenario that illustrates this problem is the sequencing of two patients, one healthy and another sick; if the marking system has characteristics similar to C2, the sequences of interest can be exchanged between the patients, which causes errors in the results.

Regarding the poor quality of data from C2, among the possible causes of this failure, it can be highlighted the fluctuation of electric power on the sequencing unit. This particular failure could affect the capture of beads, as these are captured through an electromagnetic field, which is extremely sensitive to power quality.

6. Conclusions

The lack of literature on summary measures able to characterize the sequencing with regards to the marking system is one of the aspects that motivated this work. Also, we stress the importance of a high degree of reliability of these data; in particular, because the marking system failures can cause the shuffle of the sequences of interest, which implies on a waste of computational processing for genome analysis and errors in the results.

Seeking to fill these gaps, we developed a statistical analysis that had the following summary measures: median, mode, variance and sequences to barcodes ratio. This allows, among other analysis: the evaluation of protocols used in the preparation of the libraries for labeling in multiplex sequencing; assessment of possible discard of the generated data; and the initial guiding in the process of developing a custom filter for barcodes.

References

- [1] H. Brey, "A theoretical understanding of 2 base color codes and its application to annotation, error detection, and error correction," *White Paper SOLiD System*, 2010.
- [2] A. Biosystem, "Solid system barcoding," *application note SOLiD*, 2008.
- [3] A. Valouev and et al., "A high-resolution nucleosome position map of *c. elegans* reveals a lack of universal sequence-dictated position," *Genome Res.*, vol. 18, pp. 1051–1063, 2008.

- [4] A. Sassom and T. P. Michael, "Filtering error from SOLiD output," *Bioinformatics*, vol. 26, pp. 849–850, 2010.
- [5] L. Salmela, "Correction of sequencing errors in a mixed set of reads," *Bioinformatics*, 2010.
- [6] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using phred. ii. error probabilities," *Genome Res*, vol. 8, pp. 175–185, 1998.
- [7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "*Introduction to Algorithms*." McGraw-Hill, 2002.

Enhancement on the predictive power of the prediction model for human genomic DNA methylation

Hao Zheng¹, Shi-Wen Jiang², and Hongwei Wu^{1*}

Abstract—DNA methylation is an important type of epigenetic modification that plays an instrumental role in organogenesis, cellular differentiation, suppression of deleterious elements, and carcinogenesis. In addition to the experiment-based approaches, computational prediction provides guidance in an effective, fast and cheap way to the genome-wide DNA methylation profiling. In this paper, we describe the development of support vector machine-based models for the prediction of the CpG island methylation. The features used for prediction include those that have been previously demonstrated effective (e.g., CpG island specific attributes, DNA sequence composition patterns, DNA structure patterns, distribution patterns of functional and evolutionarily conserved elements, and histone methylation status) as well as those that have not been extensively explored but are likely to contribute additional information from a biological point of view (e.g., nucleosome positioning propensities, gene functions, and histone acetylation status). Statistical tests were performed to identify the features that are significantly correlated with the methylation status of CpG islands, and principal component analysis was subsequently performed to decorrelate the selected features. The CpG island methylation profile data from the Human Epigenetic Project were used to train, validate and test our predictive models. Specifically, the models were trained and validated by using the data of the CD4 lymphocyte, and were then further tested for generalizability using the data of the other 11 tissues and cell types. The experiments showed that (1) an eight-dimensional feature space that was selected via the principal component analysis and that combines all categories of information was effective for predicting the CpG island methylation status, (2) by incorporating the information regarding the nucleosome positioning, gene functions, and histone acetylation, the model could achieve a higher specificity and accuracy than the existing model while maintaining a comparable sensitivity, (3) the histone modification information contributed significantly to the prediction, without which the performance of the model deteriorated, especially in terms of sensitivity, and, (4) the predictive models generalized well to different tissues and cell types, no matter whether the histone modification information was incorporated or not.

I. INTRODUCTION

Epigenetics refers to a somatically inheritable pattern of gene expression that is determined by mechanisms other than those encoded in DNA sequences. DNA methylation is an important type of epigenetic modification, implicated in critical cellular functions including genetic imprinting, X-chromosome inactivation, suppression of retroviral elements, and carcinogenesis. DNA methylation involves the

addition of a methyl group to DNA via DNA methyltransferase, and typically occurs at the cytosine residues in a CpG dinucleotide context [1]. CpG dinucleotides in human genome are relatively rare but are enriched in short DNA segments known as CpG islands [2]. Most CpG dinucleotides are methylated in human somatic cells [3], but the CpG dinucleotides residing within CpG islands tend to remain unmethylated.

DNA methylation can be determined experimentally via biochemical assays or sequencing. On the other hand, computational modeling can effectively complement the wet chemistry approach in identifying critical factors or pathways controlling DNA methylation patterns, as well as to provide valuable information when methylation data are unavailable for certain genome regions. Computational prediction of DNA methylation has been conducted at two levels – CpG dinucleotides and CpG islands, respectively. At the CpG dinucleotide level, DNA fragments of fixed length with a cytosine in the center were used for the prediction. Each nucleotide was represented by a 5-bit binary sparse code, so that each DNA fragment was represented by a series of codes and the difference between DNA fragments could be quantified. With the optimal DNA fragment length (39 nucleotides), a ~75% of accuracy could be reached for predicting whether a CpG dinucleotide is methylated or not [4]. At the other level, computational models have been developed to distinguish between methylated and unmethylated CpG islands (or DNA fragments with high CpG density). For example, Feltus et al. used DNA sequence patterns to distinguish methylation-prone and methylation-resistant CpG islands under *de novo* methylation, and reached an 82% accuracy [2]. Bock et al. augmented the feature space by including DNA sequence patterns, DNA repeats and predicted DNA structure. Their experiments on the Human Epigenome Project (HEP) data set showed a ~90% accuracy for predicting the methylation status of DNA fragments of high CpG density [5] [6]. The MethCGI used both the DNA sequence composition and transcription factor binding site (TFBS) features to characterize CpG islands and reached an 84% specificity and 84% sensitivity on human brain data [7]. Fan et al. augmented the feature space of the CpG island by including histone methylation information, which is highly correlated with DNA methylation, and reported a 94% sensitivity and 74% specificity on the HEP data [8].

In this study, we considered various attributes that are possibly related to the CpG island methylation. These attributes include those that have been previously investigated (e.g., the CpG island specific attributes, DNA sequence

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA. {hzheng7, hongwei.wu}@gatech.edu

²Department of Biomedical Sciences, Mercer University School of Medicine-Savannah Campus, GA, USA. Jiang.s@mercer.edu

*corresponding author

patterns, DNA structure patterns, distribution patterns of functional and evolutionarily conserved elements, and the histone methylation status) as well as those that have not been extensively investigated (e.g., nucleosome positioning propensities, gene functions, and histone acetylation status). The contribution of each individual feature was evaluated by statistical tests; and the correlation between features was reduced by principal component analysis (PCA). These DNA methylation-relevant yet non-intercorrelated features were then used to build support vector machine-based classifiers to predict the methylation status of CpG islands. The predictive models were evaluated by using the HEP data set. Specifically, the CpG island methylation profiles in the CD4 lymphocyte were used to train and validate the models, while the CpG island methylation profiles in the other 11 tissues/cell types were used to test the generalizability of the models. Through these experiments, we assessed the individual and combinational influence of the newly added features (nucleosome positioning propensities, functions of nearby genes, and the acetylation status of nearby histones) and the impact of histone modification information.

II. DATA SETS

Methylation profiles were obtained from HEP. HEP aims to provide the high-resolution data set regarding genome-wide DNA methylation patterns in human tissues and cell lines [9]. It currently covers chromosomes 6, 20 and 22, and provides 1.9 million CpG methylation values of 2,524 amplicons derived from 12 different tissues and 43 different samples using bisulfite DNA sequencing. The methylation values of the analyzed CpGs range from 0 to 100 inclusive, where 0 corresponds to the lowest and 100 to the highest methylation intensity.

CpG islands can be defined in a number of ways, one of which is based on the Gardiner-Garden criteria: (i) with at least 200 base pairs (bp), (ii) with a GC content >50%, and (iii) with an observed/expected CpG ratio >60% [10]. When applying the Gardiner-Garden criteria on the human genome, we also excluded the repetitive sequence fragments (such as the Alu repeats, which are GC rich and with high CpG observed-to-expected ratio). The methylation intensity of a CpG island was considered as the average methylation intensities of all CpG dinucleotides contained in the island. For statistical reliability, we only considered those CpG islands with more than 10% CpG dinucleotides being measured the methylation intensity levels, and defined *unmethylated* CpG islands as those whose average methylation intensities are less than 10% while *methylated* CpG islands as those whose average methylation intensities are larger than 50% [8].

III. METHODS

A. Feature Extraction

It has been shown that the CpG island methylation status is correlated with the following features: CpG island specific attributes (e.g. length, GC content, GC observed/expected ratio) [11] [12] [7], patterns of DNA sequence composition [2] [12] [5], patterns of predicted DNA structure [11] [5],

patterns of repetitive elements [11] [12] [7] [5], patterns of TFBS, patterns of evolutionarily conserved elements [11], as well as the methylation status of nearby histones [8]. Computational prediction of CpG island methylation status based on the statistical properties of these features could render fairly reasonable accuracy (e.g., ~89% [2] [8]). In this study we incorporated three more sets of attributes that have not been extensively explored, including (i) the nucleosome positioning propensities of the CpG island, (ii) the acetylation status of nearby histones, and (iii) the functional roles of nearby genes. These attributes are promising to add more dimensions of information, because an accumulating body of evidence has shown that DNA methylation is influenced by nucleosome positioning [13], associated with histone acetylation [14], and involved in biological processes such as gene imprinting, X chromosome inactivation, and tumor suppressor gene silencing [15] [16]. In the following paragraphs of A.1 to A.6., we describe how these features were extracted.

A.1. The CpG island specific attributes, including the GC content, length and observed/expected CpG ratio, were directly obtained from UCSC human genome browser.

A.2. We considered the DNA composition and structure of each CpG island. For the DNA compositional features, we focused on the frequencies of the tetramer oligonucleotides and their z-scores; and, for the DNA structural features, we focused on those basic characteristics capturing the DNA 3-D conformation as well as the nucleosome positioning propensities.

The z-score of a tetramer oligonucleotide fragment, $N_1N_2N_3N_4$, was calculated as:

$$Z(N_1N_2N_3N_4) = \frac{O(N_1N_2N_3N_4) - E(N_1N_2N_3N_4)}{\sigma(N_1N_2N_3N_4)} \quad (1)$$

where $O(\cdot)$ represents the observed frequency, $E(\cdot)$ and $\sigma(\cdot)$ represent the expected frequency and standard deviation. $E(N_1N_2N_3N_4)$ was estimated empirically based on a maximal-order Markov model [17]:

$$E(N_1N_2N_3N_4) = \frac{O(N_1N_2N_3)O(N_2N_3N_4)}{O(N_2N_3)} \quad (2)$$

and $\sigma(N_1N_2N_3N_4)$ was approximated as:

$$\sigma(N_1N_2N_3N_4) = E(N_1N_2N_3N_4) * \frac{[O(N_2N_3) - O(N_1N_2N_3)][O(N_2N_3) - O(N_2N_3N_4)]}{O^2(N_2N_3)} \quad (3)$$

The DNA conformation related attributes include twist, tilt, roll, shift, slide and rise, which were estimated based on a model of dinucleotide stiffness [18]. For each of these six attributes, the average value over all dinucleotides of the CpG island was used.

Nucleosome positioning propensities of the CpG islands were estimated based on the genome-wide prediction of the nucleosome organization map [19]. There were two types of predictions, one at the nucleotide level, and the other at the DNA fragment level. The nucleotide level prediction regards the probability of each nucleotide being covered by

any nucleosome, based on which we calculated the mean and standard deviation over the entire CpG island. The fragment level prediction regards the nucleosome positioning potential of each 147 bp (the typical length of a nucleosome) DNA fragment, based on which we calculated the mean and standard deviation over all fragments overlapping with the CpG island.

A.3. We also considered the distribution patterns of the functional or evolutionarily conserved elements in the chromosomal region flanking the CpG island, where the functional elements refer to the TFBS that are conserved in human, mouse and rat genomes [20], and the evolutionarily conserved elements are those that are conserved across vertebrate, insect, worm and yeast genomes [21]. To account for both the short- and long-range association between these elements and CpG islands, we considered flanking regions of various lengths, ranging from 100 bps to 2,000 bps (with step size of 100 bps) upstream and downstream of the CpG island. Each TFBS or evolutionarily conserved element is characterized by a score quantifying its degree of conservativeness across genomes. We counted the number of these elements overlapping with the CpG island, and calculated their average score.

A.4. We examined whether a CpG island's nearby genes are involved in any cancer-related biological processes. A CpG island's nearby genes refer to those whose promoter region (from the 1,000 bps upstream to the 200 bps downstream of the transcription start site) overlaps with the CpG island. 37 biological processes (30 oncogene related, 11 tumor suppressor related, and 4 common) were determined through gene ontology enrichment analysis of the genes retrieved from the Cancer Gene Census [22]. If the gene ontology annotations of a gene include one or more of these processes, the corresponding gene function feature is 1 and 0 otherwise.

A.5. We considered the methylation and acetylation statuses of each CpG island's nearby histones. The histone methylation information was obtained from Barkski et al.'s data set, which characterizes the genome wide distribution of 20 histone methylations as well as histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF in CD4 lymphocytes [23]. The histone acetylation information was obtained from Wang et al.'s data set [24], which characterizes the genome-wide patterns of 18 histone acetylations in CD4 lymphocytes. In both data sets, a nucleotide is tagged if its nearby histone undertakes a methylation or acetylation modification; hence, the number of tags at each nucleotide can be interpreted as being proportional to the modification level of nearby histones. We used the average and standard deviation of the number of tags over all nucleotides of a CpG island to represent the methylation (or acetylation) level of the CpG island's nearby histones.

B. Feature Selection through Statistical Tests and Principal Component Analysis

A total number of 841 features were extracted for each CpG island, including three CpG island-specific attributes,

512 DNA compositional features and 10 DNA structural features of the CpG island, 230 about the distribution of TFBS and two about the distribution of the evolutionarily conserved elements in the flanking chromosomal region, two about the involvement of the neighboring genes in oncogene or tumor-suppressor related processes, and 82 about the methylation and acetylation status of nearby histones. The extraction of these features was biologically motivated. However, from a statistical point of view, the correlations of these features to the CpG island methylation status vary from one feature to another. For instance, it was reported that DNA sequence composition patterns, distribution of repeat elements, and DNA structure properties are highly or moderately correlated with the CpG island methylation status; whereas the distribution of genes, single nucleotide polymorphism, and CpG island distribution are only weakly correlated with the CpG island methylation status [5]. To screen out the features of predictive power, we performed various statistical tests, including the Fisher's exact test [25], Chi-squared test [26], and Kolmogorov-Smirnov (KS) test [27]. The Fisher's exact test was used for functional roles of nearby genes, for which the feature variable is categorical and some expected values in the contingency tables are extremely small (<5); the Chi-squared test with Yates corrections [28] was used for the other categorical features (i.e., the number of functional and evolutionarily conserved elements in the flanking chromosomal region); and, the KS test was used for those features whose values are continuous, including CpG island specific attributes, tetramer frequencies and z-scores, DNA structural features, scores of functional and evolutionarily conserved elements, and scores of histone methylation and acetylation. For each of these statistical tests, a feature was considered to be statistically significantly correlated with the methylation status of CpG islands if its p -value was less than 0.05.

Besides their correlations with the CpG island methylation status, these features might be inter-correlated. For example, the histone methylation and acetylation status are likely to be correlated, because some acetylation and methylation (e.g. histone H3 at lysine 9) play opposite roles in gene activity [29]; DNA sequence and structure properties are likely to be correlated, because most DNA structures are predicted based on DNA sequences; and, the distribution of functional/evolutionarily conserved elements in a short flanking neighborhood (e.g., ± 200 bps) is likely to be correlated with the distribution in a longer flanking neighborhood (e.g., ± 2000 bps). The correlation between features makes the feature space unnecessarily high-dimensional. To minimize the redundancy in the features, we performed the PCA on those CpG island methylation-related features that were selected via the above statistical tests.

C. Prediction Test

The features selected through statistical tests and PCA were used to build support vector machine-based models to predict the CpG island methylation status. To examine the contribution of the newly added features as well as the impact

of the inhibitive-to-acquire histone modification information, we established the following predictive models, (1) M_1 : a model with all information being incorporated, (2) M_2 : a model with all but the histone modification information being incorporated, (3) M_3 – M_9 : seven models with individual or combinations of the newly added features being excluded, and (4) M_{10} – M_{16} : seven models with individual or combinations of the newly added features as well as the histone methylation information being excluded. We used the CD4 lymphocyte data for training and validating the models, while the data of the other 11 tissues/cell types for generalizability testing.

Training/Validation (based on the CD4 lymphocyte data): All these models were trained and validated by using a 10-fold cross validation scheme. That is, all CpG islands were partitioned randomly into 10 approximately equally-sized folds, each of which was used in turn for validation while the remaining folds were used for training. The performance of the classifiers was assessed by using three metrics defined in Eqns. (4)–(6), namely, sensitivity (SE), specificity (SP), and accuracy (ACC). This partition-training-and-validation procedure was repeated for 20 times, and the classifier performance was averaged over the 200 validation folds.

$$SP = \frac{\text{\#correctly classified unmethylated CpG islands}}{\text{\#unmethylated CpG islands}} \quad (4)$$

$$SE = \frac{\text{\#correctly classified methylated CpG islands}}{\text{\#methylated CpG islands}} \quad (5)$$

$$ACC = \frac{\text{\#correctly classified CpG islands}}{\text{\#CpG islands}} \quad (6)$$

For fair comparisons with the existing method, a leave-one-out cross-validation (LOOCV) scheme was also used. That is, each CpG island was in turn used for validation while the remaining CpG islands were used for training. The performance of the model in the LOOCV scheme was also assessed by the three metrics averaged over all validation CpG islands.

Generalizability testing (based on data of other tissue/cell types): Two predictive models built on the CD4 lymphocyte data were tested for generalizability using the data of the other 11 tissues and cell types: one (M_1) relying on all information, while the other (M_2) relying on all but the histone modification information. For the former model, because the genome-wide histone methylation and acetylation profiles are not available for these 11 tissues and cell types, we used the genome-wide histone modification profiles in the CD4 lymphocytes, assuming that histone modifications in various cell types are moderately or even highly correlated [41].

IV. RESULTS AND DISCUSSIONS

A. Statistical Tests and PCA

Out of a total number of 841 features, 342 features were retained whose p -values in the statistical tests were less

than 0.05. These features include two of the CpG island specific attributes, 217 DNA compositional and eight DNA structural features, 35 functional element features and two evolutionarily conserved element features, two features regarding the functional roles of the neighboring genes, and 76 features related to the modification status of nearby histones. Particularly, among the newly added features, two out of the four nucleosome positioning features, all of the 36 histone acetylation features, and both of the features regarding the functional roles of the neighboring genes were retained after statistical tests.

PCA was performed to decorrelate these 342 selected features. Table I summarizes the number of principal components that must be retained to keep a certain percentage of the variance of the original feature space. Observe that the first eight principal components together can account for the $\sim 99.90\%$ of the variance in the original feature space and were therefore used to build the predictive models. Fig. 1 depicts the contribution of each of the 342 original feature dimensions to the eight principal components. Observe from Fig. 1 that each of the following eight categories of features, (i) the CpG island specific attributes, (ii) DNA sequence patterns, (iii) DNA structure patterns, (iv) distribution of TFBS, (v) distribution of the evolutionarily conserved elements, (vi) gene functions, (vii) histone methylation and (viii) histone acetylation status, makes substantial contributions to one or more principal components, suggesting that these categories of information, though correlated, are complementary to a certain extent for predicting the CpG island methylation.

TABLE I
NUMBER OF PRINCIPAL COMPONENTS (PCs) REQUIRED TO RETAIN A CERTAIN PERCENTAGE (PCNT) OF THE TOTAL VARIANCE.

Pcnt	100%	99.99%	99.90%	99.00%
PCs	342	10	8	6
Pcnt	95.00%	90.00%	75.00%	50.00%
PCs	5	4	3	2

B. Performance of the Predictive Models Based on the CD4 Lymphocyte Data

The specificity, sensitivity, and accuracy measures of our predictive model M_1 that incorporates all information are summarized in Table II. Observe that both cross-validation schemes rendered similar results, indicating that these measures can reliably characterize our model. The performance of our classifier was compared to that of Fan et al.'s [8] method. Note that both models incorporated the histone modification information. Observe that our model showed an improved specificity and accuracy than Fan et al.'s model while maintaining a comparable sensitivity. Furthermore, it was reported in [8] that when evaluated on the human brain data, Fan et al.'s method could outperform Epigraph [6].

We could argue that the improvement of our model M_1 over the existing model was partly due to the incorporation of the three new types of features – nucleosome positioning propensities, gene functions, and histone acetylation status.

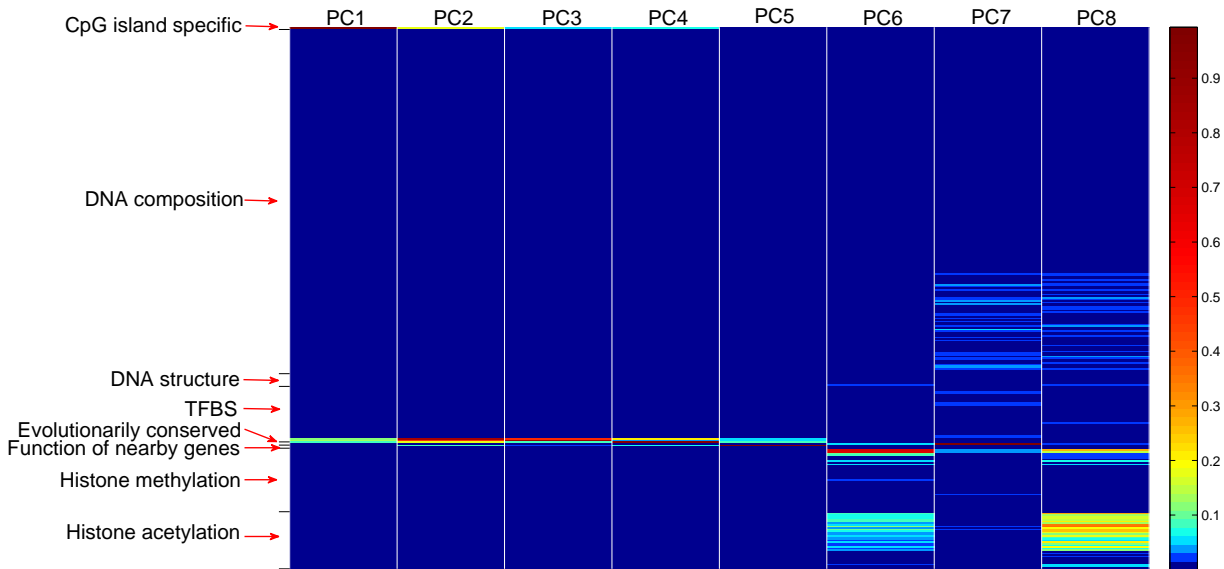


Fig. 1. Contribution of the 342 features to the eight principal components. Each column corresponds to a principal component, and each row corresponds to an original feature dimension.

The performance of our models M_3 through M_9 , each with an individual or a combination of the new types of features being excluded, are summarized in Table III. Observe that the performance of the predictive model deteriorated to different extents when individual or combinations of the newly added features were excluded. Specifically, the models without histone acetylation information (M_3 , M_6 , M_7 , and M_9) deteriorated more than those models with histone acetylation information but without the other two types of newly added features (M_4 , M_5 , and M_8). Therefore, histone acetylation appears to be the most influential feature to the performance of the predictive model among the newly added features.

We suspected that the information carried by the histone methylation features was too dominant to fairly assess the influence of these newly added features; and therefore excluded the histone methylation features and repeated the above experiments excluding individual or combinations of the newly added features. The resultant models were M_{10} through M_{16} , and their performance was summarized in Table III. Similarly, the models without an individual or a combination of the newly added features deteriorated. It is noteworthy that (1) the histone methylation and acetylation information greatly affected the sensitivity of the models, and (2) the loss of histone methylation information could largely be made up by including the histone acetylation information. This is not surprising, given that these two forms of histone modifications are closely related as repeatedly observed in various tissues and cell types [29].

C. Classifier Generalizability

The two predictive models, one with the histone modification information (M_1) and the other without (M_2), that were both built on the human CD4 lymphocyte data were tested on the data of the other 11 tissue and cell types for their

TABLE II
PERFORMANCE OF OUR CLASSIFIERS M_1 ON CD4 LYMPHOCYTES WITH COMPARISON TO THE EXISTING METHOD.

Method	SP	SE	ACC
M_1 (10-fold)	0.9405	0.9257	0.9313
M_1 (LOOCV)	0.9429	0.9307	0.9403
Fan et al.'s [8]	0.7400	0.9428	0.8994

generalizability. The sensitivity, specificity, and accuracy of M_1 and M_2 during these testing experiments are summarized in Tables IV and V.

When the histone modification information was incorporated, the classifier model built on the CD4 lymphocyte data can be applied to most of the other tissues and cell types (except for sperm) with little or no performance deterioration. When the histone modification information was not used, the performance of the predictive model on the data of the other tissues and cell types deteriorated substantially, especially in terms of the sensitivity. However, if compared to the validation results where the histone modification information was not used, the performance on the testing data was not unexpected. Therefore, with or without the histone modification information, the predictive model established on the CD4 lymphocyte data can well generalize to the other tissue or cell type data.

Considering that DNA methylation is heavily involved in cellular differentiation, our results in Tables IV and V look suspicious. We therefore calculated the correlations of the CpG island methylation levels between different tissue and cell types, as depicted in Fig. 2. Observe that the correlation coefficients between the somatic/placenta cells are very high (mean: 0.9455, standard deviation: 0.0229), where the correlation coefficients between the somatic/placenta and sperm

TABLE III

PERFORMANCE OF THE PREDICTIVE MODELS (M_3 THROUGH M_{16}), EACH WITH AN INDIVIDUAL OR A COMBINATION OF THE NEWLY ADDED CATEGORIES OF FEATURES BEING EXCLUDED.

	Features	SP		SE		ACC	
		LOOCV	10-fold	LOOCV	10-fold	LOOCV	10-fold
Histone Methylation Retained	All retained	0.9429	0.9405	0.9307	0.9257	0.9403	0.9313
	Acetylation (M_3)	0.9048	0.9012	0.9010	0.8965	0.9175	0.9046
	Functional roles (M_4)	0.9319	0.9302	0.9315	0.9265	0.9362	0.9210
	Nucleosome (M_5)	0.9285	0.9270	0.9276	0.9250	0.9205	0.9205
	Acetylation + Functional roles (M_6)	0.8876	0.8791	0.8912	0.8903	0.8915	0.8897
	Acetylation + Nucleosome (M_7)	0.8805	0.8698	0.8815	0.8835	0.8902	0.8826
	Functional roles + Nucleosome (M_8)	0.9208	0.9186	0.9107	0.9116	0.9202	0.9186
All three (M_9)	0.8775	0.8685	0.8810	0.8822	0.8806	0.8786	
Histone Methylation Excluded	All but histone methylation	0.9321	0.9318	0.5941	0.5932	0.8593	0.8575
	Acetylation (M_{10})	0.9701	0.9670	0.2277	0.2247	0.8102	0.8001
	Functional roles (M_{11})	0.9109	0.9092	0.5720	0.5670	0.8369	0.8312
	Nucleosome (M_{12})	0.9088	0.9078	0.5682	0.5660	0.8298	0.8296
	Acetylation + Functional roles (M_{13})	0.9402	0.9320	0.2289	0.2279	0.7885	0.7862
	Acetylation + Nucleosome (M_{14})	0.9381	0.9266	0.2302	0.2304	0.7752	0.7641
	Functional roles + Nucleosome (M_{15})	0.9012	0.8990	0.5520	0.5519	0.8252	0.8232
	All three (M_{16})	0.9098	0.8972	0.2341	0.2338	0.7406	0.7352

TABLE IV

PERFORMANCES OF THE CLASSIFIER MODEL BUILT ON THE DATA OF 11 DIFFERENT TISSUES AND CELL TYPES: WITH HISTONE MODIFICATION.

Procedure	Tissue/Cell Type	SP	SE	ACC
Validation	CD4 (10-fold)	0.9405	0.9257	0.9313
	CD4 (LOOCV)	0.9429	0.9307	0.9403
Testing	CD8	0.9608	0.8932	0.9448
	liver	0.9680	0.8762	0.9465
	heart muscle	0.9462	0.9479	0.9466
	skeletal muscle	0.9542	0.9451	0.9524
	embryonic skeletal	0.9395	0.9367	0.9389
	embryonic liver	0.9259	0.9342	0.9277
	placenta	0.9695	0.9130	0.9571
	dermal melanocytes	0.9663	0.8785	0.9446
	dermal fibroblasts	0.9525	0.9239	0.9467
	dermal keratinocytes	0.9385	0.9341	0.9376
	sperm	0.8459	0.9778	0.8617

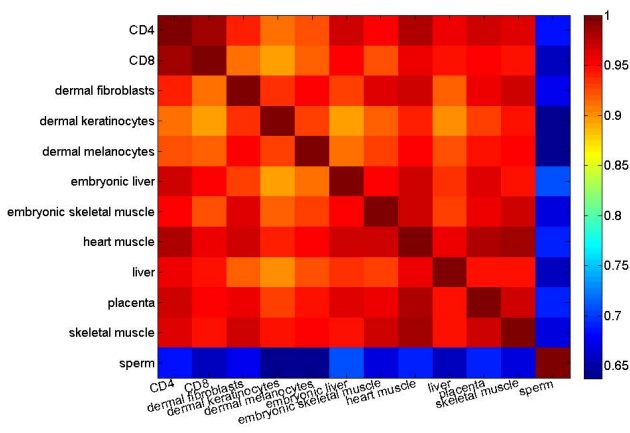


Fig. 2. Correlation coefficients of the CpG island methylation levels across different tissues and cell types.

cells are only moderate (mean: 0.6706, standard deviation: 0.0225). This suggests that the methylation status of CpG islands are highly correlated in various somatic/placenta cells, and therefore do not represent tissue-specific differentially methylated regions. Our observations are consistent with recent studies [30] [31] that there are few variance in methylation levels of autosomal CpG island promoters, and there is only a relatively small fraction of CpG islands with tissue-specific methylation. The difference between the somatic/placenta and sperm cells, as reflected by their moderate cross-correlations and the performance deteriorations of our prediction models being applied to the sperm cell data, suggests that gametes are epigenetically more deviated from somatic cells than somatic cells themselves. This difference is likely related to the meiotic process, the special conditions and gene expression required for gamete production [32].

V. CONCLUSIONS AND FUTURE WORKS

The establishment of DNA methylation pattern is a crucial part of cell differentiation and organ development, suppres-

TABLE V
PERFORMANCES OF THE CLASSIFIER MODEL ON THE DATA OF 11 DIFFERENT TISSUES AND CELL TYPES: WITHOUT HISTONE MODIFICATION.

Procedure	Tissue/Cell Type	SP	SE	ACC
Validation	CD4 (10-fold)	0.9670	0.2247	0.8001
	CD4 (LOOCV)	0.9701	0.2277	0.8102
Testing	CD8	0.9722	0.2108	0.8104
	liver	0.9678	0.2143	0.8122
	heart muscle	0.9562	0.2386	0.8186
	skeletal muscle	0.9594	0.2364	0.8306
	embryonic skeletal	0.9425	0.2298	0.8100
	embryonic liver	0.9389	0.2306	0.8054
	placenta	0.9655	0.2184	0.8276
	dermal melanocytes	0.9700	0.2186	0.8156
	dermal fibroblasts	0.9605	0.2200	0.8237
	dermal keratinocytes	0.9425	0.2204	0.8095
	sperm	0.8524	0.2365	0.7625

sion of viral genes and deleterious elements, and carcinogenesis. Computational prediction of DNA methylation levels provides an effective, fast and cheap alternative approach for studying the DNA methylation patterns. In this study, we performed the computational prediction of the CpG island methylation by incorporating additional features and effec-

tively selecting and decorrelating the features. We incorporated the information regarding the nucleosome positioning propensity, acetylation status of nearby histones, and the functional roles of nearby genes. These features were first screened through statistical tests and PCA. The most DNA methylation-relevant yet non-intercorrelated features were subsequently used to build computational models to predict the methylation status of CpG islands. Our experiments on the HEP data set demonstrated that (1) an eight-dimensional feature space, which combines all the eight categories of information, was effective in predicting the methylation status of CpG islands; (2) by incorporating the information regarding the nucleosome positioning propensities, gene functions, and histone acetylation, our predictive model achieved a higher specificity and accuracy than the existing model while maintaining a comparable sensitivity; (3) the histone modification attributes carry a weight of information for the prediction, without which the performance of the predictive model deteriorated substantially in terms of sensitivity; (4) with or without the histone modification information the performance of the predictive models are consistent on the validation and testing data. This computational model, with its evidently high specificity and sensitivity, provides an effective tool for identification of new methylation targets and therefore lays foundation for our future endeavors in the regulation mechanisms of DNA methylation.

REFERENCES

- [1] A. Bird, "CpG-rich islands and the function of DNA methylation," *Nature*, vol. 321, pp. 209–213, 1986.
- [2] F. Feltus, E. Lee, J. Costello, C. Plass, and P. Vertino, "Predicting aberrant CpG island methylation," *Proceedings of the National Academy of Sciences USA*, vol. 100, pp. 12253–12258, 2003.
- [3] M. Ehrlich, M. Gama-Sosa, L. Huang, R. Midgett, K. Kuo, R. McCune, and C. Gehrke, "Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells," *Nucleic Acids Research*, vol. 10, pp. 2709–2721, 1982.
- [4] M. Bhasin, H. Zhang, E. Reinherz, and P. Reche, "Prediction of methylated CpGs in DNA sequences using a support vector machine," *FEBS Lett*, vol. 579, pp. 4302–8, 2005.
- [5] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter, "CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure," *PLoS Genetics*, vol. 2, p. e26, 2006.
- [6] C. Bock, J. Walter, M. Paulsen, and T. Lengauer, "CpG island mapping by epigenome prediction," *PLoS Computational Biology*, vol. 3, p. e110, 2007.
- [7] F. Fang, S. Fan, X. Zhang, and M. Zhang, "Predicting methylation status of CpG islands in the human brain," *Bioinformatics*, vol. 22, pp. 2204–2209, 2006.
- [8] S. Fan, M. Zhang, and X. Zhang, "Histone methylation marks play important roles in predicting the methylation status of CpG islands," *Biochemical and Biophysical Research Communications*, vol. 374, pp. 559–564, 2008.
- [9] F. Eckhardt, J. Lewin, R. Cortese, V. Rakyan, J. Attwood, M. Burger, J. Burton, T. Cox, R. Davies, T. Down, C. Haefliger, R. Horton, K. Howe, D. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck, "DNA methylation profiling of human chromosomes 6, 20 and 22," *Nature Genetics*, vol. 38, pp. 1378–1385, 2006.
- [10] M. Gardiner-Garden and M. Frommer, "CpG islands in vertebrate genomes," *Journal of molecular biology*, vol. 196, pp. 261–282, 1987.
- [11] C. Previti, O. Harari, I. Zwir, and C. del Val, "Profile analysis and prediction of tissue-specific CpG island methylation classes," *BMC Bioinformatics*, vol. 10, p. 116, 2009.
- [12] R. Das, N. Dimitrova, Z. Xuan, R. Rollins, F. Haghghi, J. Edwards, J. Ju, T. Bestor, and M. Zhang, "Computational prediction of methylation status in human genomic sequences," *Proc Natl Acad Sci USA*, vol. 22, pp. 10713–10716, 2006.
- [13] R. Chodavarapu, S. Feng, Y. Bernatavichute, P. Chen, H. Stroud, Y. Yu, J. Hetzel, F. Kuo, J. Kim, S. Cokus, D. Casero, M. Bernal, P. Huijser, A. Clark, U. Kramer, S. Merchant, X. Zhang, S. Jacobsen, and M. Pellegrini, "Relationship between nucleosome positioning and DNA methylation," *Nature Letter*, vol. 466, pp. 388–392, 2010.
- [14] J. Dobosy and E. Selker, "Emerging connections between DNA methylation and histone acetylation," *Cell Mol Life Sci*, vol. 58, pp. 721–727, 2001.
- [15] Y. Yamada, H. Watanabe, F. Miura, H. Soejima, M. Uchiyama, T. Iwasaka, T. Mukai, Y. Sakaki, and T. Ito, "A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q," *Genome Research*, vol. 14, pp. 247–266, 2004.
- [16] D. Hanahan and R. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, 2000.
- [17] S. Schbath, B. Prum, and E. Turckheim, "Exceptional motifs in different markov chain models for a statistical analysis of DNA sequences," *Journal of Computational Biology*, vol. 2, pp. 417–437, 1995.
- [18] J. Goñi, A. Pérez, D. Torrents, and M. Orozco, "Determining promoter location based on DNA structure first-principles calculations," *Genome Biology*, vol. 8, p. R263, 2007.
- [19] N. Kaplan, I. Moore, Y. Fondudfe-Mittendorf, A. Gossett, D. Tillo, Y. Field, E. LeProust, T. Hughes, J. Lieb, J. Widom, and E. Segal, "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature Letter*, vol. 458, pp. 362–366, 2009.
- [20] D. Karolchik, R. Baertsch, M. Diekhans, T. Furey, A. Hinrichs, Y. Lu, K. Roskin, M. Schwartz, C. Sugnet, D. Thomas, R. Weber, D. Haussler, and W. Kent, "The UCSC genome browser database," *Nucleic Acids Res*, vol. 31, pp. 51–54, 2003.
- [21] A. Siepel, G. Bejerano, J. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. Hillier, S. Richards, G. Weinstock, R. Wilson, R. Gibbs, W. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, vol. 15, pp. 1034–1050, 2005.
- [22] P. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. Stratton, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, pp. 177–183, 2004.
- [23] A. Barski, S. Cuddapah, K. Cui, T. Roh, D. Schones, Z. Wang, G. Wei, I. Chepelev, and Z. K., "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, pp. 823–837, 2007.
- [24] Z. Wang, C. Zang, J. Rosenfeld, D. Schones, A. Barski, S. Cuddapah, K. Cui, T. Roh, W. Peng, M. Zhang, and K. Zhao, "Combinatorial patterns of histone acetylations and methylations in the human genome," *Nature Genetics Letter*, vol. 40, pp. 879–903, 2008.
- [25] A. Agresti, "A survey of exact inference for contingency tables," *Proceedings of the National Academy of Sciences USA*, vol. 7, pp. 131–153, 1992.
- [26] N. Turner, "Chi-squared test," *Journal of Clinical Nursing*, vol. 9, p. 93, 2000.
- [27] G. Marsaglia, W. Tsang, and J. Wang, "Evaluating kolmogorov's distribution," *Journal of Statistical Software*, vol. 8, pp. 1–4, 2003.
- [28] J. Freeman and S. Julious, "The analysis of categorical data," *Scope*, vol. 16, pp. 18–21, 2007.
- [29] K. Zhang, J. Sino, P. Jones, P. Yau, and E. Bradbury, "A mass spectrometric western blot to evaluate the correlations between histone methylation and histone acetylation," *Proteomics*, vol. 4, pp. 3765–3775, 2004.
- [30] M. Weber, I. Hellmann, M. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler, "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome," *Nature Genetics*, vol. 39, pp. 457–466, 2007.
- [31] R. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. Potash, S. Sabuncyan, and A. Feinberg, "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores," *Nature Reviews Cancer*, vol. 41, pp. 178–186, 2009.
- [32] P. Nawapen, S. Junpen, H. Dion, D. Michael, C. Bernie, and T. Mongkol, "Different DNA methylation patterns detected by the Amplified Methylation Polymorphism Polymerase Chain Reaction (AMP PCR) technique among various cell types of bulls," *Acta Veterinaria Scandinavica*, vol. 52, p. 18, 2010.

Paralellization of Needleman-Wunsch String Alignment Method

Jaime Seguel

Department of Electrical and Computer Engineering
University of Puerto Rico at Mayaguez
Mayaguez, Puerto Rico
Jaime.seguel@upr.edu

Carlos Torres

Department of Electrical and Computer Engineering
University of Puerto Rico at Mayaguez
Mayaguez, Puerto Rico
Carlos.torres9@upr.edu

Abstract— The identification of homologies between protein sequences is a central problem in molecular biology and several algorithms have been proposed for accomplishing this task. The Needleman-Wunsch Algorithm and its close variant, the Smith-Waterman dynamic programming methods, solve the problem exactly in quadratic time and space. However, due to the massive amount of data involved in sequence-to-database of sequences comparisons, heuristic methods such as BLAST, are used instead. This article explores the design of a parallel version of Needleman-Wunsch based on a divide-and-conquer idea derived from an algorithm first proposed by Hirschberg, for the identification of the longest common substring of two strings.

Keywords: string alignment, dynamic programming, divide-and-conquer, recursion, parallel computation.

I. INTRODUCTION

The identification of homologies between protein sequences is a central problem in molecular biology. In computational terms, the problem is stated as the search of an optimal alignment between a pair of strings over the 20-character amino acid alphabet. An alignment of a pair of strings $S_1=x_1\dots x_m$ and $S_2=y_1\dots y_n$, is a $2 \times q$ matrix $A[i, j]$, where $i = 1, 2$ and $q \geq \max\{m, n\}$; whose entries are the characters of the amino acid alphabet or a gap symbol “-”. In addition, A satisfies:

- i. For each $i = 1, 2$ and j , $A[i, j]$ is a character of the protein alphabet or either $A[1, j] = -$ or $A[2, j] = -$, but not both;
- ii. If all gap symbols are removed and each empty cell is filled by left shifting by one all the cells to its right, then $A[1, j] = x_j$ and $A[2, j] = y_j$

The similarity is quantified with respect to a substitution matrix that establishes the rate at which a character in the amino acid alphabet changes to each of the other characters in the same alphabet, over time. We denote the substitution rate of pair of characters x_k and y_j as $r(x_k, y_j)$. A gap penalty, in turn, is a numerical cost imposed for the insertion of the gap symbol in either $A[1, j]$ or $A[2, j]$. Gap penalties may be assigned as a constant per gap symbol inserted, or as an affine gap penalty containing separate costs for initiating and for

extending the gap with consecutive gap symbol insertions. For the sake of simplicity, we consider solely constant gap penalties. The score of an alignment is the sum of the rate of substitution of each pair of aligned characters minus a gap penalty per each character that is aligned with the gap symbol. The higher the score, the better the alignment. For example, let's consider the alignment of the pair (S_1, S_2) ; with $S_1 = \text{PAWHEAE}$ and $S_2 = \text{HEAGAWGHEE}$ [1]. The next alignment is optimal with respect to the BLOSUM50 substitution matrix and a penalty gap of $\gamma = -8$.

TABLE I. AN OPTIMAL ALIGNMENT OF S_1 AND S_2

H	E	A	G	A	W	G	H	E	-	E
-	-	P	-	A	W	-	H	E	A	E

This alignment, which reaches the optimal score of 1, is not necessarily unique. Indeed, in this particular case as in many others, there are other alignments that reach the same score.

A. Dynamic Programming

An exhaustive search for an optimal alignment takes exponential time. Needleman and Wunsch [2] introduced a quadratic time and space dynamic programming method known today as the Needleman-Wunsch algorithm (NWA). NWA computes an optimal alignment in two main steps. The first step uses a recursive formula to fill in, usually row-by-row or column-by-column; a dynamic programming matrix D . The next pseudo-code describes this process.

Step 1: Computation of the Dynamic Programming Matrix

For each k , $0 \leq k \leq m$;

$D[k, 0] \leftarrow k \times \gamma$

For each j , $0 \leq j \leq n$;

$D[0, j] \leftarrow j \times \gamma$

For each k , $1 \leq k \leq m$;

For each j , $1 \leq j \leq n$;

$D[k, j] \leftarrow \max\{D[k-1, j-1] + r(x_k, y_j), D[k-1, j] + \gamma, D[k, j-1] + \gamma\}$

Matrix D contains the optimal scores for the alignment of all subsequences of S_1 and S_2 that start at x_1 and y_1 ;

respectively. The last entry in this matrix, this is $D[m, n]$, is the optimal score for the alignment of S_1 and S_2 . The second step of NWA backtracks the path of solutions of the subsequence alignments that led to the optimal score $D[m, n]$. We refer to a path of indices of D that results from this process as backtrack-path or b-path. Backtracking is summarized in the next pseudo code.

Step 2: Computation of a b-path

$b\text{-path} \leftarrow \{[m, n], [0, 0]\}$

While $[k, j] \neq [0, 0]$ is in $b\text{-path}$

 If $D[k, j] = D[k-1, j-1] + r(x_k, y_j)$

$b\text{-path} \leftarrow b\text{-path} \cup \{[k-1, j-1]\};$

 Else if $D[k, j] = D[k-1, j] + \gamma$

$b\text{-path} \leftarrow b\text{-path} \cup \{[k-1, j]\};$

 Else if $D[k, j] = D[k, j-1] + \gamma$

$b\text{-path} \leftarrow b\text{-path} \cup \{[k, j-1]\};$

The actual alignment follows from the b-path by applying the next rule:

If $[k-1, j-1]$ and $[k, j]$ are in the b-path, x_k is aligned with y_j

If $[k-1, j]$ and $[k, j]$ are in the b-path, x_k is aligned with –

If $[k, j-1]$ and $[k, j]$ are in the b-path, y_j is aligned with –

NWA solves the global alignment problem, which is, the optimal alignment of the whole S_1 and S_2 input sequences. However, in many biological instances, sequences share only segments of meaningful similarity. These may vary from short regions to large domains of recognizable similarity. The so-called local alignment problem is the problem of identifying the segments in S_1 and S_2 with the highest alignment score. An exact algorithmic solution of the local alignment problem is the Smith-Waterman algorithm (SWA) due to Smith and Waterman [3]. The SWA is a variant of the NWA that replaces negative scores with zeroes masking thus, segments of low similarity score. SWA dynamic programming and backtracking steps are similar to the ones in NWA except that backtracking starts from the indices of the entry with the highest score in the whole matrix D , and ends right before the first 0 encountered while performing the backtracking process. Thus, unlike b-paths, local backtracking paths or lb-path are not necessarily anchored in $[m, n]$ and $[0, 0]$. This difference has deep algorithmic consequences.

B. Some Previous Parallelization Attempts

Filling D with either NWA or SWA takes $O(mn)$ time and space. Backtracking, in turn, is accomplished in $O(m+n)$ time. Although most implementations do not store D , the information needed to backtrack the optimal alignment still takes $O(mn)$ space. Due to the large size of protein sequences and protein databases, SWA is often replaced with the heuristic BLAST [4] (Basic Local Alignment Sequence Tool) algorithm. Unlike BLAST, whose heuristics is well suited for parallelization; the parallelization of NWA and SWA is limited by the recursive nature of their core processes. Some NWA and SWA parallelization attempts exploit the fact that each entry in D depends solely on its northern, western and northwestern

neighbors. Thus, entries lying in the same anti-diagonal of D can be computed in parallel. This method is referred as the wave-front computation of D and is sometimes attributed to Gotoh [5] in the literature. Hsien-Yu Liao [6] et. al. use the wave-front computation to pipeline the search for an optimal alignment of a query sequence and the sequences in a database. The idea is to slide the front-wave across an enhanced scoring matrix whose rows correspond with the query sequence while the columns, to the concatenation of all the sequences in the database. Such concatenation is expected to diminish the latency incurred in starting each new comparison. Another method, introduced by Fa Zheng [7] et. al. splits the query sequence S_1 into a fixed number of sequences of approximately the same length. Each sub-sequence of S_1 is aligned with S_2 independently. A drawback in this method is that the scores matrices, which are computed in parallel, do not always correspond to sub-problems of the original alignment problem. Thus, in order to retrieve the alignment from the computed matrices, the authors propose what they called *combine and extend* method; which compromises the sensitivity of the result.

Most attempts to speedup NWA or SWA concentrate in the computation of the score matrix. This article takes a more integral approach. The starting point in our search for a parallel method is a non-recursive alternative to backtracking. The proposed alternative is based on symmetry properties that arise when matrix D is compared D^* , the dynamic programming matrix of the alignment of S_1^* and S_2^* , which are the original sequences S_1 and S_2 but in reversed order.

II. PROPERTIES OF THE DYNAMIC PROGRAMMING MATRIX

Hirschberg [8] introduced an $O(m+n)$ space algorithm for finding the longest common sub-string of a pair of strings. His method relies on the rather obvious fact that the longest common sub-string of S_1 and S_2 is the same as the longest common sub-string of S_1^* and S_2^* . Hirschberg's idea has been extended to the calculation of the edit distances between pairs of sequences and to the global alignment of a pair of sequences. The authors did not find in the literature any extension of Hirschberg's method to the parallelization of NWA. This section develops the mathematical foundations of Hirschberg-NWA space saving algorithm, and describes it in a way that makes it more suitable for the parallel NWA discussed in section III.

A. The D - D^* symmetry

The mathematical facts that allow the use of Hirschberg's ideas in the solution of the problem of the global alignment of a pair of sequences are stated below.

Lemma 1. *The optimal alignment of a pair (S_1, S_2) in reversed order is an optimal alignment for the pair (S_1^*, S_2^*) ; and vice versa.*

The next theorem is crucial in the parallelization and space saving strategy of NWA to be discussed in the next section.

Theorem 1. Let D and D^* be the score matrices produced with NWA for the pairs (S_1, S_2) and (S_1^*, S_2^*) , respectively. Then, for each $0 \leq k \leq m$ and each $0 \leq j \leq n$,

- i. $D[k, j] + D^*[m - k, n - j] \leq D[m, n]$
- ii. $D[k, j] + D^*[m - k, n - j] = D[m, n]$ if and only if $[k, j]$ is in a b -path.

Proof. The proof of assertion i. is by induction on k and j . For the base case we set $k = 0$ and $j = 0$. Thus, the statement to be proved is $D[0, 0] + D^*[m, n] \leq D[m, n]$. This statement is true because, after Lemma 1, $D[m, n] = D^*[m, n]$ and $D[0, 0] = 0$. We assume now that there is a pair of indices k, j for which $D[k, j] + D^*[m - k, n - j] \leq D[m, n]$ and prove that under this assumption the statement:

- (a) $D[k + 1, j] + D^*[m - k - 1, n - j] \leq D[m, n]$, and
- (b) $D[k, j + 1] + D^*[m - k, n - j - 1] \leq D[m, n]$, and
- (c) $D[k + 1, j + 1] + D^*[m - k - 1, n - j - 1] \leq D[m, n]$,

is also true. The proof of claim (a) is as follows. Since by definition of the NW recursion $D[k + 1, j] \leq D[k, j] + \gamma$, we have that

$$\begin{aligned} & D[k + 1, j] + D^*[m - k - 1, n - j] \leq \\ & D[k, j] + \gamma + D^*[m - k - 1, n - j] \leq \\ & D[k, j] + D^*[m - k, n - j] \leq D[m, n]. \end{aligned}$$

Sub-statement (b) is proved similarly. In order to demonstrate claim (c) let $B[k, j]$ be the entry for the pair (x_k, y_j) in the substitution matrix. Then,

$$\begin{aligned} & D[k + 1, j + 1] + D^*[m - k - 1, n - j - 1] \leq \\ & D[k, j] + B[k + 1, j + 1] + D^*[m - k - 1, n - j - 1] \leq \\ & D[k, j] + D[m - k, n - j] \leq D[m, n]. \end{aligned}$$

Assertion ii is a direct consequence of Lemma 1. \square

The index relation $([k, j], [m - k, n - j])$ is referred as D - D^* symmetry and the entries $D[k, j]$ and $D^*[m - k, n - j]$ as D - D^* symmetric entries.

B. The $O(m + n)$ Space Hirschberg-NWA

Theorem 1 provides the mathematical basis for a space saving Hirschberg-NWA (HNWA). This method, which follows the divide-and-conquer paradigm, uses Step 1 of NWA repeatedly, each time over sequences of approximately half the size of the previous ones, to divide the problem in sub-problems, until a predetermined sub-problem size is reached. Only the last column of each intermediate sub-problem dynamic programming matrices D and D^* are temporarily stored to determine the indices $[k, j]$ that satisfy statement ii of Theorem 1. Once $[k, j]$ is known, the problem is split in two sub-problems. Indeed, because of the general form of a b -path, the indices $[r, s]$ of the b -path segment from $[0, 0]$ to $[k, j]$ must satisfy $0 \leq r \leq k$. Similarly, the indices $[r, s]$ of the b -path segment from $[k, j]$ to $[m, n]$ must satisfy $k \leq r \leq m$. Thus, the search for the next indices that satisfy statement ii of Theorem 1 is reduced to the upper leftmost $k \times j$ block $D[r, s]$, $0 \leq r \leq k$, $0 \leq s \leq j$; and the lower rightmost $(m - k) \times (n - j)$ block, this

is, $D[r, s]$, $k \leq r \leq m$, $j \leq s \leq n$. This splitting identifies a pair of subsequences of S_1 and S_2 , which are aligned in the global alignment of S_1 and S_2 , except perhaps for the introduction of gaps. By selecting j as close as possible to the middle of the subsequence of S_2 , the corresponding blocks in the dynamic programming matrix are of similar size in the average case. This process of splitting sequences in subsequences, which we refer as the divide phase, is repeated on each of the newly identified subsequence up until a predetermined subsequence length is reached. The divide phase ends with the application of NWA to each of the subsequences of predetermined size. It is easy to demonstrate that the divide phase can still be performed in $O(mn)$ time, although with a higher constant. The cost in memory space, in turn, is reduced in the best case (i.e. one point subsequences) to $O(m + n)$. The conquer phase of the method is also linear in time and space as it consists basically in pasting together the b -path segments computed at the end of the divide phase.

C. A Case Study

We illustrate the base divide and conquer technique of HNWA with the problem of finding an optimal alignment for $S_1 = \text{HEAGAWGHEE}$ and $S_2 = \text{PAWHEAE}$. Before the illustration of the divide phase it is worth remarking that Step 1 of NWA can be modified to compute the rightmost column of D in-place, this is using only the storage space of one column. This is an essential element in HNWA memory space reduction strategy. The next pseudo code, which illustrates such in-place computation, uses a one-dimensional array $C[k]$, $0 \leq k \leq m$; to store intermediate and final result.

Step 1.a. In-place computation of the rightmost column of D

```

For  $k = 1$  to  $m$ 
   $Aux2 \leftarrow C[k]$ 
  If  $Aux2 + \gamma > Aux1 + r(x_k, \text{character})$ 
     $C[k] \leftarrow Aux2 + \gamma$ 
  Else
     $C[k] \leftarrow Aux1 + r(x_k, \text{character})$ 
  If  $C[k] < C[k - 1] + \gamma$ 
     $C[k] \leftarrow C[k - 1] + \gamma$ 
  If  $C[k] < 0$ 
     $C[k] \leftarrow 0$ 
   $Aux1 \leftarrow Aux2$ 
Return  $C$ 

```

We return now to the example. In the first step of the division phase, we select $j = 5$ and split sequence S_1 in two halves each of length 5. The second half is written in reversed order. Thus, we split the original problem of aligning the pair (HEAGAWGHEE, PAWHEAE) into two independent problems, namely; that of aligning the pair (HEAGA, PAWHEAE) and that of aligning the pair (EEHGW, EAEHWAP). Now, using the Step 1.a described above, we compute the rightmost columns of the dynamic programming matrices of each of these pairs of sequences. Although in practice these matrices are not store, for the sake of clarity we present in Table II the full dynamic programming matrices for these two pairs of subsequence alignments.

TABLE II. DYNAMIC PROGRAMMING MATRICES FOR (HEAGA, PAWHEAE) AND (EEHW, EAEHWAP)

Ind		0	1	2	3	4	5
	Char		H	E	A	G	A
0		0	-8	-16	-24	-32	-40
1	P	-8	-2	-9	-17	-25	-33
2	A	-16	-10	-3	-4	-12	-20
3	W	-24	-18	-11	-6	-7	-15
4	H	-32	-14	-18	-13	-8	-9
5	E	-40	-22	-8	-16	-16	-9
6	A	-48	-30	-16	-3	-11	-11
7	E	-56	-38	-24	-11	-6	-12

Ind		0	1	2	3	4	5
	Char		E	E	H	G	W
0		0	-8	-16	-24	-32	-40
1	E	-8	6	-2	-10	-18	-40
2	A	-16	-2	5	-3	-10	-18
3	E	-24	-10	4	5	-3	-11
4	H	-32	-18	-4	14	6	-2
5	W	-40	-26	-12	6	11	21
6	A	-48	-34	-20	-2	6	13
7	P	-56	-42	-28	-10	-2	3

Two border columns and rows have been added to keep track of the matrix indices and their corresponding characters in the sequences. By adding the D-D* symmetric entries of the rightmost columns of the dynamic programming matrices (i.e. the fifth column of each matrix in this case) we find that:

$$2 = \arg \max \{D[k, 5] + D^*[7 - k, 5]: 0 \leq k \leq 7\}. \quad (1)$$

Therefore, [2, 5] is in the b-path and the search for the segment of the b-path to the left of [2, 5] is reduced to the set of indices $\{[r, s]: 0 \leq r \leq 2, 0 \leq s \leq 5\}$; while the search for the segment to right of [2, 5] is reduced to $\{[r, s]: 2 \leq r \leq 7, 5 \leq s \leq 10\}$ or, in terms of D*, to $\{[r, s]: 0 \leq r \leq 5, 0 \leq s \leq 5\}$. The next step is to reduce the sequences accordingly. This gives the reduced pairs (HEAGA, PA) and (EEHW, EAEHW). At this point, the algorithm checks whether the lengths of all the latter sequence segments are less than or equal to the predetermined maximal length. If this is not the case, subsequences HEAGA and EEHW are split into two new sequences and the above process is repeated to get two new reduced pairs out of each of (HEAGA, PA) and (EEHW, EAEHW). This decomposition generates a binary tree that at each leaf has a pair of segments of the original sequences whose length is less than or equal to the predetermined length. At this point, a b-path for each pair

of segments is computed and the conquer phase started. For the sake of simplicity, let's assume that the predetermined length is 5 in the example. Then, the following dynamic programming matrices need to be computed and stored and process with Step 2 of NWA.

TABLE III. DYNAMIC PROGRAMMING MATRICES FOR (HEAGA, PA) AND (EEHW, EAEHW)

Ind		0	1	2	3	4	5
	Char		H	E	A	G	A
0		0	-8	-16	-24	-32	-40
1	P	-8	-2	-9	-17	-25	-33
2	A	-16	-10	-3	-4	-12	-20

Ind		0	1	2	3	4	5
	Char		E	E	H	G	W
0		0	-8	-16	-24	-32	-40
1	E	-8	6	-2	-10	-18	-40
2	A	-16	-2	5	-3	-10	-18
3	E	-24	-10	4	5	-3	-11
4	H	-32	-18	-4	14	6	-2
5	W	-40	-26	-12	6	11	21

By applying Step 2 of NWA to each of these matrices we get the b-paths are $\{[2, 5], [1, 4], [1, 3], [0, 2], [0, 1], [0, 0]\}$ and $\{[5, 5], [4, 4], [4, 3], [3, 2], [2, 1], [1, 1], [0, 0]\}$, respectively. And by applying the previously discussed rules for constructing alignments to each b-path we get the alignments,

H	E	A	G	A
-	-	P	-	H

and

E	-	E	H	G	W
E	A	E	H	-	W

Finally, by reversing the second alignment and concatenating it to the first one we retrieve the optimal alignment of Table 1.

It can be easily proved that all other optimal alignments are obtained from combinations of the alternative optimal alignments of each of the leaf pairs of sequence segments.

III. PARALLELIZING HIRSCHBERG-NWA

The parallelization of the previously discussed method exploits all independent computations in HNWA divide and conquers phases. These are, in summary, the computation of

the rightmost column of the dynamic programming matrix for the optimal alignment of each pair of subsequences, the computation of the b-path of each pair of subsequences of length less than or equal to the predetermined maximum length, and the production of the corresponding alignments. We use the master-workers paradigm with 2^p workers for describing the parallel method. Each worker is identified by a worker's identification number q , $1 \leq q \leq 2^p$. The master's identification number is 0.

A. Parallel HNWA

The following pseudo code is a high level description of a parallel HNWA.

Master:

$p \leftarrow 1$ (global variable)

On input (S_1, S_2)

$Aux \leftarrow S_1$

If $(length(Aux) > L$ or $length(S_2) > L)$ and $p < P$

$S_1 \leftarrow$ first half of Aux

Send (S_1, S_2) to Worker 1

$S_1 \leftarrow$ reversed second half of Aux

$S_2 \leftarrow S_2^*$

Send (S_1, S_2) to Worker 2

Receive $(Al(2^p - 1), Al(2^p))$

Concatenate $Al(2^p - 1)$ and $Al(2^p)^*$

Return

Else perform $NWA(S_1, S_2)$

Worker q :

If $1 \leq q \leq 2^{p-1}$

Receive (S_1, S_2)

Step a: Compute column C with Step 1.a on (S_1, S_2)

If $2^{p-1} < q \leq 2^p$

Send C to Worker $q - 2^{p-1}$

Else, Receive C from Worker $q + 2^{p-1}$

Compute index k in formula (1)

Send $k \leftarrow m - k$ to Worker $q + 2^{p-1}$

$S_2 \leftarrow$ First k characters of local S_2

$Aux \leftarrow$ local S_1

$p \leftarrow p + 1$ (global update)

If $(length(Aux) > L$ or $length(S_2) > L)$ and $p < P$

Local $S_1 \leftarrow$ first half of Aux

$S_1 \leftarrow$ reversed second half of Aux

$S_2 \leftarrow S_2^*$

Send (S_1, S_2) to Worker $q + 2^{p-1}$

Go to Step a

Else $Al(q) \leftarrow$ NWA alignment of (S_1, S_2)

If $1 \leq q \leq 2^{p-1}$

Send $Al(q)$ to Worker $q + 2^{p-1}$

Else Receive $A(q - 2^{p-1})$

$Al(q) \leftarrow$ Concatenation of $Al(q - 2^{p-1})$ and $Al(q)^*$

$p \leftarrow p - 1$

Send $Al(q)$ to Worker $q + 2^{p-1}$

The pseudo code imposes an additional condition for the halting of the HNWA divide phase. The divide phase stops when the length of all subsequences is less than or equal to a predetermined length $L > 0$ or when all workers are busy. If the conditions for splitting a local subsequence are met at Worker q , then Worker q keeps the first half of its local S_1 segment and

the first k characters of its local segment of S_2 to repeat the processes on them, and sends the second half and $m - k$ (local m) remaining characters of S_2 , both in reversed order, to Worker $q + 2^{p-1}$. Therefore, if for instance, $P = 2$ and the conditions for splitting the sequences are always met, the divide phase will involve 2 parallel steps. First, the master sends tasks to Worker 1 and Worker 2. When these parallel tasks are completed, Worker 1 sends a sub-task to Worker 3 and Worker 2 a sub-task to Worker 4. All four workers process their sub-tasks in parallel. So, ideally, the parallel tasks in the divide phase spawn a binary tree of height 2, rooted at the master's task. There is P parallel communications, as well. The conquer phase, in turn, traverses this tree from the leaves up in P additional parallel steps. First, workers 1, 2, 3 and 4 produce their local alignments in parallel. Then, Worker 1 sends its alignment to Worker 2 and Worker 3 sends its alignment to Worker 4. At this point, Worker 2 and Worker 4 concatenate their alignments in parallel and send the result to the master.

B. Performance Estimations

The next analysis, which is based on a highly simplified performance model, shows that the proposed parallelization has the potential to speed up the execution time of NWA. Let $t(N)$ be the execution time of the NWA on a problem of size N . Then, $t(N) = d(N) + b(N)$, where $d(N)$ is the time for the computation of the dynamic programming matrix and $b(N)$, the time for computing the b-path and forming the alignment. The P steps in the divide phase of the parallel HNWA will take approximately

$$(\frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^P})d(N) = (1 - \frac{1}{2^P})d(N) \text{ time.} \quad (2)$$

Since local b-paths and alignments are computed in parallel, the time for producing them can be estimated as approximately $\frac{1}{2^P} \times b(N)$. We also estimate the parallel communication overheads as a linear function of N , which we represent as $c \times N$. Thus, the speed up formula is:

$$[d(N) + b(N)] / [(1 - \frac{1}{2^P})d(N) + \frac{1}{2^P}b(N) + cNP]. \quad (3)$$

Now, taking into account that $d(N)$ is in general much larger than $b(N)$, the quotient $f(N) = d(N) / b(N)$ is always a fraction $0 < f(N) < 1$, which tends to 0 as N grows to infinity. By dividing equation (3) by $d(N)$ we get,

$$[1 + f(N)] / [(1 - \frac{1}{2^P}) + \frac{1}{2^P}f(N) + cPN/d(N)]. \quad (4)$$

Therefore, as N grows, the theoretical speed up approaches $S = 1 / [1 - \frac{1}{2^P}]$.

C. Pipelining

After returning their local alignments Worker 1 and Worker 2 are idle. Therefore, a new pair of sequences (S_1, S_2) can be received from the master. Subsequent returns from workers liberate the necessary processors for the new sequences to spawn the binary tree of tasks, if required. This is especially suitable for the parallel processing a query sequence; let's say S_1 , against a database of sequences.

IV. CONCLUSIONS

The exploitation of D-D* symmetries, which are derived from the original ideas of Hirschberg, renders a parallel version

of the NWA. The parallel method has a theoretical speed up over the serial NWA and allows for the pipelined processing of a query sequence against a database of sequences. The speed up formula obtained from a simplified performance model seems to indicate that for large problems, the parallel method is more advantageous for small number of processors, this is, a coarse grain parallelization.

REFERENCES

- [1] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis* Cambridge University Press, 2007.
- [2] Needleman S, Wunsch C, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, Vol. 48, Issue 3, pp. 443-453, 1970.
- [3] T.F Smith and M.S. Waterman, "Identification of common molecular subsequences", *Journal of Molecular Biology*, 147(1), 195-197, 1981.
- [4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool", *J. Mol. Biol.*, 215, pp.403-410, 1990.
- [5] O. Gotoh, "An improved algorithm for matching biological sequences", *J. Mol. Biol.*, 162, 705-708, 1982.
- [6] Hsien-Yu Liao, Meng-Lai Yin, Yi Cheng, "A parallel implementation of the Smith-Waterman algorithm for massive sequences searching", *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, San Francisco, CA, USA; September 1-5, 2004.
- [7] F. Zhang, X. Z. Qiao, Z. Y. Liu, "A parallel Smith-Waterman algorithm based on divide and conquer", *Proceedings Fifth International Conference on Algorithms and Architectures for Parallel Processing (ICA3PPi02) 2002*.
- [8] D. Hirschberg, "A linear space algorithm for computing maximal common subsequences", in *Communications of ACM*, Vol. 18, No. 6, pp. 341-343, 1975.

Genetic Matching: An Efficient Algorithm to Adjust Covariate Imbalance for Data Analysis and Modeling

Kao-Tai Tsai and Karl E. Peace

Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, Georgia

Abstract - *In causal-effect relationship research, similarity of groups being compared in terms of covariates or patient/disease characteristics is critical to ensure fairness of the comparison and unbiasedness of the findings. When dissimilarity is suspected, one can either adjust for imbalance or match the groups according to certain important covariates or characteristics. Regression analysis is commonly used to adjust the imbalance and matching techniques are usually used to match subjects between groups. Diamond and Sekhon [2] proposed a genetic matching algorithm to maximize the covariate balance. We describe the theory and conduct a simulation study to compare the relative performance of propensity score matching, Mahalanobis matching, and Genetic matching. Generally, Genetic matching achieves better covariate balance and produces more stable and unbiased treatment effect estimates. We also apply Genetic matching to a clinical study to investigate the treatment effects on rheumatoid arthritis.*

Keywords: propensity score, Mahalanobis matching, Genetic matching, Robbins-Munro stochastic approximation, randomized controlled clinical trials.

1 Introduction

In causal-effect relationship research, similarity of groups being compared in terms of covariates or patient/disease characteristics is critical to ensure fairness of the comparison and unbiasedness of the findings. When dissimilarity is suspected, one can either adjust for imbalance or match according to certain important covariates or characteristics. Regression analysis is commonly used to adjust for imbalance and matching techniques are usually used to match subjects between comparison groups. Therefore, matching has become an important method of causal-effect relationship inference in many fields including biomedicine, economics, social science, and statistics, to name a few.

Several matching procedures have been proposed in the literature by researchers since the early 1970s. Important differences between these proposed methods are the

efficiency of the algorithms utilized and the effectiveness of the methods to reduce imbalance prior to subsequent inferences.

Propensity score matching based on logistic regression and multivariate matching based on Mahalanobis distance are among the more commonly used methods for this purpose. Several variations and combinations of these methods are also used frequently by practitioners.

When covariates have spherical or ellipsoidal distributions, these methods generally perform quite well; however, these methods can perform poorly when the distributions deviate substantially from this family of distributions. Therefore, it is highly desirable to have alternatives that can perform well even when the distributions of the covariates deviate substantially from this family of distributions.

Diamond and Sekhon [2] and Sekhon [15] proposed a genetic matching algorithm that imposes additional properties and generalizations to propensity score and Mahalanobis matching methods and maximizes the balance of observed covariates between the subject groups being compared. The method is nonparametric and does not depend on knowing or estimating the propensity score; however, when a propensity score is incorporated, the method can sometimes be improved by taking advantage of the information embedded in the propensity scores.

Genetic matching has been successfully utilized in social sciences to investigate causal-effect relationships (Diamond and Sekhon [3], Hopkins [5]); however, it has rarely been used in biomedical research to investigate between treatment group differences with covariate imbalance among subjects in the groups.

As stated by Peto, et al. [7], "There is simply no serious scientific alternative to the generation of large-scale randomized evidence. If trials can be vastly simplified, . . . , and thereby made vastly larger, then they have a central role to play in the development of rational criteria for the planning of health care throughout the world." Recruitment of a large number of eligible patients from

a general population is both a major strength and weakness of large pragmatic trials.

Deliberately broadening the entry criteria means that the overall result can be difficult to apply to particular groups. However, in modern medical practice, physicians are often interested in individualized medicine and how best to use results of randomized clinical trials to maximize the wellbeing of each patient. Therefore, proper analyses of targeted subgroups of patients to investigate treatment efficacy has become increasingly necessary if heterogeneity of treatment effects is likely to occur.

Theoretically, the covariates of subjects should be well balanced in randomized controlled trials. However, in actual practice with small to moderate sample sizes, it is not uncommon to find subgroups of patients under study with covariate imbalance. This issue is a particular concern in many observational studies with long-term follow-up due to subject attrition. Therefore, it is critical to ensure similarity between subjects on important covariates in order to make the efficacy comparison of treatments meaningful and unbiased.

In section 2 of this article, we describe the theory of the propensity score, Mahalanobis distance matching and Genetic matching methods. In section 3, we describe a simulation study we conducted to compare the relative performance of these matching methods. In section 4, we apply Genetic matching to a dataset from a clinical study to investigate the relative effectiveness of two treatments for rheumatoid. Discussion and conclusions are presented in section 5.

2 Theory of Propensity Score, Mahalanobis Distance, and Genetic Matching

2.1 Propensity Score Matching

The concept of propensity scores is thoroughly discussed by Rosenbaum and Rubin [10] as well as by other authors. In the following, we describe a few key points for analytical purposes. Let Y_{i1} denote the response of the active treatment of subject i , ($1 \leq i \leq N$), and Y_{i0} denote the response of the control treatment of subject i . Let X_i denote the vector of covariates associated with subject i and $T_i = 1(0)$ if subject i receives active (control) treatment. The observed outcome for subject i is then $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.

If subjects were well randomized between treatment and

control groups, then

$$E(Y_{ij}|T_i = 1) = E(Y_{ij}|T_i = 0), \quad j = 0, 1, \quad (1)$$

even though $E(Y_{i0}|T_i = 1)$ of the treated group and $E(Y_{i1}|T_i = 0)$ in the control group cannot be estimated from the data since each subject can receive only either control or active treatment, but not both.

Under the well-randomized situation, the average treatment effect can be estimated using the observed data by

$$\tau = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) = a_1 \tau_1 + b_1 \tau_0, \quad (2)$$

where $a_1 > 0$, $b_1 > 0$, $a_1 + b_1 = 1$, and

$$\begin{aligned} \tau_1 &= [E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1)], \\ \tau_0 &= [E(Y_{i1}|T_i = 0) - E(Y_{i0}|T_i = 0)] \end{aligned} \quad (3)$$

are the (unobserved) treatment effects from the treated and control groups, respectively.

When imbalance in covariates is suspected between the patient groups under study, proper matching of covariates is needed prior to subsequent inference in order to obtain a fair estimate of treatment effect or difference. Given covariate X_i , and following the results of Rubin [12, 14], one can show that

$$E(Y_{ij}|X_i, T_i = 1) = E(Y_{ij}|X_i, T_i = 0). \quad (4)$$

Therefore, the treatment effect of the treated group can be estimated by

$$\tau_1 = E_{\{X_i|T_i=1\}} \{E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0)\} \quad (5)$$

where the expectation is taken over $\{X_i|T_i = 1\}$.

Define the propensity score as

$$e(X_i) = P(T_i = 1|X_i) = E(T_i|X_i), \quad (6)$$

namely, the probability of patient i being assigned to active treatment given the covariate. Assume, given the subjects covariates, treatment assignments are not deterministic and are independent among study subjects, Rosenbaum and Rubin [9] had shown that

$$\begin{aligned} \tau_1 &= E_{\{e(X_i)|T_i=1\}} \{E(Y_i|e(X_i), T_i = 1) \\ &\quad - E(Y_i|e(X_i), T_i = 0)|T_i = 1\}, \end{aligned} \quad (7)$$

where the expectation is taken over $\{e(X_i)|T_i = 1\}$, and τ_0 can be estimated similarly. Therefore, the average treatment effect can be estimated by combining the results of τ_1 and τ_0 . More details about the propensity score can be found in Rosenbaum [9] in addition to the papers mentioned herein.

Let $X_i = (x_{i1}, x_{i2}, \dots, x_{ik})'$ and $m \leq k$ be the vector of covariates. A common method to estimate $e(X_i)$ is via the logit function, i.e.,

$$\text{logit}(e(X_i)) = \beta_0 + h_1(\eta_{1i}) + h_2(\eta_{2i}), \quad (8)$$

where h_1 and h_2 are known functions and $\eta_{1i} = \sum_{r=1}^m f_r(x_{ir})$, $\eta_{2i} = \sum_{r,q=1}^m f_r(x_{ir})f_q(x_{iq})$ represent the main effects and interactions, respectively. The parameters in Eq(1) can be estimated using MLE. Goodness-of-fit can be checked graphically via Landwehr, *et al* [6] or Tsai [16].

According to Rosenbaum & Rubin [10], it is advantageous to sub-classify or match not only on $e(x)$ but for other functions of x as well. In particular, such a refined procedure may be used to obtain estimates of the average treatment effect in a subpopulation defined by the components of X ; for example, gender or different disease classifications.

2.2 Mahalanobis Matching and Genetic Matching

Given two covariates, X_i and X_j , the Mahalanobis and Genetic Matching are defined as following in terms of the distance between the covariates

$$md(X_i, X_j) = \{(X_i - X_j)'S^{-1}(X_i - X_j)\}^{1/2}, \quad (9)$$

and

$$gmd(X_i, X_j) = \{(X_i - X_j)'S^{-1/2}WS^{-1/2}(X_i - X_j)\}^{1/2}, \quad (10)$$

respectively, where $S^{1/2}$ is the Cholesky decomposition of the covariance matrix of X , and W is a diagonal positive definite weight matrix. The elements of W are chosen to simultaneously minimize the distributional difference and location difference of covariates between the treatment and control groups based on the Kolmogorov-Smirnov test and t -test, respectively (Sekhon [15]).

The conventional test of covariate balance based on the t -test focuses only on location and can miss distributional differences between covariates. On the other hand, the Kolmogorov-Smirnov test compares distributional differences and can miss differences in locations. By combining these two tests, the covariates can be better matched in both location and other properties of the distributions.

3 Comparison of Matching Methods - a Simulation Study

3.1 Design of a Simulation Study

To investigate the performance of various matching methods, a simulation of 500 iterations was conducted under various scenarios. Specifically, the simulation plan was designed as follows:

1. Sample size: assume equal sample size ($N = 20, 30, 50, 100$) between treatment and control groups.
2. Assume 3 covariates (x_{i1}, x_{i2} , and x_{i3}) will be matched between treatment and control groups. The covariates were assumed to have somewhat different distributions between treatment and control. Four different distributions were assumed and are shown in the following table. They consist of standard normal distributions with possibly different means and variances, or contaminated normal distributions with either symmetric or asymmetric contaminations from either tail. The list of distributions is shown in the table below.

X_i	Group	$F:\#1$	$F:\#3$
x_{i1}	treated	$N(0, 1)$	$0.9N(1, 1) + 0.1N(1, 3)$
	control	$N(0, 1)$	$0.9N(0, 1) + 0.1N(0, 3)$
x_{i1}	treated	$N(0, 1)$	$0.9N(0, 2) + 0.1N(0, 3)$
	control	$N(0, 1)$	$0.9N(1, 2) + 0.1N(1, 3)$
x_{i1}	treated	$N(0, 1)$	$0.9N(1, 3) + 0.1N(1, 4)$
	control	$N(0, 1)$	$0.9N(0, 3) + 0.1N(0, 4)$
X_i	Group	$F:\#2$	$F:\#4$
x_{i1}	treated	$N(1, 1)$	$.9N(1, 1) + .1 N(1, 3) $
	control	$N(0, 1)$	$.9N(0, 1) + .1 N(0, 3) (-1)$
x_{i1}	treated	$N(0, 2)$	$.9N(0, 2) + .1N(0, 3)$
	control	$N(1, 2)$	$.9N(1, 2) + .1N(1, 3)$
x_{i1}	treated	$N(1, 3)$	$.9N(1, 3) + .1 N(1, 4) (-1)$
	control	$N(0, 3)$	$.9N(0, 3) + .1 N(0, 4) $

3. The response variable (Y) was assumed to follow two different models. The first model is

$$Y_i = \text{treatment effect} + \sum_{j=1}^3 x_{ij} + \text{error}, \quad (11)$$

and the second model is

$$Y_i = \text{treatment effect} + \sum_{j=1}^3 x_{ij} + \sum_{j \neq k=1}^3 x_{ij}x_{ik} + \text{error}. \quad (12)$$

The treatment effect difference between treatment and control is assumed to be a constant, e.g., 1. The purpose of assuming two different models is to compare these methods when the model is incorrectly specified.

4. The statistical methods to be compared are:
 - (a) Empirical mean difference,
 - (b) Least squares (LS) fit (assuming the first model is correct),
 - (c) LS fit (assuming the second model is correct),
 - (d) Matching on the propensity score,

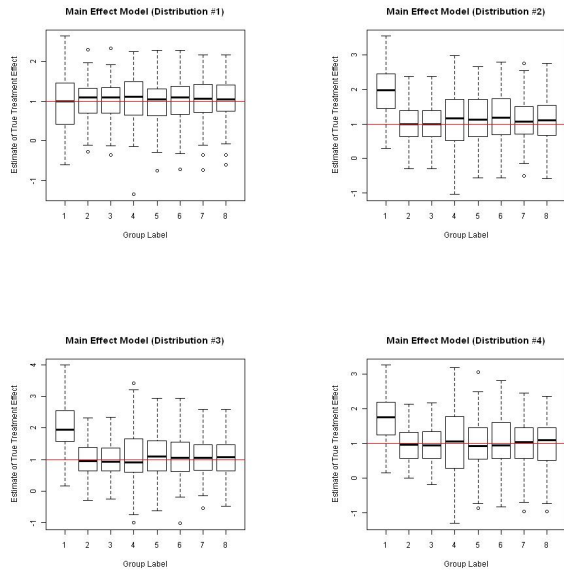


Figure 1: Estimation of treatment effect ($=1$): Main effect only. (Labels 1-8 = a-h of item 4 in Sec. 3.1)

- (e) Matching on x_{i1}, x_{i2} , and x_{i3} with all available data,
 - (f) Matching on x_{i1}, x_{i2}, x_{i3} , and the propensity score with all available data,
 - (g) Matching on x_{i1}, x_{i2} , and x_{i3} but excluding data in either tail outside of 2 times MAD (MAD is defined as $1.483 \text{ med}_i\{|x_{iu} - \text{med}_j(x_{ju})|\}$) from the median for each covariate (to mimic Tukey's robust trimmed estimate),
 - (h) Matching on x_{i1}, x_{i2}, x_{i3} , and the propensity score but excluding data in either tail outside of 2 times MAD from the median for each covariate.
5. Two criteria for comparisons are examined:
- (a) The estimates of the true treatment effect and the variation of the estimates,
 - (b) Balancing the covariates between treatment and control groups. This will be assessed by examining the minimum p -value of the Kolmogorov-Smirnov test for equality of treatment and control groups distributions for each covariate, respectively, before and after matching. Large p -values are consistent with greater comparability of the treatment and control groups in terms of the covariates, and hence reflect better covariate balance among treatment and control groups.

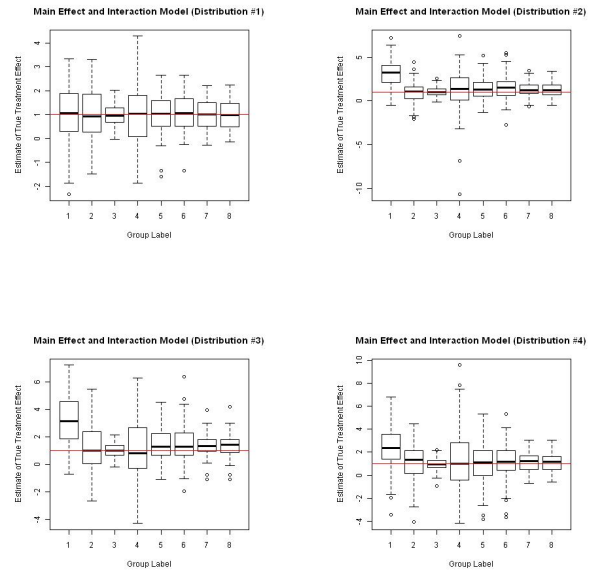


Figure 2: Estimation of treatment effect ($=1$): With interactions. (Labels 1-8 = a-h of item 4 in Sec. 3.1)

3.2 Summary Results of the Simulation Study

By examining the median, the inter-quartile distance, and the overall range of the box plots of the estimated treatment effect, we make the following conclusions:

1. The simple observed treatment difference can be a very poor estimate when the covariate distributions are different and deviate from standard normal distributions as shown in panels 2 to 4 of Figures 1 and 2.
2. For the main effect model, the LS fit (when the model is correctly specified or even over-fitted with interaction terms) is generally better than other methods in estimating the treatment effect. But the LS fit with main effect only can perform poorly if the true model includes interactions; however, the LS fit with interactions (correct model) outperforms other methods.
3. Matching purely based on propensity scores usually performs worse than Genetic matching either with all available data or with the trimmed dataset in estimating the true treatment effect. The trimmed estimate using Genetic matching to match both covariates and propensity scores performs almost uniformly better than any other method regardless of model specification, except for the LS fit when the model is correctly specified as discussed in (b) above.

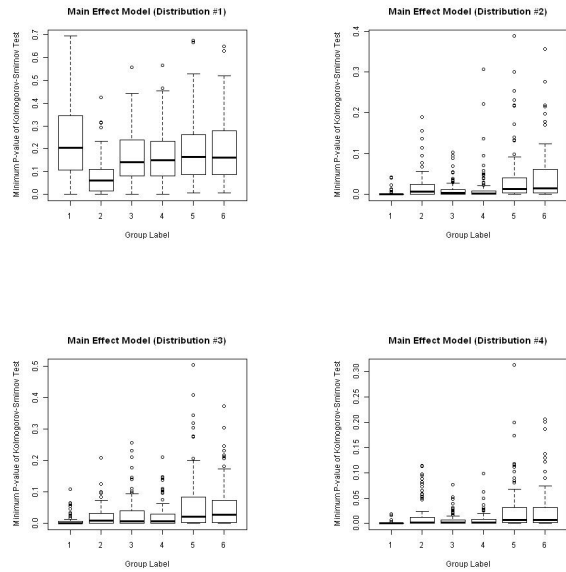


Figure 3: Minimum p-value of K-S test of equality: Main effect. (Labels 2-6 = d-h of item 4 in Sec. 3.1)

4. When the covariates of treatment and control groups have identical normal distributions, the LS method outperforms all other methods since there is no need for matching. Any effort to match is redundant. The propensity score matching seems to make the covariate matching worse more often than not. However, the Genetic matching seems to perform reasonably well, especially when the outliers were trimmed away (Panel 1 of Figures 3 and 4).
5. However, when the covariate distributions are different between treatment and control groups and deviate from the standard normal, the effect of matching from all methods becomes very visible. This can be seen in Panels 2-4 of Figures 3 and 4. Genetic matching with trimmed outliers tends to outperform all other methods either matched only on all covariates or with propensity score included. This is true for all distributions tested here.

As discussed above, when the model is correctly specified, the simple LS method outperforms other methods as expected. However, generally when analyzing data, one rarely knows the correct model or the distribution from which the data was generated. Therefore, the performance of LS method can be expected to diminish in the analysis of real data. On the other hand, the performance of Genetic Matching seems to be almost always comparable to the LS method when the model is correctly specified, and performs much better when the model is mis-specified as shown in Panels 1, 2, and

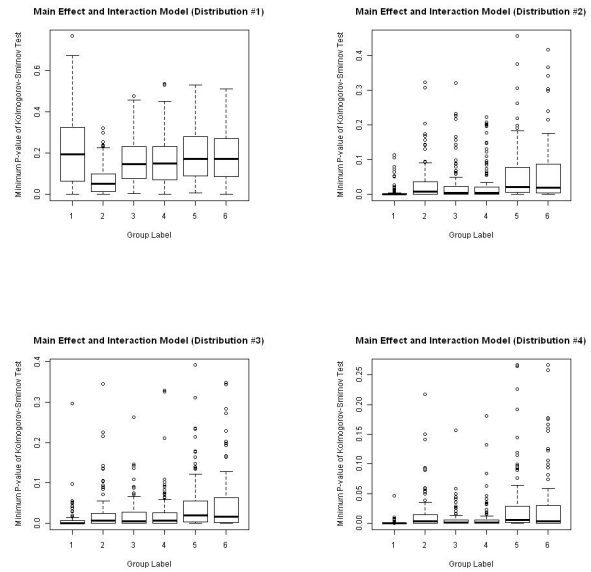


Figure 4: Minimum p-value of K-S test of equality: Interactions. (Labels 2-6 = d-h of item 4 in Sec. 3.1)

4 of Figure 2. Therefore, the Genetic Matching seems to serve as a “model mis-specification proof” tool for general data analysis.

It is interesting to note that Diamond, et al. [2] concluded that Genetic Matching is preferred over other matching methods because it is more efficient (smaller MSE) and is less biased.

4 Example

A phase III, multi-national randomized, double blind, placebo controlled clinical trial was conducted to compare the treatment effect of drug A and drug B to placebo in controlling disease activity in subjects with rheumatoid arthritis having an inadequate clinical response to methotrexate. The study was not originally designed to compare drug A and drug B directly. However, a post hoc analysis to compare these two drugs in a subgroup of countries of the original study is of clinical interest and also to meet the regulatory request. A total of 156 and 165 patients were randomized to drugs A and B in these countries, respectively. The primary endpoint of the study was the disease activity score based on 28 joints (DAS28).

Comparisons of several baseline covariates using the t -test did not show particular imbalance between the two treatment groups. However, a more in-depth investigation of the baseline distributions by quantile-quantile plots showed some deviations between the two popu-

lations. The objective in this analysis is to properly estimate the treatment difference under the situation of baseline imbalance.

The first step in this analysis is to match the patients from drugs A and B. Both the propensity score and the Genetic matching methods were used so that we can compare the relative performance of these two matching methods.

Several covariates were examined to compare the performance of propensity score and Genetic matching. The baseline pain scores between the treatment groups are compared and shown in Figure 5. The original Q-Q plot of pain scores between drug A and drug B is shown in Panel 1. The Q-Q plots of this covariate using propensity score matching and Genetic matching are shown in Panels 2 and 3, respectively. One can clearly see substantial improvement in covariate balance of Genetic matching over the propensity score matching.

Empirical permutation distributions of the treatment effect before and after Genetic matching were generated to determine the level of significance of the observed treatment effect among the randomly permuted samples. The observed treatment difference prior to matching is about -0.19. However, the magnitude of the treatment difference was reduced to -0.048 after matching. The treatment effect estimated after matching indicates the treatment difference is not as big as the original estimate. In other words, without this matching step, the treatment difference may have potentially been overestimated and the medical practice may be misguided. Even though the permutation test did not show a significant treatment difference in either pre or post matching; however, the treatment effect distributions from permutations seem to have some subtle difference and the test prior to matching showed a higher significance level than post matching. The 95% confidence interval of the treatment effect difference was also estimated using the stochastic approximation proposed by Robbins and Munro [8] and implemented by Garthwaite [4]. A total of 5000 randomized samples were generated and analyzed. The estimates fluctuate substantially in the beginning of the approximation process. The process began to stabilize after about 2500 randomizations. Figure 6 shows the stochastic approximation for the upper and lower limits of the confidence interval. The resulting 95% confidence interval is (-0.110, 0.4858).

5 Discussion

Statistical modeling and data analysis are important steps in advancing innovative scientific research in the fields of medicine, economics, social sciences, etc. To

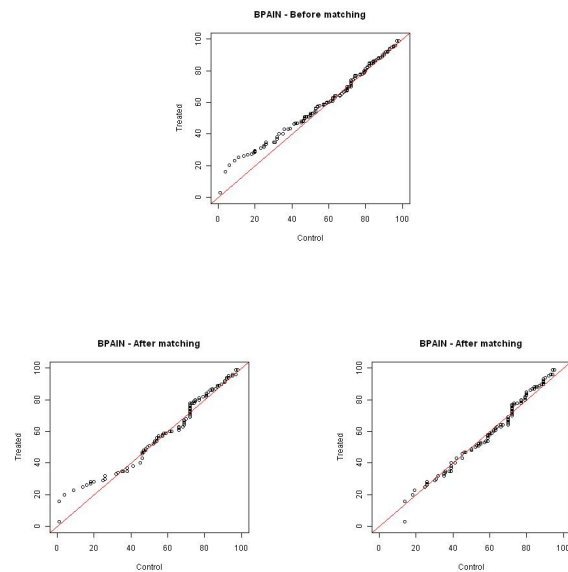


Figure 5: Comparison of covariate adjustment before and after propensity score and genetic matching, respectively.

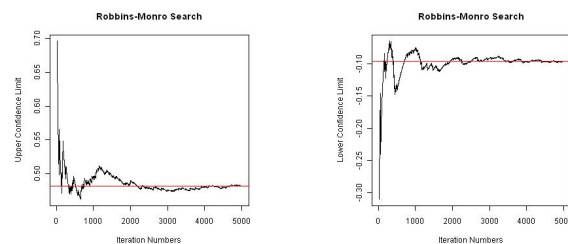


Figure 6: Stochastic approximation of the 95% confidence interval of treatment effect difference (based on 5000 simulated randomization)

translate data into useful unbiased information is a critical endeavor for scientists and researchers as well. When data do not come from well-designed experiments, statistical modeling and data analysis to extract unbiased information can become much more challenging.

In this paper, we described various matching techniques to make the subjects under consideration more comparable before statistical inference; we also conducted a simulation study to further investigate the performance of these methods under different scenarios in their relative ability to better balance the covariates between the subjects groups, and in obtaining the unbiased estimate of treatment effect. The methods we compared ranged from the usual linear regression, conventional matching techniques with all available data to more robust alter-

natives, which flexibly weights the outliers. Generally, Genetic matching is preferred to other methods under various covariate distributions in balancing the covariates and obtaining the true treatment effect.

Given its longer history, the propensity score matching has been the most well known and most commonly used method in casual-effect relationship research; however, the selection of variables to be incorporated into the logistic regression model to derive the propensity score is not a trivial matter.

Several authors have proposed various approaches to incorporate covariates to estimate the propensity score (e.g., Rubin & Thomas [11], Rubin [14], Brookhart et. al. [1]). The general findings are to incorporate covariates which are thought to be related to outcomes and are confounded with both treatment assignment and outcomes. The model which incorporates as many covariates as possible or the model which includes obvious covariates such as age, gender, and race do not seem to perform as well as one would expect. On the other hand, the Genetic matching method has the additional flexibility to allow the covariates to be assigned unequal weight and also takes into account the covariance of the variables incorporated into the distance calculation which can eliminate some modeling difficulties caused by co-linearity between covariates.

Estimation and comparison of treatment effects should only be conducted after careful examination of balance between the groups being compared. It is important to note that the research findings should be regarded as exploratory and be interpreted with care within the context of biological or scientific plausibility and relevance.

References

- [1] Brookhart, M.A. et. al. Variable selection for propensity score models. *American Journal of Epidemiology* 2006; 163: 1149-1156.
- [2] Diamond A, Sekhon, JS. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Working paper (2005).
- [3] Diamond, A. and Sekhon, J. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. University of California, Berkeley, 2010.
- [4] Gartwaite, P.H. Confidence intervals from randomization tests. *Biometrics* 1996; 52: 1387-1393.
- [5] Hopkins, D. Politicized Places: Explaining Where and When Immigrants Provoke Local Opposition. *American Political Science Review* 2010, 104 (1): 4060.
- [6] Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association* 1984; 79: 61-71.
- [7] Peto, R., Collins, R., and Gray, R. Large-scale randomized evidence: large, simple trials and overviews of trials. *Journal of Clinical Epidemiology* 1995; 48: 23-40.
- [8] Robbins, H. and Munro, S. A stochastic approximation method. *Annals of Mathematical Statistics* 1951; 22 : 400-407.
- [9] Rosenbaum, P.R. *Observational Studies*. New York: Springer-Verlag 1995.
- [10] Rosenbaum, P.R. and Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 79 : 516-524.
- [11] Rubin, D.B. and Thomas, N. Matching using estimated propensity score: relating theory to practice. *Biometrics* 1996; 52 : 249-264.
- [12] Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; 66 : 688-701.
- [13] Rubin, D.B. Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* 1977; 2 : 1-26.
- [14] Rubin, D.B. Estimating causal effects from large data sets using the propensity score. *Annals of Internal Medicine* 1997; 127 : 757-763.
- [15] Sekhon, J.S. Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference. Working Paper. <http://sekhon.berkeley.edu/papers/Sekhon-BalanceMetrics.pdf> 2006.
- [16] Tsai, K.T. Assessing Regression Modeling with Ordinal Responses. Presentation at the Joint Statistical Meetings of the American Statistical Association 2008.

Contact author: Kao-Tai Tsai
Email address: tsai0123@yahoo.com

A CAM(Content Addressable Memory)-based architecture for molecular sequence matching

P.K. Lala¹ and J.P. Parkerson²

¹Department Electrical Engineering, Texas A&M University, Texarkana, Texas, USA

²Department Computer Science & Computer Engineering, University of Arkansas, Fayetteville, Arkansas, USA

Abstract - *The development of elegant matching procedures has significantly improved the search time of a query sequence in a database of molecular sequences. However the architectural limitations of conventional computers allow only one sequence access/retrieval from the database at a time. This paper presents a digital hardware-based solution for fast comparison of molecular sequences.*

Keywords: CAM, sequence alignment, codon, dynamic programming, barrel shifter

1 Introduction

The derivation of functional information from genomic sequences has many applications in molecular biology. Since similar sequences behave in a similar manner, the characteristics of a new sequence can be predicted by comparing it with known sequences. Sequence comparison provides an indication of which parts of comparing sequences are similar and which parts are different. It is employed to identify sequences similar to a given sequence from a database [1-3]. Two sequences are said to be *homologous* if they evolved from the same ancestor sequence, and share many common features.

Sequence similarity in many ways is synonymous with the concept of sequence alignment, which identifies similarities and differences between sequences. An alignment is a correspondence between the sequences, in which each symbol in a sequence is assigned to at most one of the symbols in the other sequence while maintaining the order of the symbols in the sequences. The main objective of the alignment is to have matching symbols at maximum number of positions.

Sequences can be compared either by *global* or *local* alignment. Global alignment spans the whole length of comparing sequences, and is appropriate if the sequences are

likely to share substantial similarity. The alignment attempts to match them to each other from end to end, even though parts of the sequences may not match. Local alignment is used to identify common sub-regions of similarity between long sequences. There is no attempt to force entire sequences into an alignment, just those parts that appear to have good similarity. Thus local alignment is therefore particularly useful for comparison of long DNA sequences where only small subsequences may be related.

The number of possible alignments for two sequences of length m and n is extremely large. Therefore the obvious solution of enumerating all alignments and then choosing the one with the smallest or the highest score (depending on the scoring scheme used) is computationally impractical. An efficient alignment process needs to employ a completely automatic method e.g. *dynamic programming algorithms*, which is usually used for solving optimization problems. Needleman and Wunsch [4] were the first to propose such a method. Their motivation for developing the method was to maximize similarities between amino acid sequences. It allows global comparison of an entire query sequence with all sequences in a database. One drawback of this global alignment algorithm is that highly similar shorter subsequences with meaningful similarities may be ignored because of the overall objective of matching largest number of residues of one sequence with another sequence.

Smith and Waterman [5] proposed an algorithm, perhaps the most widely used local similarity algorithm for biological sequence comparison, based on the concept of dynamic programming. It identifies pairs of subsequences of all possible lengths in the query and the database sequence that have maximum degree of similarity. Two other dynamic programming algorithms BLAST [1] and FASTA[2] have been developed to identify possible homologues for a query sequence in a database of all other known sequences.

2 Hardware-based matching

In recent years there has been some academic and industrial efforts on the use of FPGAs (Field Programmable Gate Arrays) for enhancing the speed of sequence matching [6-8]. However all the techniques developed so far for accelerating this task sequentially compare a query sequence with sequences stored in a conventional memory system. Although this strategy results in improvement in the matching speed compared to traditional software-based algorithms, the comparison of the query sequence with the stored sequences still needs to be done one sequence at a time because of the sequential nature of information retrieval from the memory system.

The only way significant improvement in the matching speed can be achieved if a query sequence can be compared with all the stored sequences in parallel, and if all the matched sequences can be accessed simultaneously. This paper presents a digital hardware-based sequence matching procedure based on this principle. It is assumed that a query sequence is a subsequence of one or more sequences stored in the computer memory system. The bases and the *gap* in a sequence are represented by binary patterns as shown below:

A = 000, C = 010, G = 100, T = 110, – (*dash*) = $xx1$

where x is a *don't care*. Thus, a 1 in the least significant bit of a 3-bit binary representation indicates a *dash*.

Fig.1 shows the block diagram of the proposed matching system. It consists of a dedicated Content Addressable Memory (CAM) block and a bi-directional barrel shift register. A CAM unlike a traditional RAM (Random Access Memory) is addressed by the desired content, and an address that stores the content is obtained as the output of the CAM [9]. It is assumed that a CAM-based memory system will store a large number of sequences, and the content of a stored sequence has to have a few common codons i.e. all three bases identical, for it to be accessible. However, in order to avoid a large number of “hits” in the CAM, the number of codons considered for matching has to be properly selected to keep the number of accessed sequences to a realistic number. For the sake of simplifying the explanation of our approach we consider matching of only one codon for accessing relevant sequences in a CAM.

While comparing a query sequence with the stored sequences in a CAM, a better match may be obtained by shifting the query sequence left (or right) by one or more

bases at a time, and then comparing it with the stored sequences. Traditionally the alignment process starts after some partially-matched sequences have been retrieved from the memory block. Then by proper placement of *dashes* in the query sequence, matching of bases in the comparing pair is maximized.

In the proposed hardware-based sequence matching approach, one of the goals is to retrieve sequences with a high degree of similarity whenever possible. The possibility of better matching is explored by shifting the query sequence to the left (or right) before accessing the stored locations. The shifting of a single base in the query sequence requires simultaneous shifting of three bits. The barrel shifter in Fig.1 is used for shifting pre-selected multiple data bits. It is used to store the query sequence, and has a length of $3b$ where b is the number of bits in the sequence. It has a shift left and shift right capability, and can be used to shift three bits i.e. one base at a time.

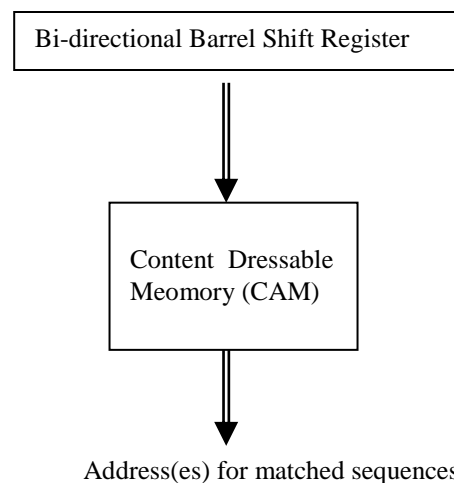


Fig. 1 Hardware based system for sequence match

Fig.2 shows the structure of an n -bit barrel shifter; the control logic determines the direction of shifting (i.e. left or right). It is assumed that that the shift-in data is 001 i.e. a *dash*. Although the proposed system is designed for local sequence alignment, it could also be used for global sequence alignment.

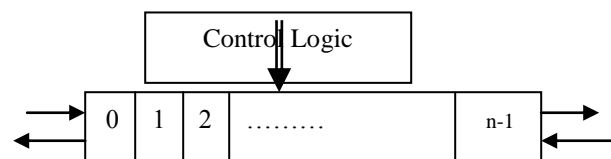


Fig. 2 Bi-directional Barrel Shift Register

A dedicated CAM-based architecture has been designed for simultaneous comparison of a pre-defined number of codons in a query sequence with the same number of codons in stored sequences in the CAM. Fig.3 shows the proposed architecture. If a majority (pre-selected) of the number of codons match with the corresponding codons in the stored sequences, then the addresses of these stored sequences are retrieved; they identify the locations of sequences that partially match the query sequence. The address corresponding to a matched location is activated via the output of an *k-out-of-n* detector. A *k-out-of-n* detector produces an output of 1 if at least *k* out of *n* inputs is at 1.

For example, if the first five codons of a query sequence are compared with a stored sequence and a match is assumed if any three of the five codons are similar, then a 3-out-of-5 detector is used to identify the address of the stored location. To illustrate let us assume that the first five codons in the following six codon query sequence

out-of-5 detector is used to identify the address of the stored location. To illustrate let us assume that the first five codons in the following six codon query sequence

ACG AT- CGT –GA TCG ATG

are compared with a stored sequence

ACG CAG CGT TTC TCG AC- C-T ATC

Since three codons in the comparing sequences match, a 3-out-of-5 detector circuit will produce an output of 1. On the other hand if four codons have to be similar for a match, then a 4-out-of-5 detector has to be used; this detector will produce an output 0 in this case. Thus, a programmable detector that allows pre-selection of the *k* value in the *k-out-of-n* detectors will enable comparison of pre-selected parts of the query sequence to stored sequences.

As shown in Fig. 3 the CAM architecture has *m* rows (addresses) and *n* columns (codons). Each row consists of *n* comparators, *n* two-input AND gates, and a *k-out-of-n* programmable detector. One of the inputs to a two-input AND gate is the output of a comparator, the other input is programmable. The programmable input to the AND gate is the output of a comparator, the other input is programmable. The programmable input to the AND gate is set at 1 if a codon in the query sequence is being matched with the codon at the identical position in the stored sequence.

A matching circuit is used for comparing a codon in the query sequence with that in a stored sequence/. Fig. 4 shows the circuit for matching of two codons where *lmn* and *pqr* are the codons in the query sequence and a stored sequence respectively. Note that this circuit includes a 2-out-of-3 detector. Thus as long as two bases in the comparing codons match, the codons are assumed to be matched. The output the 2-out-of-3 detector is connected to the input of the programmable detector associated with the stored sequence address decoder via a two-input AND gate.

To illustrate the application of the system of Fig.1 for sequence comparison let us assume the following query sequence in the barrel shifter:

ACT –GAT-CGAA

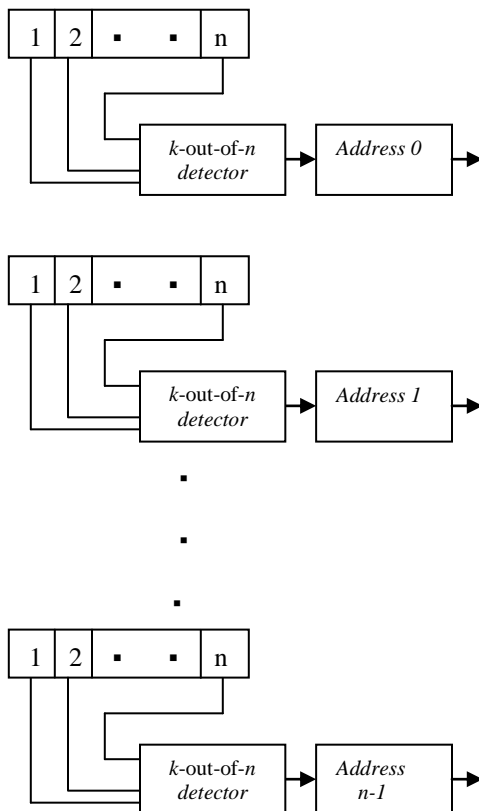


Fig. 3. Proposed $m \times n$ CAM Architecture

For example, if the first five codons of a query sequence are compared with a stored sequence and a match is assumed if any three of the five codons are similar, then a 3-

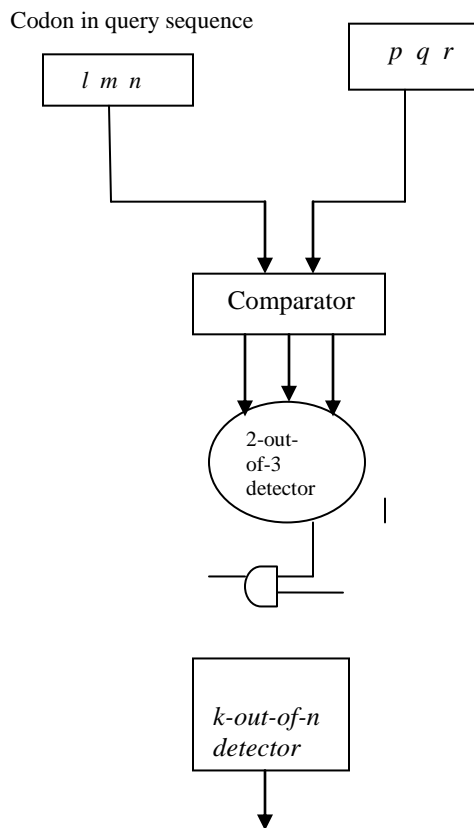


Fig . 4 Matching circuit

:Suppose the contents of the CAM (assuming it has four locations) are as follows

Content	Address
A-CT-GTCGCGA	00
ACTG-CT- GGAA	01
- CTATGT - CACT	10
AC-- G -- T--AG	11

If the programmable detectors driving the address identifiers in the CAM are considered to be 3-out-of-4 (n=4 in this example), then only address 01 will be identified as the CAM location that contains the sequence with most matched codons. However if k=2 (i.e. 2-out-of-4 detector) is used instead, both addresses 01 and 10 will be identified.

One particular advantage of the proposed strategy is that it simplifies the identification of shorter subsequences that may be common among several stored sequences. To illustrate let us determine whether -- TAAG is part of the following stored sequences:

Address	Contents
00	ATCT-A-CAGCG
01	-TAAGC-CAGAG
10	-TGCTAAGCTGA

First -- TAAG is transferred to the barrel shift register as the query sequence. As shown below, the subsequence occupies the first six most significant positions in the barrel shifter, the remaining positions are filled with *dashes*.

-- TAAG -----

This results in matching with the contents of address 01.

A shift to the right by two bases

----- TAAG -----

will result in matching with the contents of address 10.

In certain cases a desired subsequence may appear in a stored sequence in bits and pieces. For example, it is not immediately obvious that the above subsequence is part of the sequence at location 00. However, a number of shift operations of the query sequence as shown below verifies its presence:

--- TAAG -----
 --- TA -AG -----
 --- T-A-AG -----
 --- T-A--AG --

Certain shift operations in this case required shifting of individual bases in the query sequence. Similar situation will arise when a query sequence has certain similarity with a stored sequence. Once the stored sequence has been retrieved a number of shift operations of the individual bases or subsequences in the query sequence may be needed to increase the number of positions with matching symbols. Thus a major objective will be how to design the barrel shifter such that one or more bases can be shifted left or right without necessarily shifting the whole query sequence,

3 Conclusions

Currently available tools for molecular sequence matching are in general software-based. The computation time needed for matching is dependent on search sensitivity. A low sensitivity search requires short computation time,

whereas for a high sensitivity search the computation time is very long. The conventional i.e. Von Neumann architecture based computers can process information only serially, thus limiting their speed of computation. This paper presented a digital hardware-based sequence matching technique that allows simultaneous comparison of a query sequence with all the stored sequences in a database, thereby achieving significant improvement in the matching speed compared to software-based techniques.

4 References

- [1] S. F. Altschul, W.Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic Local alignment Search Tool", *Jour. Molecular Biology*, vol.215(3), pp.403-410, 1990
- [2] D. J. Lipman and W.R.Pearson, "Rapid and sensitive protein similarity searches", *Science* 227 (4693), pp. 1435–41, 1985
- [3]. W.R. Pearson , "Using the FASTA program to search protein and DNA sequence data base", *Methods in Molecular Biology*, 25, pp.365--389, 1994
- [4]. S.D. Needleman and C.D. Wunsch., " A general method applicable to the search for similarities in the amino acid sequences of two proteins", *Jour. Mol. Biol.*, vol.48, pp.443-453, 1970.
- [5]. T. F.Smith and M.S.Waterman, "Identification of common molecular subsequences", *Adv. Appl. Math.*, vol.2, pp.482- 489, 1981.
- [6]. E. Sotiriades, C. Kozanitis and A. Dollas, "FPGA based Architecture for DNA Sequence Comparison and Database Search", *Proc.20th IEEE International Parallel & Distributed Processing Symposium*, pp 186-, 2006
- [7] T. Oliver, B. Schmidt, D. Maskell, D. Nathan, and R. Clemens, " High-speed Multiple Sequence Alignment on a reconfigurable platform", *International Journal of Bioinformatics Research and Applications*, Vol. 2, No.4, pp.394 – 406, 2006
- [8] S. Lloyd and Q.O. Snell, "Hardware Accelerated Sequence Alignment with Traceback", *International Journal of Reconfigurable Computing*, Vol. 2009, Article ID 762362, , 10 pages, 2009
- [9] S.M. Jalaeledine and L.G. Johnson, " Associative IC memories with relational search and nearest match capabilities", *IEEE Jour. Solid. State Circuits*, vol.27, no.6, pp. 892-900,1992.

Acknowledgement

This work was supported in part by the National Science Foundation, USA under Grant 0925080

Effective Algorithms for Altering Human Chromosomes Shapes

Wei Wu¹, Xiaoli Yang¹, and Charles C. Tseng²

¹Department of Electrical and Computer Engineering, Purdue University Calumet, Hammond, IN, USA

²Department of Biological Sciences, Purdue University Calumet, Hammond, IN, USA

Abstract - Learning human cytogenetics is important for biology education and training of clinical cytogenetics technologists. To increase the resources of metaphase spreads with different chromosomes shapes for student practice, we describe effective algorithms for such purposes to enhance our cytogenetics tutorial program.

Keywords: human chromosome, cytogenetics, algorithms

1 Introduction

The function of altering human chromosome shape from curved to straight is available in the commercial karyotyping software [Cytovision, Genetix Corp., San Jose, CA] which is used by clinical genetics laboratories for chromosome analysis from patient samples. The purpose of chromosome straightening in the commercial software is to provide viewers with the appearance of neatly arranged chromosomes on the karyotype sheet. However, few cytogenetics technologists actually use this function because the straightened chromosomes do not provide more information from the G-banded chromosomes. In fact, the straightened chromosomes appear to be unnatural.

For education purposes, however, a versatile program with the capability of altering the chromosome shape can enrich the metaphase resources for student practice. A basic human chromosome modeling program has recently been introduced as a tool for cytogenetics education [1, 2]. In the early model, however, the changeable parts are limited to certain points along the chromosome. In this paper, we describe an improved method for chromosomal skeletonization which serves as a basis for programming the inter-conversion of chromosome shapes from curved to straight and *vice versa*.

The functionality of the chromosome shape alteration is to reconstruct original chromosome images into different shapes. Pixels on the original images are mapped to new positions on destination images using a transformation function. The new algorithms are capable of changing chromosome shapes at any points along the entire length of the chromosome. The new function meets the following two requirements: First, the length and width of the original chromosome image remain unchanged when the altered chromosome is created, and second, the original information

including the grayscale values and G-bands information are kept as close as possible to the newly generated model. Throughout our beta tests, there were no missing or adding G-bands on the altered chromosomes.

2 Chromosome Midline

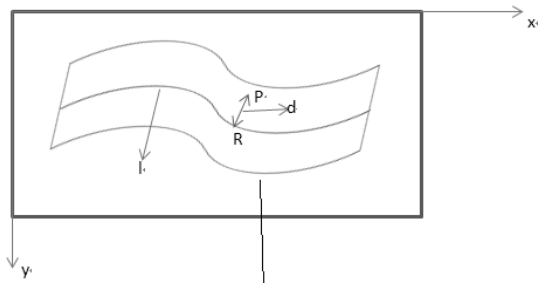
The chromosome midline is to obtain a curve to represent the approximate curvature of the chromosome. This is the pre-processing step for chromosome shape alteration. The human chromosome shape in metaphase is basically linear with the length much larger than the width; therefore, in practice, the width may be neglected so that a midline can be created to represent the chromosome curvature (Fig. 1). The midline curve is in parallel to the two sides with the same distances. The midline curve can be obtained in three steps: boundary detection, thinning, and curve fitting. Boundary detection is to acquire the boundary of chromosome image. Thinning is to convert binary shape obtained from boundary detection to a 1-pixel wide curve. Curve fitting is used to find the "best fit" line or curve for a series of data points.



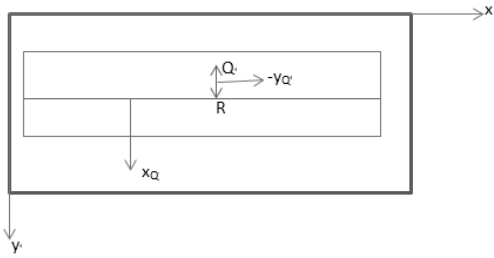
Fig. 1. Midline after curve fitting for human chromosome 2

3 Chromosome Straightening

Most of original chromosome images under the microscope are curved in various degrees. For beginners, it is easier to learn G-banded chromosome characteristics from straight chromosomes. Therefore, metaphase spreads with all straight chromosomes are ideal for beginning students. Fig 2.a is a schematic of an originally curved chromosome: The red line is the midline which can be calculated by chromosome midline introduced in session 2. Fig 2.b is a schematic of the same chromosome after straightening.



Schematic of originally curved chromosome



Schematic of the chromosome after straightening

Fig. 2. Conversion of curved to straight chromosome

After the midline is calculated, the next step is to map the pixels. There are two mapping algorithms: forward mapping and inverse mapping [3]. The forward mapping is to map the source pixel (x, y) to an appropriate destination place (u, v) . However, since the image source pixel is represented as an integer, when applying the translation function on the source pixel, the value of destination pixel may be represented by a decimal number, which must be rounded up to an integer. It may result in one target pixel having multiple source pixels or some missing destination pixels which lead to Mosaic effects. The inverse mapping is that, with the given location of the target pixel (u, v) in the destination image, we can calculate an appropriate location of the source pixel (x, y) in the original image based on the transformation function. Thus, all the destination image pixels are mapped to pixels in the source image. Therefore, the inverse mapping method was chosen for our modeling. Below is the detailed inverse mapping calculation:

In Fig 2.a, pixel $P(x_p, y_p)$ is the source image, and line PR is perpendicular to the midline on pixel $R(x_r, y_r)$. In Fig 2.b, $Q(x_q, y_q)$ is the destination pixel of P . In the inverse mapping, the position of Q is given, and the position of P must be calculated for mapping from P to Q . In Fig 2.a, d is the perpendicular distance from P to midline, and l is the length from the beginning of the midline to R ; therefore:
 $-y_Q = d$ and $x_Q = l$

Assume the equation of midline is $y = f(x)$, then

$$l = \int_1^{x_R} \sqrt{1 + f(x)^2} dx \tag{1}$$

$$d = \sqrt{(x_R - x_P)^2 + (y_R - y_P)^2} = \sqrt{(x_R - x_P)^2 + (F(x_R) - y_P)^2} \tag{2}$$

Because $PR \perp$ midline, then

$$\frac{y_R - y_P}{x_R - x_P} \times F'(x_R) = -1 \tag{3}$$

Because R is in the midline, then

$$f(x_R) = y_R \tag{4}$$

Through solving the above functions, $P(x_p, y_p)$ can be calculated. However, the values of x -axis and y -axis must be integers. An interpolation technique is applied here to obtain the pixel value of P . The bicubic interpolation [4] was used to determine a more accurate pixel value of P and make the destination image smoother. P can be written as $P(i+u, j+v)$. The pixel value of $P f(i+u, j+v)$ is calculated from its 16 neighbors according to the bicubic interpolation. Fig. 3 shows the original image of human chromosome 2 on the left side and the straightened model on the right side.

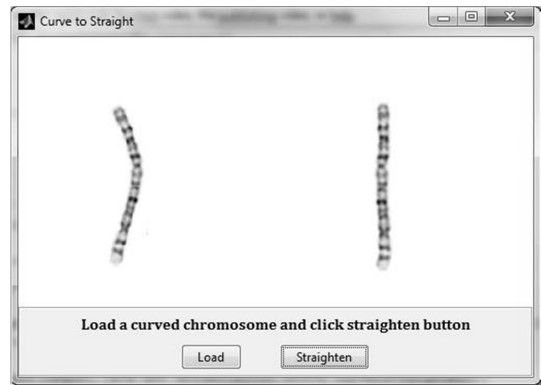


Fig. 3. Result of chromosome 2 straightening

4 Chromosome Curving

The algorithm for converting a straight to a curved chromosome is similar to that described in the precious section 3, but the midline of the destination chromosome is now a curved line. The shape of the curve can be set manually. Theoretically this algorithm is capable of changing a straight chromosome to any shape. A painting program (Fig 4) is provided to allow users to draw a curve as the midline of the altered chromosome. Through the curve fitting algorithm using interpolation method, an equation of the curve drawn by users can be calculated. Based on the equation, the originally straight chromosome can be converted to the designated shape. In Fig. 5, the originally straight human chromosome 2 is on the left, and the curved chromosome on the right is created according the curve drawn in Fig 4.

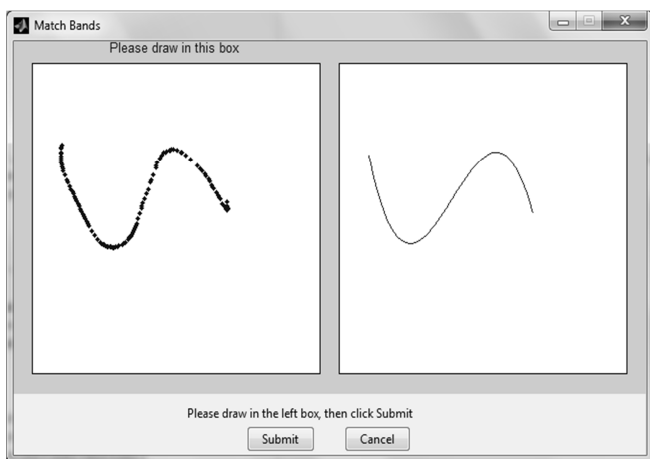


Fig. 4. Painter program accepts input such as a curve from users.

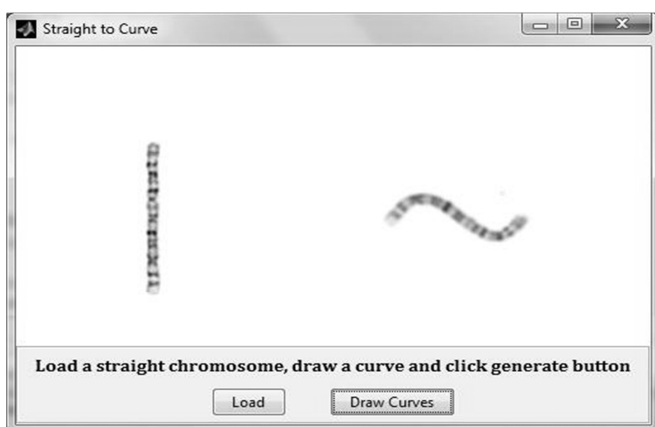


Fig. 5. Conversion of straight to curved chromosome 2

5 Results and Conclusions

Using the algorithms described above, metaphase spreads with all or mostly straightened chromosomes (Fig. 6) and with all or mostly curved chromosomes (Fig. 7) can be created from the original metaphase spread (Fig. 8). Likewise, karyotypes with all or mostly straightened chromosomes (Fig. 9) and with all or mostly curved chromosomes (Fig. 10) can be generated from the original karyotype (Fig. 11).

This paper describes effective algorithms for alteration shapes of human chromosome images, including straightening and curving at any points along the entire chromosome. With these algorithms, we are able to model and generate a wide range of human chromosome images with different shapes, thus increasing the teaching resources for learning human cytogenetics. The new program can be used in conjunction with our computer based cytogenetics learning programs [1, 2] as a new teaching tool.



Fig. 6. Metaphase spread with all or mostly straightened chromosomes



Fig. 7. Metaphase spread with all or mostly curved chromosomes



Fig. 8. Original metaphase spread

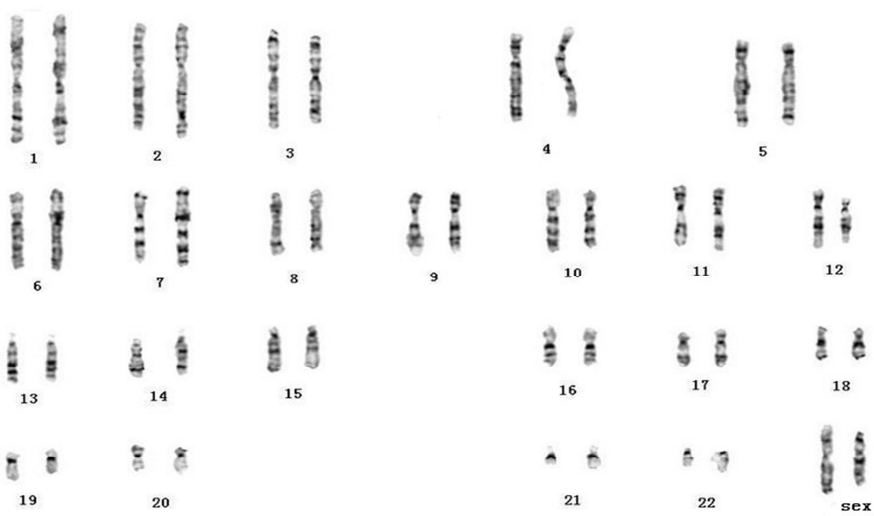


Fig. 9. Karyotype with all or mostly straightened chromosomes

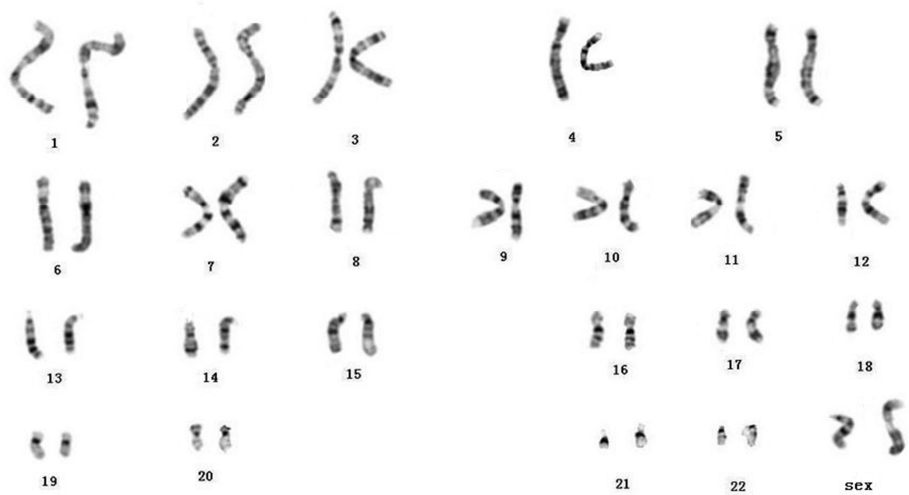


Fig. 10. Karyotype with all or mostly curved chromosomes

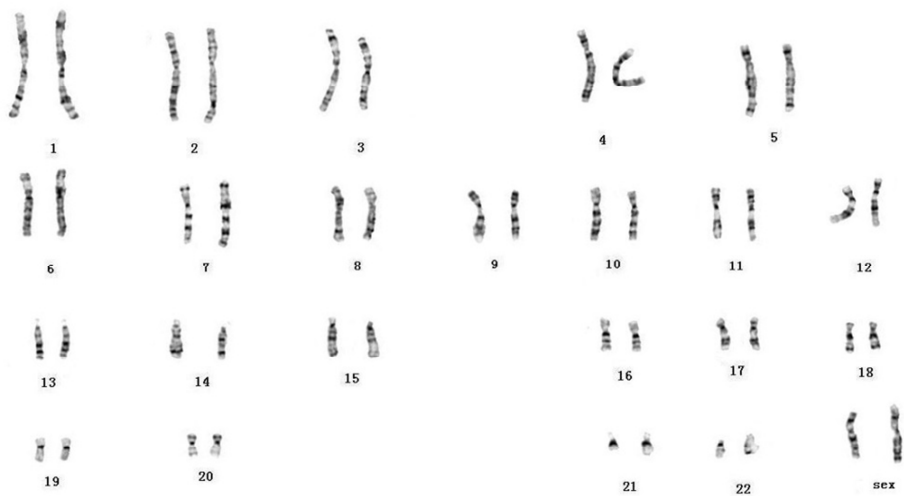


Fig. 11. Original karyotype

6 References

- [1] Yang X., D. Wen, X. Wu, Z. Zhao, J. Lacny and C. Tseng, A Comprehensive Cytogenetics Tutorial Program, Encompassing Changeable G-band Resolutions, *Computer Methods and Programs in Biomedicine*, July 2010/Volume 99/Number 1/ISSN 0169-2607.
- [2] Yang, X., Z. Zhao, D. Wen, X. Wu, Y. Cui, Y. Zhao, X. Cao, J. Lacny and C. Tseng, Virtual Reality Based Human Chromosome Learning Program, *ISAST Transactions on Electronics and Signal Processing (International Society for Advanced Science and Technology)*.2(3): 41-50, 2008. (ISSN 1797-2329).
- [3] P. S. Heckbert. Survey of texture mapping. *IEEE Computer Graphics and Applications*, 6(11):56–67, November 1986.
- [4] Keys, R., Cubic convolution interpolation for digital image processing, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(6), Dec 1981 Page(s):1153 – 1160.

Identification of Minimum Redundancy Tagging SNPs via Gibbs Sampling

Gaolin Zheng

Department of Math and Computer Science, North Carolina Central University, Durham, NC 27707, USA

Abstract—Single nucleotide polymorphisms (SNPs) are genetic changes that can occur within a DNA sequence. Due to the high frequency of SNPs in the human genome, it is desirable to select a small set of SNPs (tagging SNPs) that can be used to represent the majority of SNPs. We propose a Gibbs sampling approach to find a small set of SNPs with minimum redundancy for tagging purposes. Pre-clustering is added in the basic Gibbs sampling procedure to avoid the disturbance caused by local optima. We also propose two general purpose correlation measures that are able to accommodate SNPs with three or more alleles. Our experimental results show that Gibbs sampling process converges faster and finds better optimum if pre-clustering is conducted before the sampling process. While our tagging process is not guided by any prediction algorithm, we are able to obtain comparable results as the SNP prediction guided algorithm SVM/STSA [1] while requiring much less time.

Keywords: minimum redundancy, Chi-squared statistic, mutual information, single nucleotide polymorphism, Gibbs sampling

1 Introduction

Single nucleotide polymorphism (SNPs) are the most frequent variations in the human genome [2], and many SNPs show correlated genotypes because of their shared evolutionary history [3]. Many known polymorphic sites need not be genotyped when testing for genotype-phenotype associations because of this redundancy. There is considerable interest in finding an informative and minimal set of common polymorphisms (tagging SNPs) to detect genetic associations while controlling cost [1, 4-7]. Halldorsson et al. gave an in-depth review of these approaches [8].

Popular tagging SNP selection algorithms are typically based on block-based heuristics such as LD-Select [9], MultiPop-TagSelect [10]. The main drawback of block-based approaches is that the definition of blocks is not always straightforward and there is no consensus on how blocks must be formed [11]. Several researchers have focused on looking for tagging SNPs using block-free methods [1, 8, 11-13]. Most of these methods are based on some greedy deterministic searching procedures that are susceptible to local optimum. Furthermore, most of these

methods are using the r^2 similarity/correlation measure between two SNPs. This measure is not able to handle three or more alleles. SNPs with three or more alleles are usually ignored for processing conveniences. To accommodate more alleles, we propose two correlation measures that are more general purpose for handling nominal data. The first one is mutual information and the second one is the Chi-squared statistic.

Finding a set of k tagging SNPs out of a total set of n SNPs requires evaluating $\binom{n}{k}$ different combinations. It is computationally infeasible to exhaustively search the optimal solution when n is usually large. In this study, we describe a global search heuristic based on a randomized procedure (Gibbs sampling) that aims to find a set of tagging SNPs with minimum redundancy. Although the stochastic nature of Gibbs sampling is presumed to prevent it from becoming completely trapped in local optima, it still requires a better initial value due to strong disturbance from the local optima. We propose a pre-clustering approach to obtain a better initial SNP set. The effect of pre-clustering will be investigated.

The paper is organized as follows. In Section 2, we explain our Gibbs sampling approach to obtain the minimum redundancy SNP set. The experiments and results will be discussed in Sections 3 & 4. We conclude our paper in Section 5.

2 Methods

2.1 Redundancy Measures

Consider two biallelic loci, locus 1 with alleles a and A , locus 2 with alleles b and B . Suppose the frequencies for alleles A and a are P_A and $1 - P_A$, the frequencies for alleles B and b are P_B and $1 - P_B$, and the the frequency of genotypes having allele A at locus 1 and allele B at locus 2 is P_{AB} . The commonly used *linkage disequilibrium measure* r^2 [14] is defined as

$$r^2 = \frac{(P_{AB} - P_A P_B)^2}{P_A(1 - P_A)P_B(1 - P_B)} \quad (1)$$

The mutual dependency of two random variables can also be used as a redundancy measure. Here redundancy and

correlation are used interchangeably. The *mutual information* between SNP X and SNP Y is defined as

$$MI(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2)$$

where both X and Y are discrete variables, $p(x, y)$ is the joint probability and $p(x)$ and $p(y)$ are marginal probabilities.

Chi-squared test of independence is adopted here to measure the correlation between two SNPs. For SNP X and SNP Y , we first obtain a contingency table between the two SNPs. The *Chi-squared statistic* is defined as

$$\chi_s^2(X, Y) = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

where r is the number of alleles for SNP X , and c is the number of alleles for SNP Y , O_{ij} is the observed joint frequency for i^{th} allele of SNP X and j^{th} allele of SNP Y , and E_{ij} is the expected frequency which is given by

$$E_{ij} = \frac{\sum_{k=1}^c O_{ik} \sum_{k=1}^r O_{kj}}{N} \quad (4)$$

where N is the total number of samples. A higher value of χ_s^2 indicates a stronger association between the two SNPs.

For a set S consisting of k SNPs, the total pair-wise mutual information is defined as

$$MISUM(S) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k MI(SNP_i, SNP_j) \quad (5)$$

The total pair-wise Chi-squared statistics is defined as

$$CHISUM(S) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k \chi_s^2(SNP_i, SNP_j) \quad (6)$$

The total pair-wise r^2 measure is give by

$$R2SUM(S) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k r^2(SNP_i, SNP_j) \quad (7)$$

2.2 Clustering of SNP Data

Due to the nominal nature of SNP data, the commonly used K-means clustering and its many variants are not suitable. In this study, we first obtain a similarity matrix using a similarity measure that is applicable for nominal data such as Chi-square statistic or mutual information. The distance matrix is then obtained by subtracting each entry from the maximum of all the values. We then apply the agglomerative clustering procedure with complete linkage to obtain the desired number of clusters.

2.3 Gibbs Sampling

Gibbs sampling is a special case of the Metropolis–Hastings algorithm. It is a stochastic global search heuristic for optimization problems. However, it still requires a better starting set to avoid being trapped in local optima. A pre-clustering is proposed to avoid the disturbance from local optima. To achieve this, we first cluster the SNPs into K groups, and randomly select an SNP from each group to form the initial SNP set. We then follow a Gibbs sampling procedure to find a set of K SNPs that minimize a goal function. The goal function can be one of the functions defined in equation 5-7. The detail of our approach is summarized in Figure 1.

```

Input: S is the total set of N SNPs, ε is a predefined threshold value
Output: C is the set of K chosen SNPs
minRedundancySNP(S, C, ε)
    Cluster the set S into K groups via customized hierarchical clustering
    Form set G of K members where Gi is the ith cluster
    Form initial set C by randomly pick one SNP from each of the K clusters
    while( a predefined maximum iteration is not reached)
        Randomly pick a number n from 1 to K
        Find a SNP x in Gn that minimizes MISUM/CHISUM/R2SUM
        Replace Cn with x
        Return C if the improvement of MISUM/CHISUM/R2SUM is less than ε
    end
Return C

```

Figure 1. The pseudocode for finding the minimum redundancy SNP set via Gibbs sampling with pre-clustering.

2.4 Prediction of Non-tagging SNPs with Tagging SNPs

Once the tagging SNP set is found, they can be used to predict the genotype values of the non-tagging SNPs. Many machine learning and statistical models can be used for this goal, including logistic regression [15], neural networks, support vector machines (SVM) [16], and random forest [17]. In this study, we conduct our experiments using logistic regression and SVM. We choose a K -fold cross validation to evaluate the effectiveness of our method. Our K -fold cross validation procedure is similar to the leave-one-out cross validation procedure for SNP prediction described in [1] where K is equal to the number of observations in the original sample.

3 Experimental Data

The following datasets are used to validate our method.

IBD 5q31: This data set is from an inflammatory bowel disease study of father-mother-child trios [18]. The original data set contained 103 SNPs in 387 subjects. Using the PHASE 2.0.2 software to derive haplotypes resulted in 103 non-singletons from 774 phased chromosomes.

TRPM8: The phased haplotype data was downloaded from Hapmap Data release 24. It contains 101 SNPs from 119 phased chromosomes.

4 Results and Discussion

4.1 Effect of Pre-clustering on the Convergence of the Gibbs Sampling Process

In order to test how fast the Gibbs sampling process converges, we obtained the convergence curve using all three measures introduced in Section 2.1 (i.e., the linkage disequilibrium measure, mutual information, and Chi-

squared statistic). Figure 2 shows the convergence process while attempting to find 10 tagging SNPs.

In each case, the Gibbs sampling process converged within 100 iterations regardless of whether or not pre-clustering was applied. However, the resulting set of SNPs had smaller redundancy measures when pre-clustering was used. Without pre-clustering, there is still some disturbance from local optima that affect the global minimum search process through Gibbs sampling.

4.2 Tagging Results

We conducted our experiments on the three distance measures to find tagging SNPs using the randomized algorithm mentioned above. The tagging results for IBD data set using r^2 , Chi-squared statistic and mutual information are shown in Table I, II, III respectively. The tagging results for TRPM8 data set using r^2 , Chi-squared statistic and mutual information are shown in Table IV, V, VI respectively.

For IBD data, pre-clustering is able to improve prediction performance. With pre-clustering, our 10-fold cross validation results are comparable with published leave-one-out cross validation results obtained by He et al. [1] and better than the results obtained by FSFS [11] (Table II, III). SVM and logistic regression show similar performance. Although our method does not present significant advantages over He's method [1], our method is simpler and does not rely on specific machine learning model to guide the selection process which is susceptible to over-fitting. In addition, the prediction based selection method SVM/STSA [1] requires calling SVM model during each stepwise selection process. This can be expensive due to the overhead of the prediction algorithm.

Among the three distance measures, both Chi-squared statistic and mutual information performed better than r^2 measure. This proves both of them can be used to study SNP association, and they are particularly useful for genotype data that sometimes involve more than two alleles.

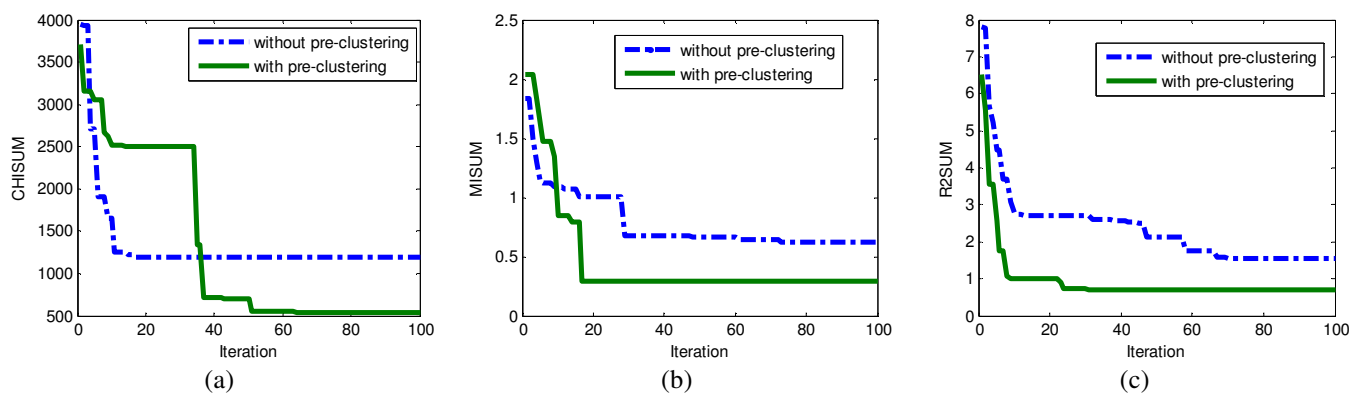


Figure 2. Convergence curve for the Gibbs sampling process based on three redundancy measures. (a) minimization process of CHISUM (b) minimization process of MISUM (c) minimization process of R2SUM.

For TRPM8, our ten-fold cross validation prediction performance is better than SVM/STSA when mutual information is used as correlation measure (Table V). Similar performances are observed between r^2 measure and Chi-squared statistic. The pre-clustering does not improve the prediction performance significantly. This indicates that the disturbance from local optima in this data set is not as strong as in IBD data set.

TABLE I. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON IBD DATA SET (CORRELATION IS MEASURED WITH r^2 MEASURE, K IS THE NUMBER OF TAGGING SNPs).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	81.1%	80.7%	76.6%	76.6%
5	81.5%	81.2%	78.8%	77.4%
10	93.5%	91.7%	77.4%	77.4%
20	98.2%	97.8%	85.7%	86.0%
30	98.5%	98.3%	93.5%	93.5%

TABLE II. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON IBD DATA SET (CORRELATION IS MEASURED WITH CHI-SQUARED STATISTIC, K IS THE NUMBER OF TAGGING SNPs).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	85.6%	85.5%	79.3%	79.3%
5	86.0%	85.1%	81.5%	81.1%
10	95.0%	93.3%	94.4%	93.5%
20	98.2%	97.9%	98.1%	97.5%
30	98.5%	98.5%	97.8%	97.4%

TABLE III. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON IBD DATA SET (CORRELATION IS MEASURED WITH MUTUAL INFORMATION, K IS THE NUMBER OF TAGGING SNPs).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	86.6%	86.5%	80.0%	79.9%
5	87.3%	86.1%	79.8%	79.8%
10	96.0%	95.0%	77.4%	77.3%
20	98.2%	97.8%	89.8%	88.4%
30	98.5%	98.3%	97.3%	96.4%

TABLE IV. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON TRPM8 DATA SET (CORRELATION IS MEASURED WITH r^2 MEASURE, K IS THE NUMBER OF TAGGING SNPs).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	89.1%	87.9%	87.6%	81.3%
5	88.9%	86.2%	87.5%	92.3%
10	91.7%	91.3%	91.3%	92.4%
20	99.5%	99.7%	99.2%	99.2%
30	99.7%	99.7%	99.3%	99.7%

TABLE V. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON TRPM8 DATA SET (CORRELATION IS MEASURED WITH CHI-SQUARED STATISTIC, K IS THE NUMBER OF TAGGING SNPs).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	89.1%	87.9%	87.6%	81.3%
5	88.9%	86.2%	87.5%	92.3%
10	91.7%	91.3%	91.3%	92.4%
20	99.5%	99.6%	99.2%	98.6%
30	99.7%	99.7%	99.3%	99.7%

TABLE VI. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON TRPM8 DATA SET (CORRELATION IS MEASURED WITH MUTUAL INFORMATION, K IS THE NUMBER OF TAGGING SNPs).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	96.7%	90.1%	89.3%	81.8%
5	97.3%	92.1%	96.3%	92.3%
10	97.3%	92.3%	96.0%	92.2%
20	99.7%	99.7%	97.4%	98.5%
30	99.7%	99.7%	98.1%	99.7%

4.3 Running Time Results

The running time required to select different number of tagging SNPs using our Gibbs sampling procedure is shown in Table VII. Our Gibbs sampling code is implemented with R statistical programming language.

The running time increases as the number of tagging SNPs increases. The running time results are similar between the Chi-squared statistic and mutual information. The program often ran a little faster with r^2 as correlation measure. Comparing with prediction guided SNP selection SVM/STSA [1] which takes up to 1 day to find 10 tagging SNPs for IBD data set, and 23 hours to find 10 tagging SNPs for TRPM8 data. It even took several hours to find 1 tagging SNPs[1], our Gibbs sampling procedure runs much faster and usually completes within 5 minutes for up to 30 tagging SNPs.

TABLE VII. RUNNING TIME REQUIRED (SECONDS) TO SELECT TAGGING SNPs USING DIFFERENT CORRELATION MEASURES (K IS THE NUMBER OF TAGGING SNPs, ALL EXPERIMENTS ARE PERFORMED ON A COMPUTER WITH AMD ATHLON II X4 620, 2.61 GHZ PROCESSOR AND 2 GB OF RAM)

Data set	IBD			TRPM8		
	Correlation measure			Correlation measure		
	r^2	χ^2	MI	r^2	χ^2	MI
3	11.57	11.75	11.28	9.11	10.52	9.25
5	20.68	19.83	23.11	16.62	10.86	21.32
10	53.44	63.20	53.00	25.56	27.30	34.15
20	94.22	136.23	120.62	80.38	107.73	91.69
30	196.94	195.83	191.46	121.27	157.03	157.81

5 Conclusions

We investigated a block-free stochastic global search heuristic to find a set of minimum redundancy tagging SNPs. It is a randomized search technique based on Gibbs sampling. We modified the basic Gibbs sampling procedure by adding a pre-clustering step to find a better starting set. In order to properly cluster the SNP data, we applied hierarchical clustering with a distance measure that is applicable for nominal data. The Gibbs sampling process typically converges faster and reaches lower minimum if a pre-clustering is used. Pre-clustering improves the tagging prediction accuracy if there is a disturbance from local optima. If there is little disturbance from local optima, pre-clustering at least does no harm.

Although our tagging process is driven by a simple objective function that aims to minimize redundancy among a set of SNPs instead of being driven by a prediction method such as SVM, we are able to obtain comparable prediction results while running much faster than prediction based SNP selection method [1].

We also proposed two correlation measures to study SNP association. They proved to be as effective as the commonly used r^2 measure. These two measures can be useful for genetic features (e.g. genotypes) that could have more than two alleles.

6 Acknowledgments

This work was supported by the National Institutes of Health [5T36GM008789-08].

7 References

- [1] J. He and A. Zelikovsky, "Informative SNP Selection Methods Based on SNP Prediction," *Nanobioscience*, vol. 6, pp. 60-67, July 18, 2006 2006.
- [2] L. Kruglyak and D. Nickerson, "Variation is the spice of life.," *Nat Genet*, vol. 27, pp. 234-236, 2001.
- [3] D. Reich, *et al.*, "Linkage disequilibrium in the human genome.," *Nature*, vol. 411, pp. 199-204, 2001.
- [4] H. Ackerman, *et al.*, "Haplotypic analysis of the TNF locus by association efficiency and entropy.," *Genome Biology*, vol. 4, p. R24, 2003.
- [5] X. Ke and L. R. Cardon, "Efficient selective screening of haplotype tag SNPs," *Bioinformatics*, vol. 19, pp. 287-288, January 22, 2003 2003.
- [6] Z. Meng, *et al.*, "Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes.," *The American Society of Human Genetics*, vol. 73, pp. 115-130, June 5 2003.
- [7] K. Zhang, *et al.*, "A dynamic programming algorithm for haplotype block partitioning.," *Proc Natl Acad Sci*, vol. 99, pp. 7335-7339, 2002.
- [8] B. Halldorsson, *et al.*, "Optimal selection of SNP markers for disease association studies.," *Hum Hered*, vol. 58, pp. 190-202, 2004.
- [9] C. Carlson, *et al.*, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.," *Am J Hum Genet.*, vol. 74, pp. 106-120, 2004.
- [10] B. Howie, *et al.*, "Efficient selection of tagging single-nucleotide polymorphisms in multiple populations," *Human Genetics*, vol. 120, pp. 58-68, 2006.
- [11] T. M. Phuong, *et al.*, "Choosing SNPs Using Feature Selection," *Proceedings IEEE Computational Systems Bioinformatics Conference*, pp. 301-309, 2005.
- [12] P. Sebastian, *et al.*, "Minimal Haplotype tagging," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 9900-9905, 2003.
- [13] J. He and A. Zelikovsky, "MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression," *Bioinformatics*, vol. 22, pp. 2558-2561, October 15, 2006 2006.
- [14] W. G. Hill and A. Robertson, "Linkage disequilibrium in finite populations," *TAG Theoretical and Applied Genetics*, vol. 38, pp. 226-231, 1968.
- [15] A. Agresti, *Categorical Data Analysis*. New York: Wiley-Interscience, 2002.
- [16] C. Cortes and V. Vapnik, "Support Vector Network," *Machine Learning*, vol. 20, pp. 273-297, 1995.
- [17] L. Breiman, "Random Forests.," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [18] M. J. Daly, *et al.*, "High-resolution haplotype structure in the human genome," *Nature Genetics*, vol. 29, pp. 229-232, 2001.

Development Approach and Architecture of GenSAS: the Genome Sequence Annotation Server

T. Lee¹, I. Cho², C. Peace¹, S. Jung¹, P. Zheng¹, and D. Main¹

¹Horticulture and Landscape Architecture, Washington State University, Pullman, WA, U.S.A.

²Computer Science, Saginaw Valley State University, Saginaw, MI, U.S.A.

Abstract - *Advances in DNA sequencing technology have significantly reduced the costs associated with sequencing an organism's genome. However, the operating costs of hardware, software, and labor to analyze the sequence data are still too high for most users to process in house. Henceforth, most of the current bioinformatics applications used by bench scientists will be accessible through a Web environment. This paper presents GenSAS, the **Genome Sequence Annotation Server**, a JavaScript-based framework of gene prediction and comparative sequence similarity applications for structural and functional sequence annotation. Among other web-based genome annotation pipelines, GenSAS is unique in that it offers a one-stop website with a single graphical interface for running multiple structural and functional annotation tools, visualization and manual curation of genome. We present its functionality, the technology used in implementing each functionality, and software architecture of the overall implementation.*

Keywords: genomics tool, genome, sequencing, annotation, architecture

1 Introduction

An important component of specialized genomic databases that serve a specific community is to provide useful online tools for researchers to conduct web-based sequence analysis. These web-based tools can include BLAST (Basic Local Alignment Search Tool) [1] and FASTA [2] servers for pair wise comparison of clade-specific datasets, sequence assembly tools to assemble EST transcripts and microsatellite detection and primer identification tools. With the advances of sequencing technology, more and more clade-specific databases have started to store and display whole genome sequences with automatic gene annotation data using graphic viewers such as GBrowse [3]. Automatic gene annotation often needs to be refined by further analysis. There are many gene prediction algorithms and pipelines available to help in gene identification, but there are no online tools that easily allow biologists to readily combine the evidence from several gene prediction tools and create curated gene models within the same graphic interface. We have implemented a flexible online tool called GenSAS (Genome Sequence Annotation Server, www.bioinfo.wsu.edu/gensas) for genome sequence

annotation that can assist researchers in identifying genes in genomic sequences for the Rosaceae family. GenSAS is implemented in a modular way to allow it to be easily used with other genome annotation projects.

Tool development is one of the key areas in bioinformatics research, along with the analysis and interpretation of genome data. A tool can be developed from the ground up and fine tuned for specific applications, but such a tool often ends up only being used by its developers rather than being offered for wider community use. In many cases, development of web-based systems has been ad hoc, lacking systematic approach, quality control and assurance procedures [4]. Such problems are inherent in most software development, but made worse in a Web environment due to the rapid growth of the Web, high demand for web-based applications, shorter time-to-market requirements, and relatively short history of the Web (less than 20 years). While not an exception for bioinformatics tool development, it is a challenge for bioinformatics tool developers to follow a disciplined engineering approach so that tools remain usable and stable against deviation from original assumptions about their optimal working condition. This paper does not attempt to provide a solution to all the problems mentioned. Instead, it shows one success story in bioinformatics tools development (GenSAS) and the approach taken to make the tool widely useful by researchers. GenSAS was developed with the following objectives:

- To develop a computational pipeline incorporating multiple genome sequence annotation tools.
- To develop a visualization tool to display the output from annotation tools graphically.
- To develop intuitive web-based user interfaces to facilitate curation by biologists

We first examine, in section 2, the progress of Web application development and Web architecture in recent years. In section 3, general and specific activities involved in the gene annotation process are presented. In section 4, we introduce the implementation details for GenSAS and software components that GenSAS supports. After this, related work is discussed. This paper concludes with a summary and future work in section 6.

2 Progress in Web Application Development and Web Architecture

When the Web was created in early 1990s, most websites were simple and served static content. Most emphasis was on content layouts and overall look, and easy maneuvering of the site. Little programming was required and no rigorous software engineering were required to build such websites. The growth of online resources soon made it necessary to implement search engines on the Web and process the user provided input from the Web browser. CGI was the first mechanism that allowed Web clients to execute programs on a Web server and receive their output [5]. The further growth of the Internet and World Wide Web led to full blown software applications available on the Web, rapid uncontrolled growth, hastily written code and a lack of Web standards. All this contributed to what is known as the Web Crisis. The wide use of Web applications from all over the world made them only more vulnerable to failure. To remedy the crisis and support the development of quality Web applications, the field of *Web Engineering* emerged to provide scientific, engineering and management principles and disciplined and systematic approaches to the successful development, deployment and maintenance of high quality web-based systems and applications [4]. Web engineering shares some of its principles with traditional software engineering, but it also has unique requirements: shorter development time, content-oriented development, greater importance of visual look and feel, and a more diverse user demographics.

A key area in software engineering is Software Architecture which is defined as "The software architecture of a program or computing system is the structure of the system, which comprises software components, the externally visible properties of those components, and the relationship among them [6]." There are many different architectural styles used in software applications, and it is important to decide which architecture is the best fit for the software development. The layered (or multi-tiered) architecture has many benefits: interoperability, flexibility, maintainability, and reusability, to name a few. The Communication Network protocol is an example of layered architecture. Structured Web applications also reveal multi layers where the presentation, the application processing, and the data management are logically separate processes, as shown in Figure 1. We will look at how GenSAS fits in this architecture in section 4.

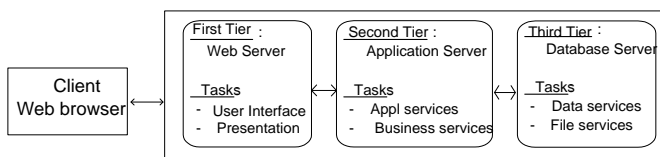


Figure 1. Multi-Tier Web Architecture

3 Genome Annotation Process

After a genome has been sequenced and assembled, the process of genome annotation starts. The purpose of genome annotation is to understand the content of the genome through locating genes and other sequence features and determining gene putative function. The annotation process can be categorized into manual and automated annotations [7], and structural and functional annotations [8, 9].

While manual annotation tends to deliver higher quality results over automated annotation, it is time consuming and expensive process, particularly impractical for large-scale whole genome sequence data. In contrast, automated annotation is a relatively inexpensive and fast process, but the output is less reliable, typically ranging from 30-70% accuracy for predicting a relatively small sample of known genes [10]. Structural annotation focuses on identifying the genomic elements on a sequence. Genomic elements include regulatory motifs, repetitive sequences, gene structure and Open Reading Frames (ORFs). Gene identification (or prediction) tools are based on statistics (*ab initio*) or sequence similarity based methods. Because each approach has its own strengths and weaknesses, it is common for gene identification tools of both types (a hybrid approach) to be used in gene annotation. Statistics based methods do not use extra information for gene prediction. Instead, they identify genomic features based on statistical patterns inside and outside of gene regions as well as patterns typical of the gene boundaries. GENSCAN is one of the most widely used statistics-based gene prediction software for human and vertebrates [11]. Other statistics based tools in wide use are FGENESH [12], GlimmerM [13], and GeneMark [14]. They use algorithms based on Markov models [15] and dynamic programming [12].

Systems have also been developed which integrate the results from several gene prediction tools and the evidence from cDNA/ESTs and protein alignments. JIGSAW, formally known as Combiner [17] is a gene prediction system that utilizes multiple sources of evidence to predict gene structure. A weight is assigned to each evidence source, and gene predictions are based on a weighted voting scheme, yielding the best consensus predictions.

Sequence similarity-based gene prediction methods are typically more reliable than statistics-based methods as experimental data are used to predict the genes. The target genome is searched for similar regions in existing sequences such as ESTs from the same species of known gene models from closely related species. The rationale behind this methodology is that homologous sequences from closely related organisms typically share evolutionarily conserved or common functions. However, sequence similarity-based tools are useful only if existing sequence data are available. For example, genes that are expressed at low levels or expressed in certain cell types, developmental stages, or growth conditions may not be adequately represented. To remedy the shortcomings of each approach, systems have been developed

to combine and integrate the results from several gene prediction tools, as in GenomeScan [16].

Functional annotation is the process of attaching biological information to the genomic elements identified during structural annotation. Such information includes, but is not limited to, biochemical function, biological function, physiological function, and Gene Ontology (GO) terms. A general approach for functional characterization of unknown genes is to infer protein functions based on significant sequence similarity to annotated proteins in sequence databases. Typically, a sequence of a gene with unknown function is compared against public databases such as Swiss-Prot [18], TrEMBL [19] or NCBI [20] using the BLAST sequence similarity algorithm.

Each of the tools mentioned have their own attributes, for instance, certain annotation tools are more robust for certain species than others, having originally been developed for those species. Generally, statistics-based methods find genes with a full-length CDS (coding sequence) but they perform poorly on finding genes with partial CDS which can be annotated more correctly with sequence similarity-based methods. Thus, quality of the results generated from each tool is not regarded as equal; some results are more reliable than others, and the result varies in different circumstances. Therefore, it is unwise to rely on only one source of evidence but rather best to combine different types of results to draw conclusions.

Often computational annotation programs generate results in text formats. Thus, several visualization tools have been developed to display the text file data graphically so that researchers can view and interpret the results more easily. The Generic Genome Browser (GBrowse) [3] is one of the most widely used products developed through the Generic Model Organism Database (GMOD) project (<http://gmod.org>). As it is a web-based application, annotation data can be easily shared with other researchers. However, it does not allow users to dynamically edit the annotated genomic elements.

In summary, to effectively apply structural and functional annotations, researchers are required to understand the different attributes of annotations tools, and how to specify proper parameters for their genome of interest. Also, the need for visualization through setup and management of genome viewers can be overwhelming for some researchers. Researchers typically use several annotation tools and obtain results for DNA sequences of interest in text format. In some cases, researchers must wait for the result by email when the process is computationally intensive or the sequence is very large. Then, they need to convert the text result to a format specifically required by a specific genome viewer using scripting languages like Perl. Finally, researchers analyze and

compare the annotation data from different sources of evidences on the viewing tracks provided by the program. As such, researchers have to go through many time-consuming steps before reaching the final analysis steps, and currently no one-stop website exists for researchers to access several gene prediction tools and have the integrated and optimized results returned to them for further analysis.

4 GenSAS

GenSAS was developed to help researchers perform structural and functional gene annotations and provides visualization curation tools. The focus of the design was on usability and effectiveness for biologists, efficient maintenance and decreased cost for IT administrators and developers. Being a web-based tool rather than a standalone tool frees users from expending effort in installation, configuration, and upgrade. The tool was made simple to use by providing all gene annotation tasks in one Web interface.

Figure 2 shows the Web front end of GenSAS. The front page consists of five panels; User Information, Sequence Information, Task Information, Retrieve Saved Data, and Task Queue. These panels are designed to assist researchers to create various tasks intuitively and efficiently. To create a task in GenSAS, users upload a genomic sequence and select one or more annotation tools. Then, the newly created task is appended to the Task Queue panel. GenSAS currently supports nine annotation tools as listed in the Tool Information panel, but more will be added in the future.

GenSAS allows researchers to save four different types of results: Sequence, Task, Output and SVG. These results can be later retrieved to reduce redundancy in the task creation process. Clicking on the third column icon (SVG button) on the Task Queue panel will generate a report page. The reporting page in Figure 3 displays the results from nine tools and a custom track for notes as a curator manually evaluates the annotations. Genomic features from the results of gene prediction programs are colored distinctively and types of these features are identified in the legend table. The reporting page allows users to zoom in and out or set the zoom ratio, and scroll left and right to examine the genomic features in the desired location of the DNA sequence.

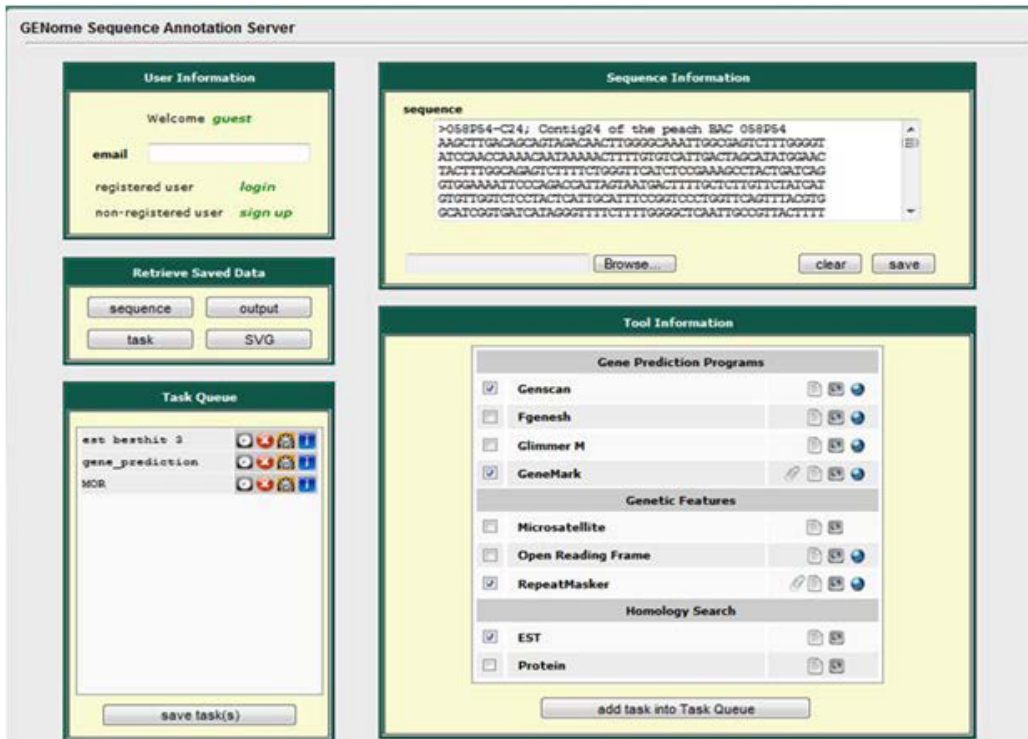


Figure 2: GenSAS Front Page

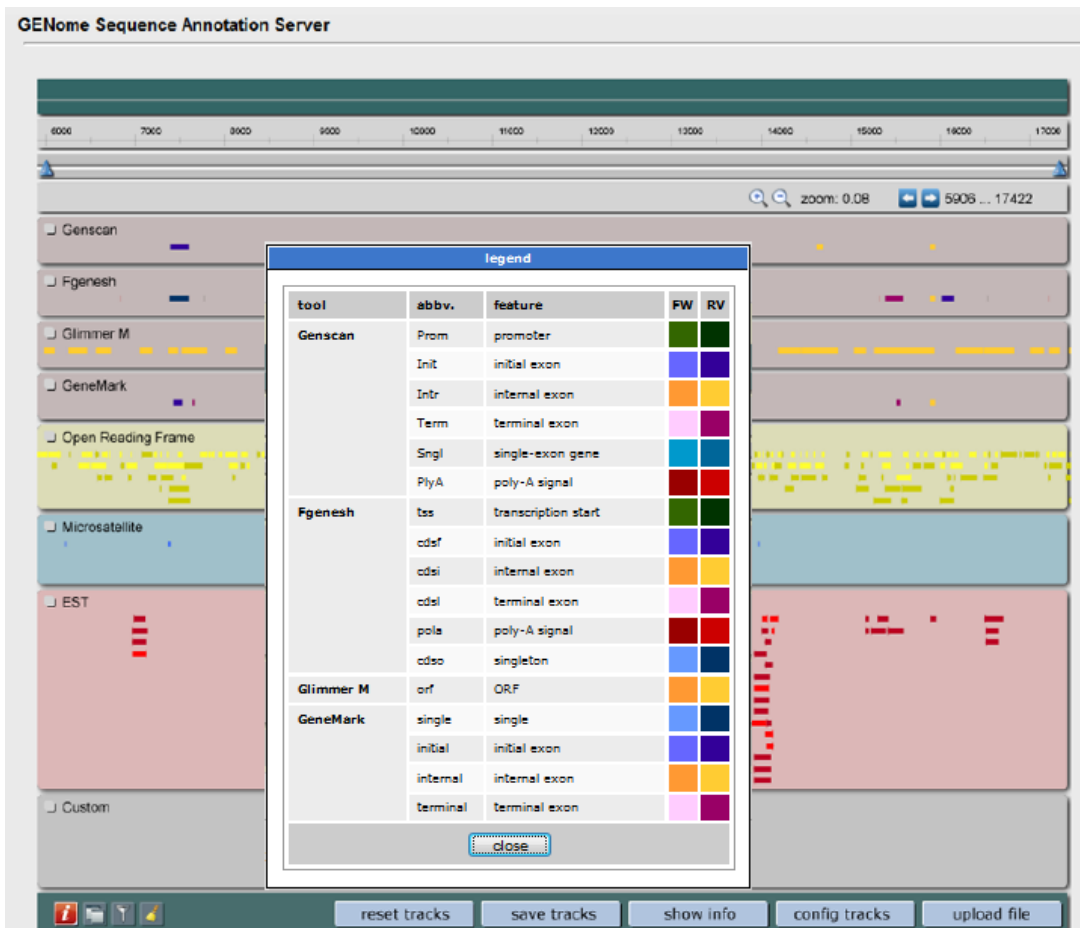


Figure 3. GenSAS Reporting Page with legend

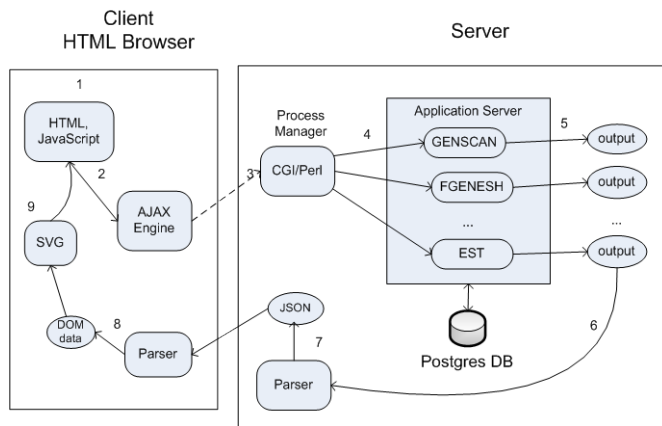


Figure 4. Overall structure of GenSAS

Figure 4 shows the overall structure of GenSAS. The software architecture conforms to the three-tier architecture covered in section 2. The rounded rectangles represent running programs or processes, and the oval shapes represent data. The typical interactions between the client and server in gene annotation are:

1. The user interacts with the GenSAS front page and creates tasks.
2. The user sends annotation job to the server (by clicking the SVG button) and waits for the result.
3. Asynchronous job requests are created and sent to the Web server running CGI/Perl (the dotted line indicates asynchronous communication).
4. CGI/Perl engine parses the input and calls exec to execute UNIX commands to run the corresponding annotation server.
5. The annotation tools execute the command and generate output in their own text formats.
6. The parser reads the text data and converts to JSON data.
7. The JSON data is sent back to the client machine.
8. The parser on the client side converts the JSON data to DOM data.
9. JavaScript works with SVG to render the data to be displayed on the Web browser.

The core technologies used in GenSAS are JavaScript, JSON, AJAX, SVG, PostgreSQL database, and CGI/Perl, all open-source and freely available tools. We now look at the details of each technology incorporated in GenSAS.

JavaScript is a scripting language commonly used in the development of client-side dynamic Web applications. It is embedded directly into HTML pages and can be readily used in most Web browsers without any further installation or configuration. It is most commonly used in creating dynamic contents into an HTML page. GenSAS uses JavaScript in various ways. It is used to manipulate HTML elements on a webpage via Document Object Model (DOM,

<http://www.w3.org/DOM>). DOM is the primary data structure by which a Web browser represents an HTML page. It provides methods and properties to retrieve, modify, update, and delete elements of a HTML document. For example, JavaScript allows users to reorder annotation tracks by drag & drop. JavaScript is also used for handling events. When an event takes place from mouse click or drag & drop, one or more corresponding JavaScript functions are called and executed to properly handle the event on the client machine. Instead of waiting for the server to respond to the user actions and reloading the entire browser page, JavaScript allows users to interact with the browser without disruption.

JSON (JavaScript Object Notation) is a common data exchange format which provides a structure to data like XML (<http://www.json.org>). It is in human readable text format and easy for machines to parse and generate. It is a native data format for JavaScript and less complicated than XML to work with for most modern programming languages. Also, JSON data can be easily transmitted between client and server machines over the network. The format of the results from each annotation tool varies and these results need to be converted into one standard format before sent to the Client Web browser. Its portability makes JSON well suited for web-based applications that intensively use JavaScript. JSON data can be easily converted to DOM format so that JavaScript can work with SVG to render the data to be displayed on the Web browser. With these reasons JSON is used as the standard text format in GenSAS; outputs from annotation tools as well as some of data types that users can save in their account in the database are formatted in JSON.

AJAX (Asynchronous JavaScript and XML) is a Web development technique that is used to create interactive Web applications on the client-side (<http://www.ajax.org>). AJAX is not a single technology but a combined technology of HTML, CSS, DOM, XML, and JavaScript. Traditionally, once a client request is sent to a server, the client has to wait for the response from the server without being able to do further work on the Web browser. When the response reaches to the client, the whole webpage needs to be refreshed to display the result, which results in disruption of user attention. Using AJAX, the client accesses the server asynchronously in the background without waiting for the response. Once the response arrives on the client machine, the Web browser displays the result without refreshing the whole page. AJAX is frequently used in GenSAS. All access requests to the database server are carried out in the background with AJAX. AJAX is also used to perform parallel processing on annotation tools on multi-processor server. Parallel processing (or concurrent processing on single processor server) is crucial for this system as the processing time of annotation tools varies significantly. It avoids having to wait for the completion of a large process before viewing the results of other smaller processes.

SVG (Scalable Vector Graphics) is a vector graphics file format and Web development language based on XML (<http://www.w3.org/TR/SVG>). Raster graphics, sometimes

called bitmap, is based on pixels and it represents an image as an array of pixels. Some genome browsers like GBrowse use raster graphics and generate image files to be sent and displayed on the client Web browser. The size of raster graphics files are relatively large compared to vector graphics and it degrades the performance of GBrowse. Also, when zoomed in, the images lose the quality with jagged lines, while vector graphics easily scale up without degrading the quality. As images for genomic features are needed to be displayed distinctively when scaled up, vector graphics is well suitable for the annotation server. Most modern Web browsers support and render SVG markup either natively (in Firefox and Safari) or with plug-ins (in Internet Explorer 8) to view SVG images correctly on Web browser. Internet Explorer 9 will natively support and render SVG. SVG is the main force behind the reporting page. Together with JavaScript, interactive graphical Web applications can be efficiently developed. All graphic features on the page are drawn with SVG images either statically or dynamically. For example, when one of exon images is clicked, it triggers the script that pops up the dialog window which shows the information about the exon such as orientation, frame and, start and stop locations. SVG graphics is also used to create GUIs such as buttons and a slider with two thumbs on the reporting page. In general, GUIs created by HTML tags are relatively plain and simple, however, the appearances of these HTML GUI components vary based on types of platforms and browsers; the looks of the buttons on the same webpage viewed by Safari and IE, for example, become different. SVG graphics allows for developers to create any shape and color, and the appearances of these GUIs will not change across browsers or platforms. Because SVG is written in XML, SVG content can be easily manipulated from JavaScript with DOM API in GenSAS.

PostgreSQL is one of the most popular relational database management systems publically and freely available (<http://www.postgresql.org>). It is used as a database server residing in the background of the annotation server system. It manages information about user accounts as well as their data. The Perl script has a module called Database Interface (DBI). DBI offers the standard database interface, which is capable of conducting primitive database functions on various types of database systems. DBI allows the database server to efficiently perform database operations online.

CGI/Perl The Common Gateway Interface (CGI) is a mechanism that allows Web clients to execute programs on a Web server and to receive their output. CGI applications are often used to produce HTML pages on the fly and process the input from an HTML form [5]. While many programming languages like C/C++, Java, Visual Basic, and Perl can be used to implement CGI, Perl is most often used to write CGI scripts for Web servers due to its long history of usage in UNIX systems and its strength in text manipulation. It is optimized for scanning arbitrary text files, extracting information from those text files, and printing reports based on that information. A project called BioPerl is supported by Open Bioinformatics Foundation (<http://www.open-bio.org>),

which further strengthens its popularity. In GenSAS, Perl script is used to build CGI pages and access the database server to manage users' data. In addition, Perl is used to execute annotation tools installed on the server using the exec functions. With this function together with the AJAX described above, various annotation tools can be simultaneously executed on the annotation server to perform parallel processing.

5 Related Work

JIGSAW [17] integrates weighted outputs from multiple gene prediction tools to predict genes. Ergatis [21] enables workflow creation with multiple bioinformatics tools to perform automated gene annotations and comparative analysis. However, these tools do not provide a graphic viewer for further annotation. MAKER [22] is a genome annotation pipeline that produces annotation results that can be viewed by GMOD browsers like GBrowse. DNA subway (<http://dnasubway.iplantcollaborative.org>) allows users to use multiple gene prediction tools, edit the gene model using Apollo [23] and view the results in GBrowse. It is the most similar tools to GenSAS, but GenSAS has its own graphic viewer that allows users to edit and view the results in the same window.

6 Conclusion and Future Work

GenSAS has been developed in close cooperation with biologist users. Interacting with users helped identify problems and issues in currently used gene annotation tools, and has brought forth new ideas for GenSAS features. Rather than developing from the ground up, GenSAS was developed with proven technologies and well supported standards-based tools. By conforming to the industry standard three-tier Web architecture, GenSAS can be easily managed and updated for future needs. The most important issues identified and put to work in GenSAS were the ease of use, prompt response, and effectiveness for the biologist user

Ease of Use: GenSAS incorporates several different annotation tools together with available customized experimental data such as cDNA, ESTs and proteins, to provide researchers with faster processing and access to the various types of generated evidence without ever leaving the GenSAS browser page. User management through accounts is supported and users can store output results which can be later retrieved for further analysis.

Prompt Response Time: AJAX allows easy implementation of concurrent and parallel processing on Web applications. GenSAS allows users to run multiple gene annotation tools, and by using asynchronous communication mechanisms in AJAX, the result will show up on the Web browser as soon as the corresponding annotation tool finishes its job. Also, AJAX allows users to continuously interact with the browser without having to wait for the server.

Effectiveness: The Web front end is very compact and the five panels of windows are well laid out for users to easily navigate. Any users with nominal experience of using gene annotation tools will be readily able to use GenSAS quite effectively. Different shapes and colors of icons with tooltip support further help users with easy navigation of the tool. Graphic features on a track can be customized with different colors, and they can be saved to the custom track for further evidence gathering.

Since its inception, GenSAS has been constantly improved and many issues have been suggested to further enhance its capability. One notable feature under development is support of multiple tracks for the same annotation tool run with different parameters. Allowing drag-and-drop for copying features onto the custom track is also being considered. GenSAS supports private and group user accounts for users to save and allow file sharing among group members, similar to UNIX file sharing, but more an advanced and versatile user account management system is desired. Utilizing a content management system like Drupal is one possibility. To further improve the performance, the annotation processes can be sent to high performance clusters or grids. We look forward to implementing these and additional features in future GenSAS versions.

Acknowledgements

We acknowledge the Department of Horticulture and Landscape Architecture of Washington State University for the support of this work. The future development of GenSAS will be supported through USDA NIFA Award #2011-67009-30030.

7 References

- [1] Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman, D. J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403-410.
- [2] Pearson W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.* 4: 1145-1160.
- [3] Stein L.D., Mungall C., Shu S., Caudy M., Mangone M., Day A., Nickerson E., Stajich J.E., Harris T.W., Arva A., and Lewis S. (2002) The Generic Genome Browser: A building block for a model organism system database. *Genome Res.* 12: 1599-1610
- [4] S. Hansen S. Murugesan, Y. Deshpande and A. Ginige. Web engineering: A new discipline for development of web-based systems. In Proceedings of the First ICSE Workshop on Web Engineering, 1999.
- [5] Deep J and Holfelder P. Developing CGI Applications with Perl. Wiley 1996.
- [6] Clements P, Bass L and Kazman R. Software Architecture in Practice. Addison Wesley, 1998.
- [7] Collins F.S., Morgan M., and Patrinos A. (2003) The Human Genome Project: lessons from large-scale biology. *Science*, 300, 286–290.
- [8] Head-Gordon, T.; Wooley, J. C. "Computational challenges in structural and functional genomics," *IBM Systems Journal*, vol.40, no.2, pp.265-296, 2001
- [9] Bright L, Burgess S, Chowdhary B, Swiderski C, and McCarthy M. *BMC Bioinformatics* 2009, 10:S8
- [10] Flicek P. (2007) Gene prediction: compare and CONTRAST. *Genome Biol.*; 8(12):233
- [11] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997, 268:78-94.
- [12] Solovyev V.V., Salamov A.A., and Lawrence C.B. (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc. Int. Conf. Intel.l Syst. Mol. Biol.*;3:367-75.
- [13] Pertea M. and Salzberg S.L. (2002) Using GlimmerM to find genes in eukaryotic genomes. *Curr Protoc Bioinformatics.* Nov;Chapter 4:Unit 4.4.
- [14] Borodovsky M. and Mcininch J. (1993) GenMark: parallel gene. recognition for both DNA strands. *Comput. & Chem.*, 17, 123–133.
- [15] Salzberg S., Delcher A., Kasif S., and White O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26:2 544-548.
- [16] Yeh R.F., Lim L.P., and Burge C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.* 11: 803-816.
- [17] Allen J.E. and Salzberg S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics.* Sep 15;21(18):3596-603.
- [18] Gasteiger E., Jung E., and Bairoch A. (2001) SWISS-PROT: Connecting Biomolecular Knowledge via a Protein Database. *Mol. Biol.*;3 (3): 47-55.
- [19] Bairoch A. and Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000". *Nucleic Acids Res.* 28: 45–48.
- [20] Maglott D, Ostell J, Pruitt KD, Tatusova T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 33:D54-8.
- [21] Hemmerich, C. et al. (2010) An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics* 26, 1122–1124.
- [22] Cantarel, B. et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18: 188-196.
- [23] Lewis S.E. et al. (2002) Apollo: a sequence annotation editor. *Genome Biol.* 3(12):RESEARCH0082.

Sequence Analysis to Predict Protein Active Sites using SSHM

P.Satheesh¹, B.Srinivas², M.Prasada Rao³, Col.Prof.Allam Apparao⁴, and G.Charles Babu⁵

¹ Associate Professor, CSE Department, MVGR College of Engineering, Vizianagaram, Andhrapradesh, India

² Assistant Professor, CSE Department, MVGR College of Engineering, Vizianagaram, Andhrapradesh, India

³ Lecturer, CSE Department, JNTU kakinada, Kakinada, Andhrapradesh, India

⁴ Vice-Chancellor, JNTU kakinada, Kakinada, Andhrapradesh, India

⁵ Professor, CSE Department, Vidya Vikas Institute of Technology, Chevella, Andhrapradesh, India

Abstract—*The advancement of bioinformatics is remarkable after the analysis of human genome is completed. The functions of the protein coded from the genome are compared with the sequence of amino-acid of unknown proteins and the sequence of the protein that is already known. It can find the similar sequence, but the global relations among the sequences cannot be extracted. Initially two protein sequences are taken one among which is the known amino acid sequence(seq A) and the other one is the unknown amino acid sequence(seq B). By Sequence mapping, comparison is made between the two sets. This method searches for the matched and the frequent set of unknown amino acid in the known amino acid sequence. The algorithm takes protein sequence as input and does the mapping. If there are zero matches of seq B in seq A then there exists noise, which is to be eliminated.*

Keywords: Sequence Mapping, Amino-acid, Frequency Set, Sequence Search Hill Climbing Algorithm(SSHM)

1. Introduction

Proteins are made up of chains known as amino acids which are bind together by the peptide bonds. Protein structure is determined by the nucleotide sequence of that protein. Amino acids are made up of carbon, hydrogen, oxygen and nitrogen which will combine to form different types of proteins which are required by the body. PHP is a scripting language which is used for web development. PHP can run on any existing platforms based on the same code. Using PHP as the scripting language fastens the execution, moreover it supports several database and HTTP server interfaces. We implemented heuristic search based on Hill Climbing Algorithm in PHP. Heuristic search method is used to find the best solution in least possible time. Hill Climbing Algorithm is an Iterative algorithm which starts with a solution and then incrementally finds a better solution by changing a single element of the solution. The matched amino acid sequence is searched and frequency set is generated.

2. Scripting Languages

Specialised scripting languages include: PHP (Hypertext PreProcessor). It is popular scripting language which has more than thousand inbuilt function support. And has nested, associative arrays features.

Perl (Practical Extraction and Report Language). This is a popular string processing language for writing small scripts for system administrators and web site maintainers. Much web development is now done using Perl. newline Hypertalk is another example. It is the underlying scripting language of HyperCard.

Lingo is the scripting language of Macromedia Director, an authoring system for develop high-performance multimedia content and applications for CDs, DVDs and the Internet.

AppleScript, a scripting language for the Macintosh allows the user to send commands to the operating system to, for example open applications, carry out complex data operations.

JavaScript, perhaps the most publicised and well-known scripting language was initially developed by Netscape as LiveScript to allow more functionality and enhancement to web page authoring that raw HTML could not accommodate. A standard version of JavaScript was later developed to work in both Netscape and Microsoft's Internet Explorer, thus making the language to a large extent, universal. This means that JavaScript code can run on any platform that has a JavaScript interpreter.

VBScript, a cut-down version of Visual Basic, used to enhance the features of web pages in Internet Explorer.

2.1 PHP scripting

PHP is an HTML-embedded scripting language. Much of its syntax is borrowed from C, Java and Perl with a couple of unique PHP-specific features thrown in. The goal of the language is to allow web developers to write dynamically generated pages quickly."

This is generally a good definition of PHP. However, it does contain a lot of terms you may not be used to. Another way to think of PHP is a powerful, behind the scenes scripting language that your visitors won't see!

When someone visits your PHP webpage, your web server

processes the PHP code. It then sees which parts it needs to show to visitors (content and pictures) and hides the other stuff (file operations, math calculations, etc.) then translates your PHP into HTML. After the translation into HTML, it sends the webpage to your visitor's web browser.

2.2 Advantages of PHP

PHP is an extremely popular scripting language. It was originally created in 1995 and designed for the web. It is free of charge and can be used on almost every operating system. There are over 20 million websites and over 1 million web servers running PHP and those numbers are growing every day. The reason why PHP is so popular and it is continuously growing is because it offers many advantages. These are:

- Fast - PHP was created to develop dynamic Web Pages so it is fast on websites. The PHP code is embedded in HTML and the time it takes to process and load the browser with HTML and create a full web page is very quick.
- Free - PHP is released under the PHP License. This licence is compatible with the GNU General Public License or GPL. Thus making PHP free software. This means that anybody can download it and use it 100
- Easy - The syntax of PHP is very easy to use and learn. PHP is usually mixed in with HTML and can be easily included in HTML files.

3. Materials and Methods

The unknown protein sequence (Q08392) is given as input to our algorithm. The known sequence which he used is (1ML6). Our algorithm does the Heuristic Search and generates the frequency set. The frequency set gives the pattern matches respective times. Algorithm is developed based on Hill Climbing algorithm and implemented in PHP. It is implemented in PHP script which generates the frequency set. We used Docking Tool to predict the protein structure.

3.1 SEQUENCE SEARCH HILLCLIMBING ALGORITHM (SSHM)

Step 1: Start

Step 2: Take Unknown Sequence Sk

Step 3: Take Known KSs Sequence set As a Search Space

Step 4: Apply Hill Climbing Technique to match Sk with KSs

$Sk[i,j](\text{Intersection})KSs[i,j] > Sk[i,k](\text{Intersection})KSs[i,k]$

Then $Sk[i,j](\text{Intersection})KSs[i,j]$ is the Result

Step 5: If Matches are not found with KSs Then Sk is New Sequence

Step 6: Stop

Known Protein Sequence 1ML6:

AGKPVLYHFNARGRMCEIRWLLAAAGVEFEEKFIQSPE
DLEKLLKKGDNLMFDQVPMVEIDGMKLVQTRAILNYIA
TKYDLYGKDMKERALIDMYTEGILDLETEMIGQLVLCPP
DQREAKTALAKDRITKNRYLPFAFEKVLKSHGQDYLVGN

RLTRVDVHILLELLLYVEELDASLLTPFPPLLKAFKSRISL
PNVKKFLQPGSQRKPPLDAKQIEEARKVFKF

Unknown Protein Sequence Q08392:

MSGKPVLYHYNTRGRMESVRWLLAAAGVEFEEKFLEK
KEDLQKLKSDGSLLFQQVPMVEIDGMKMOVQTRAILNY
IAGKYNLYGKDLKERALIDMYVEGLADLYELIMMNVV
QPADKKEEHLANALDKAANRYFPVFEKVLKDHGHDFL
VGNKLSRADVHLLLETILAVEESKPDALAKFPLLQSFKAR
TSNIPNIKKFLQPGSQRKPRLEEKDIPRLMAIFH

SCRIPT:

```
<TABLE border=1>
<? php
$str1="AGKPVLYHFNARGRMCEIRWLLAAAGVEFEEKFI
QSPEDLEKLLKKGDNLMFDQVPMVEIDGMKLVQTRAIL
NYIATKYDLYGKDMKERALIDMYTEGILDLETEMIGQLV
LCPPDQREAKTALAKDRITKNRYLPFAFEKVLKSHGQDY
LVGNRLTRVDVHILLELLLYVEELDASLLTPFPPLLKAFK
SRISLNPVKKFLQPGSQRKPPLDAKQIEEARKVFKF";
$str2="MSGKPVLYHYNTRGRMESVRWLLAAAGVEFE
EKFLEKKEDLQKLKSDGSLLFQQVPMVEIDGMKMOVQ
TRAILNYIAGKYNLYGKDLKERALIDMYVEGLADLYE
LIMMNVVQPADKKEEHLANALDKAANRYFPVFEKVL
KDHGHDFLVGNKLSRADVHLLLETILAVEESKPDALAK
FPLLQSFKARTSNIPNIKKFLQPGSQRKPRLEEKDIPRL
MAIFH";
? >
<TR>
<TD>String length</TD>
<TD><?php print_r(strlen($str1)); ?></TD>
<TD><?php print_r($str1); ?></TD>
</TR>
<TR>
<TD>String length</TD>
<TD><?php print_r(strlen($str2)); ?></TD>
<TD><?php print_r($str2); ?></TD>
</TR>
</table>
<table style="float:left" border=0>
<?php
$chars = array("");
$chars1 = array();
$red= array();
$c=0;
$high=0;
for($l = 0; $l<=strlen($str2); $l++)
for($k = 0; $k<=strlen($str2)-$l; $k++)
$string = substr($str2,$l,$k);
//echo substr($str2,$l,$k). $l. '->'. $k . "<br/>";
$chunk = substr($str2,$l,$k); if(strlen($chunk)>0)$cnt =
substr_count($str1,$chunk);
if($cnt>0)
if(isset($red[$string]))
$c++;
```

```

if($cnt>$high)
$high=$cnt;
$red[$string]=$cnt;
}
}
}
}
}
foreach($red as $i => $value)
?>
<tr><td><b><?php print_r($i); ?></b></td>
<TD> <?php echo $red[$i]; ?> </TD>
<?php
echo " <TD >".(($red[$i]/sizeof($red))*100)."
for($j=0;$j<$red[$i];$j++)
echo " <TD bgcolor='blue'>.</TD> ";
echo "</tr>";

?>
<tfoot></tfoot>
</table>
<?php
echo sizeof($red)."<span style='float:left'><b>Total
match count is ".$c." with highest frequency as
".$high."</b></span>";
?>

```

4. Results

The important of this study relates to the importance of dissimilar residues between any two proteins under study. In this case Q08392 and IML6 consider to prepare frequency table chart based on the designed algorithm. Owing to the importance of this analysis, a amino acid frequency chart representative of single and double amino acids are reported. The below tabulated values gives percentage matches with single and double amino acids matches.

From the above tables frequency of single amino acid residue between Q08392 and IML6 reported to contain 51.66% number of matches. The amino acid with highest frequency was found to be "Leucine"(L) with 32 number of matches (7.58%).Most of the percentage matches were in the range 1.42% to 2.6%.The basic amino acid "Histidine"(H) was at 0.71% .Considering the two amino acids matches ,most of the observes matches between these two proteins are not more than 2 to 3 matches with percentage number of matches been 0.5%. The overall success rate with single amino acid and double amino acid was 51.66% and 21.92% respectively.

5. Conclusions

The importance of this coding and subsequently results from this study is to emphasis the crucial amino acid residues responsible for functional attributes can be detected used molecular docking technology. In other words the presence

Table 1: single and double amino acids matches with percentages

Amino Acid	No. of matches	% of matches
M	7	1.66
S	7	1.66
G	11	2.61
K	22	5.21
P	12	2.84
V	13	3.08
L	32	7.58
H	3	0.71
Y	8	1.91
A	16	3.79
N	6	1.42
T	8	1.90
R	13	3.08
E	17	4.03
F	10	3.77
D	14	3.72
Q	9	2.13
I	10	2.13
Success rate		51.66

of either single or double amino acids with in or nearer active site region leads to gain insights towards the functional relevance of Glutathione S transferase (GST).Hence a homology modeling protein was undertaken to build the protein and subsequently molecular docking studies are initiated.

References

- [1] Aebersold R.H., Leavitt J., Saavedra R.A., Hood L.E., and Kent S.B. "Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose., Proc. Natl. Acad. Sci. Vol.84: pp.6970-6974, 1987.
- [2] Bergman T. and Jörmvall H., "Electroblotting of individual polypeptides from SDS/polyacrylamide gels for direct sequence analysis.", European Journal of Biochem. Vol.169, pp 9-12, 1987.
- [3] online Manual on PHP available online, <http://php.net/manual/en/faq.general.php>

SESSION

COMPUTATIONAL METHODS FOR FILTERING, NOISE CANCELLATION, AND SIGNAL + IMAGE PROCESSING

Chair(s)

TBA

Quantifying Phenotypic Traits in Retinal Coronary Angiography: Automated Extraction of Retinal Vascular Networks and Localization of Optic Discs in Fundus Images

Hesam Dashti*, James Driver*[§], Nader Sheibani^{†‡}, Amir Assadi*^{¶||}

[‡]Eye Research Institute, University of Wisconsin, USA

[§]Department of Physics, University of Wisconsin, USA

*Department of Mathematics, University of Wisconsin, USA

^{||}The Genome Center of Wisconsin, University of Wisconsin, USA

[¶]Comparative Biological and Medical Sciences, University of Wisconsin, USA

[†]Department of Ophthalmology and Visual Science, University of Wisconsin, USA

Abstract—Numerous retinopathies are related to the dysfunction of retinal vasculature, especially micro-vessels. Extensive research in ophthalmology has singled out critical roles of vascular morphology, and the functional dynamics of blood flow in diseases. Advances in angiography has yielded a myriad of applications for computational methods that design efficient tools to complement retinal imaging and microscopy in analytic ophthalmology. In this paper, we propose a novel mathematical approach for the design of quantitative tools that enable researchers, as well as automated vision-based systems, to perform pattern recognition, and feature extraction in retinal vasculature. The present feasibility-stage implementation of these new algorithms demonstrates the power and versatility of the set of tools we provide for the detection of morphological pathology, as well as the theoretical study of retinal neurovasculature anatomy when regarded as a complex (dynamic) system. In contrast to current state-of-the-art methods that rely on bottom-up algorithms to deal with noise and trace the vessels, we propose a top-down scheme to overcome noise and capture morphological features such as center-lines, radii, and the edge locations of circulatory blood vessels. This approach is comprised of three components. First, the algorithms for detection and measurement of the vasculature morphological structures in two-dimensional fundus images are implemented. These algorithms combine advanced kernel-based methods to extract blood vessels, and are further enhanced by variants of Canny Edge Detection algorithms. Second, a fully automated approach is provided to identify the optic disc in healthy/diseased fundus images, eliminating current bottle-necks requiring extensive human expertise. Third, we construct a hierarchical network of geometric (topological) structures of the extracted vessels, rooted in the optic disc. A notable application of our methods is to capture complex vasculature structures in noisy, blurred, and light-reflecting fundus images. Another advantage of our approach is the automation of *in vivo* quantification of complex phenotypic traits of retinal neurovasculature, which are expected to play an important role in emerging computational models for mapping genotype-phenotype relations and personalized medicine.

I. INTRODUCTION

Analysis and quantifying medical images forms an essential step in delineating practical issues in relation to the diagnoses

of systems. Extracting appropriate features to represent the content and structure of an image by precisely capturing anatomical and pathological features of the retinal tissue is the goal of quantifying fundus images. Segmentation of blood vessels and quantifying phenotypic traits, such as width, length, and distinguishing between regions of lesions, plays an important role in the diagnosis of vasculitis, malformations, vein occlusion [12], exudates, diabetes [19], glaucoma [13], and many other retinal diseases exhibiting a vascular phenotype.

Currently, almost every medical imaging technique (ultrasound, X-ray, MRI, CT, etc...) can be used to capture high resolution, two- or three-dimensional, blood vessel images. However, the complexity of the vasculature structure, unavoidable noise in the system, and faded images, challenges scientists to come up with precise, efficient, and practical approaches.

Decades of intensive research have brought a vast array of tools and methods. Comprehensive reviews and comparison of many of these accomplishments are mentioned [2], [3]. Li Wang [3] recently proposed a multi-resolution, Hermite polynomial-based model to analyze two-dimensional images, and construct a tree-type data structure of the blood vessels.

Using fuzzy methods is currently in vogue due to its ability to achieve noise removal and ease of enhancement in combination with other probabilistic methods [4], [6]. Statistical and kernel-based methods have also been proposed to overcome uncertainty in images [7], [9]. Additionally, template matching approaches have been examined [8], [10], [11]. Nonetheless, greater advancements are needed to handle unaddressed patterns of noise, reflection of light, and complex structural arrangements in images. Extensive variation in vessel width (especially in the case of arterial stenosis and aneurysms) remains an obstacle for quantifying phenotypic traits. Experts tend to have subjective variability in their identification of subtle features, creating an urgency for the ability of automated

methods to quantify phenotypic traits. Our own research [1], high throughput *in vivo* phenotyping, is needed to collect the time-series that encodes the dynamic variation of morphologies, and can predict the onset of angiogenesis in diabetic and other high-risk patients.

In this paper, we propose a new top-down method in which a kernel-based method was used to project the images to a higher-dimensional space. Using this projection, we avoid dealing with lesions, various types of noise, and reflected light. Thereafter, we applied a local to global model to extract the edges of the vessels and their bifurcated segments. A Canny-type algorithm was applied to fill the gaps along the longitudinal vessel, while closeness and varying widths of vessels were considered. Through application of the Canny based edge detection algorithm, we constructed a multi-resolution topological structure from the vessels. Whereas the blood vessels originate from the optic disc [5], we use this topological structure to identify it. In addition to the geometrical correlation between blood vessels and the optic disc, the impact of the density of blood vessels to measure the size of the optic disc and fovea has already been demonstrated [19].

In the next section, we describe our methodology and algorithms in detail. In the *Identifying the Optic Disc* section, we evaluate our algorithms by comparing them with a current standard.

II. METHODOLOGY

Measuring morphological traits of retinal blood vessels plays an important role in the screening of numerous ocular diseases. The mysterious structure of the retinal blood vessels has motivated researchers to study this topic from a computational point of view. The identification of fractals as the mathematical structure underlying vasculature has opened new branches of research [9], [10]. Utilizing fractals as a data structure for storing vessels, to study their distribution, has been examined and yielded promising results [11], [12], [13]. In this study, the method for storing vessels and finding the density of their distribution (i.e. locating the optic disc) is inspired by their fractal structure. The first part of this section explains the algorithm for extracting the vessels and generating the hierarchal structure rooted in the optic disc. The second illustrates the ability of the hierarchical structure to identify the optic disc.

A. Vasculature Structure

Using statistical learning and kernel machines in data mining is a well known approach. Vapnik introduced a new branch of data clustering approaches; whereby applying kernel machines, complex data structures could be clustered [16], [17]; 'Foundation of Analysis'. In this work we used kernel methods to project complex data objects to higher-dimensional spaces in order to efficiently distinguish meaningful information (pixels) from noise [14], [15].

1) *Kernel Mapping*: A preliminary step towards extracting the vasculature structures from the fundus images is to increase their respective pixel contrast. Mapping the images, via a kernel, allows us to individually project the color intensities of the images' respective pixels to higher and lower levels of saturation, providing greater separation of pixel values effectively "sharpening" the images. Figure 1 compares an original and kernel-mapped fundus image.

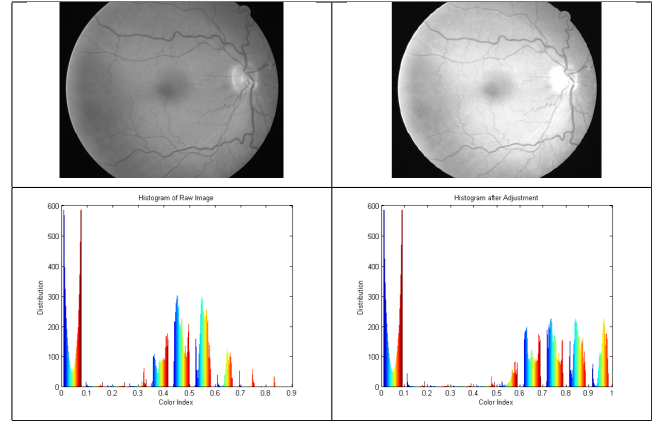


Fig. 1. The first column shows the original fundus image with its associated color intensity histogram. The second column shows the image and its histogram after adjusting the color indexes. As shown in the histograms, the range of color indexes (x-axis) and their intensities (y-axis) provides a quantifiable difference between the original and adjusted image.

2) *Canny Edge Detection*: Canny Edge Detection involves pre-filtering the image through convolution with a simple Gaussian filter to eliminate noise that might otherwise interfere with the edge detection process. Selection of a small, versus a large, filter window directly affects the observable and statistical smoothing applied to the image, and helps to reduce unavoidable noise from image acquisition. After smoothing, standard kernels G_x and G_y are applied in both the x and y directions of the image to determine edges by calculation of the image's gradient $|G|$.

$$G_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}$$

$$G_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$$

$$|G| = \sqrt{G_x^2 + G_y^2}$$

$$\theta = \arctan\left(\frac{|G_y|}{|G_x|}\right)$$

The detection of the edges after applying the 3 x 3 kernels to each pixel is determined by the angle θ and stored for comparison to determine the "strong" versus "weak" edges of the image, as specified by a double threshold intrinsic to the image based on maximum and minimum pixel values.

For example:

- 1) Round the θ value to the nearest multiple of 45 degrees, corresponding to the directional choices for the pixel's eight adjacent pixels. (0 = right, 45 degrees = upper-right, 90 degrees = upper-center, 135 = upper-left, 180 = left, etc. . .)
- 2) Compare the gradient value for each pixel based on the positive and negative θ value to obtain the next piece of the edge based on the gradient thresholding values.
- 3) If the pixel under examination, relative to its eight adjacent pixels, is largest within this threshold it is preserved as a "strong" edge. If it is "weak", as long as it is connected to a "strong" edge it is preserved. If it is neither of these, it is marked for removal.

The Canny algorithm applies a double threshold to label edges corresponding to "strong" and "weak", by referencing the value of the gradient as described in (3). It is these thresholds which ultimately determine edges detected as "strong" (i.e. pixels in the neighborhood described in (1) referring to the pixel of interest's gradient value) or "weak". Figure 2 shows the results of the edge detection algorithm with different threshold values.

3) *Dilation*: Dilation is a set operation performed over a discrete neighborhood of size n . The structuring element can be thought of as a geometric shape that overlaps and extracts the maximum pixel value lying within its boundaries; performing this operation iteratively pixel by pixel and replacing the pixel of interest with the maximum pixel value within the neighborhood.

Dilating the detected edges by $n = 1$ extends them towards filling the vessels and segmenting the vasculature structures from surrounding regions. Figure 3 shows an example of dilated edges in comparison with the original image.

B. Identifying the Optic Disc

Despite the optic disc being located in the observably blind region of the eye, known as "the blind spot", studying discs is important for diagnosing vascular disorders. The optic disc is also the gateway between the nervous and visual systems [18]. Since all blood vessels are directed towards the optic disc, extracting the geometrical distribution of the vessels is the first step towards the optic disc's location [5]. Differences in color indexes of an optic disc, relative to its surrounding regions in fundus images, has motivated researchers to develop color-index image analysis tools. However, since there is a distinct similarity between color indexes of the optic disc and exudative lesions, these color-index image analysis tools are not appropriate for quantifying affected images. In this study, we used a geometric-based algorithm to define a feasible region for the location of the optic disc and to exclude exudate regions from it. Moreover, we enhanced this algorithm, where appropriate, with image analysis tools to improve its accuracy.

Figure 4 shows the color indexes of the two-dimensional images. These graphs show the differences between the color distributions in the optic discs relative to their surrounding areas. Notice that the gradient of the surface around the optic

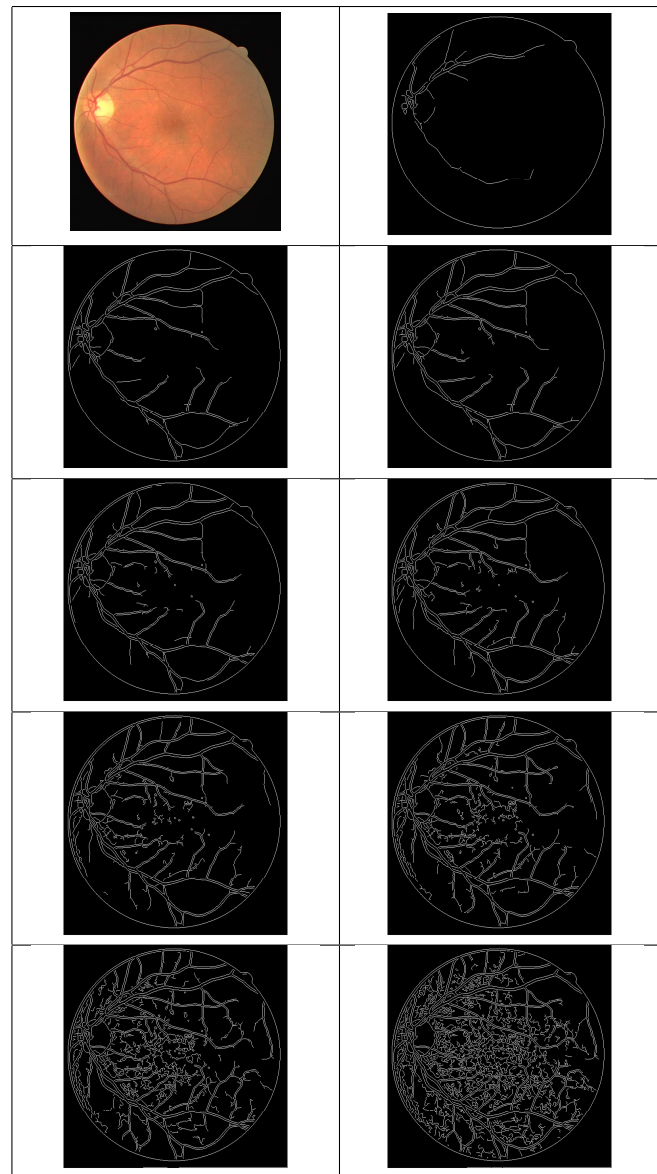


Fig. 2. Multi-resolution of vessel edges: the Canny Edge Detection algorithm with different threshold values. The threshold values decrease generally, and with respect to each other (i.e. upper versus lower threshold), left to right, and top to bottom.

disc is zero. To further emphasize the importance of analyzing color indexes, consider the contour plots of these images in Figure 5.

Previously, some healthy fundus images were considered. Next, let us consider some pathological subjects. Figure 6 shows four examples of affected fundus images and their respective color intensity surfaces. As depicted by the color intensity surfaces (second column), identifying the optic disc based on the analysis of color indexes is insufficient. However, analyzing the multi-resolution structure of the blood vessels provides satisfactory results. The results of the analysis of the multi-resolution structures are shown in the first column of Figure 7, where in the second column one can observe a strong correlation between these structures and the location of

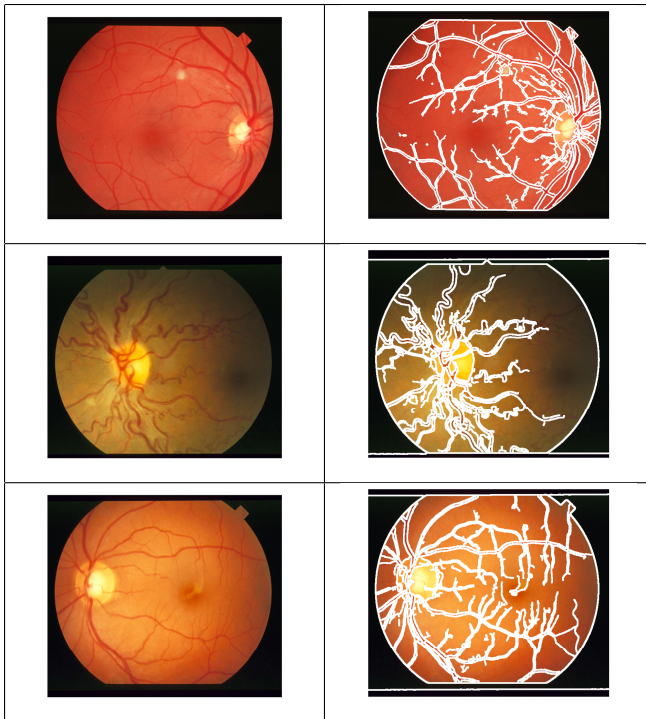


Fig. 3. The detected edges were dilated pixel by pixel over a neighborhood of radius $n = 1$.

the optic disc.

To evaluate our methodology, we applied these algorithms to two datasets: STructured Analysis of the Retina (STARE) [20], and Digital Retinal Images for Vessel Extraction (DRIVE) [21]. These datasets included fundus images of both healthy and pathological subjects. We compared the results of our algorithm with the manually extracted vessels provided by Adam Hoover [22]. As mentioned previously, instead of tracking vessels in our algorithm, the automated extraction and measurement of morphological traits is emphasized.

A direct comparison of automated and segmented vasculature structures to manually extracted vessels is shown in [23]. Moreover, some of the fundus images show different levels of hierarchical vasculature structure. To evaluate the second part of our methodology, we compared it to two well-known applications for identifying optic discs [5], [6]. The results from the application of our algorithm on the two datasets are available at [23].

CONCLUSION

In this paper, we have presented a novel computational method to quantify retinal blood vessels and identify optic discs in two-dimensional fundus images. This methodology consists of a kernel-based algorithm to extract vasculature structures. Taking advantage of the Canny Edge Detection algorithm, our methodology constructs a hierarchical structure of the blood vessels. This algorithm accurately quantifies vessel structure, length, and can capture dynamics in width by precisely detecting the edges of vessels. Moreover, analyzing

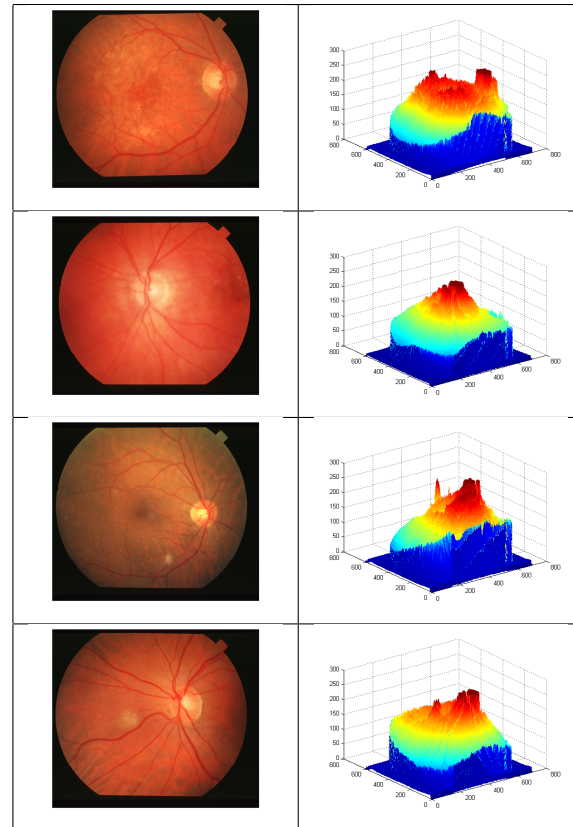


Fig. 4. The x and y-axis form the xy-plane of the image. The z-axis shows the color indexes of the image corresponding to its xy-coordinate. The optic disc is the flat region on the surface and differentiation of the color densities in these regions are zero in contrast with other parts of the image. This is an important feature to identify the optic disc.

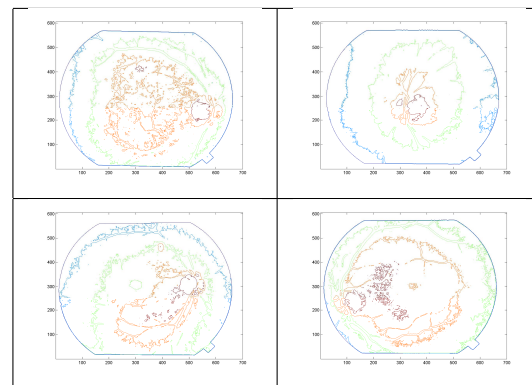


Fig. 5. These are the contour plots of the fundus images, left to right and top to bottom correspond to the images in Figure 4. The optic disc regions have the highest color indexes and are indicated by dark red. Other structures have lower color indexes, indicated by their colder colors relative to the optic disc.

this hierarchical structure, and appropriately using standard image analysis tools, we can identify the optic disc. In this manner, we have used two important biological features of angiography to detect the optic disc: the intrinsic geometry of the optic disc with respect to the vessels; and the differences of color indexes of the optic disc relative to its surrounding regions. This methodology precisely distinguishes between

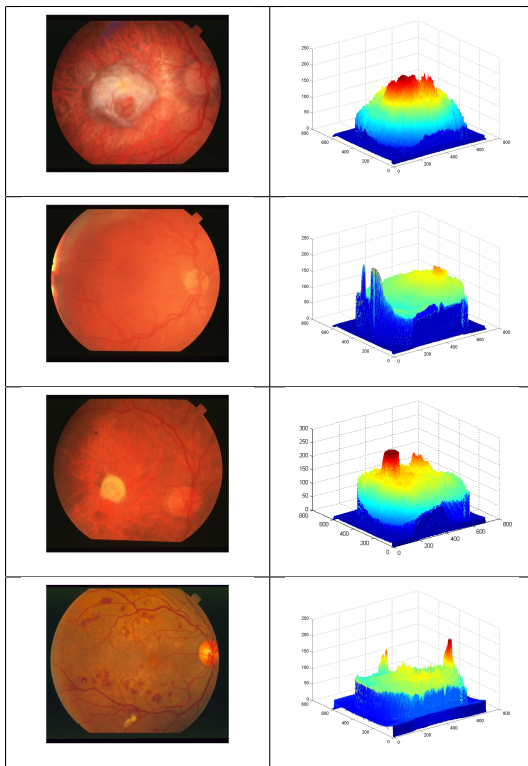


Fig. 6. These images are examples of affected retinal fundus angiography from the STARE datasets.

closely located vessels and their longitudinal gaps. It also separates the lesion regions from the optic disc. We evaluated our methodology on two datasets, where the results showed its robustness and accuracy on normal and pathological retinal fundus images. Efficiency and precision of the algorithms are demonstrated by comparing our results with manually segmented fundus images [22] and comparing them with current well-known algorithms for identifying optic discs [5], [6].

ACKNOWLEDGEMENTS

This work was supported by NSF-DMS SCREMS grants (Amir Assadi), grants EY16995 (Nader Sheibani), EY18179 (Nader Sheibani), EY21357 (Nader Sheibani and Amir Assadi), and P30-EY16665 from the National Institutes of Health and an unrestricted departmental award from Research to Prevent Blindness.

Nader Sheibani is a recipient of a Research Award from the American Diabetes Association, 1-10-BS-160, and Retina Research Foundation.

REFERENCES

- [1] Nader Sheibani, Amir Assadi Integrated Multidisciplinary Strategies for Detection of Diabetic Retinopathies. *The NIH Directors Award RC4EY021357-01*, 2010
- [2] Cemil Kirbas and Francis Quek A review of vessel extraction techniques and algorithms. *ACM Comput. Surv.*, 81–121., 2004
- [3] Li Wang, Bhalerao A, Wilson R Analysis of Retinal Vasculature Using a Multiresolution Hermite Model. *IEEE Trans Med Imaging*, 26(2):137–52, 2007

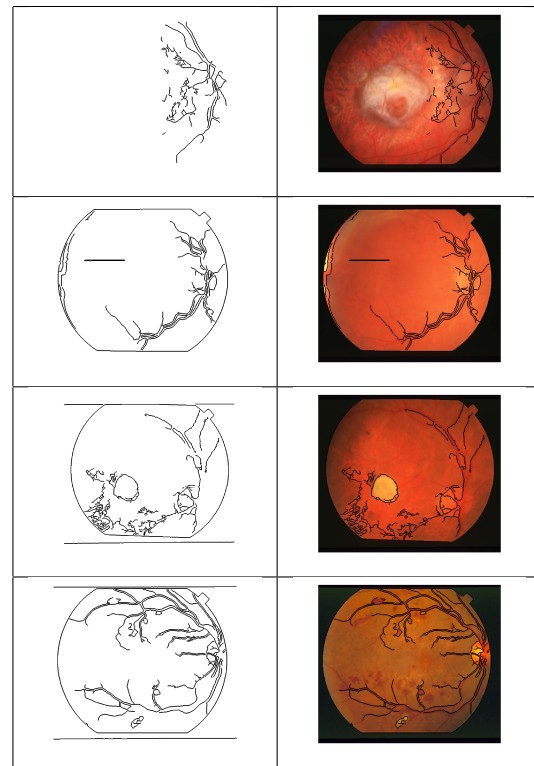


Fig. 7. These images are examples of affected retinal fundus angiography from the STARE datasets. The third image illustrates the necessity of using the geometrical and color-intensity methods together. In Figure 6, the surface of this image has two peaks, one corresponding to the lesion and the other to the optic disc. The intensity peak of the lesion is much higher than that of the optic disc. However, the density of the vasculature structure surrounding the optic disc is much greater than its density surrounding the lesion. Observably, there are no blood vessels inside the lesion. Hence, through the combination of these two observables, it is possible to accurately locate the optic disc.

- [4] Zhou Shoujun, Yang Jian, Wang Yongtian, Chen Wufan Automatic segmentation of coronary angiograms based on fuzzy inferring and probabilistic tracking. *Biomed Eng Online*, 20, 9:40., 2010
- [5] Ruggeri A., Forrachia M., Grisan E. Detecting the optic disc in retinal images by means of a geometrical model of vessel network. *Engineering in Medicine and Biology Society*, 17(1):902-905, 2003
- [6] Adam Hoover, Michael Goldbaum Locating the Optic Nerve in a Retinal Image Using the Fuzzy Convergence of the Blood Vessels. *IEEE Engineering in Medicine*, 22(8):951-958, 2003
- [7] Mouloud Adel, Aicha Moussaoui, Monique Rasigni, Salah Bourennane, and Latifa Hamami Statistical-Based Tracking Technique for Linear Structures Detection: Application to Vessel Segmentation in Medical Images. *IEEE Signal Processing Letters*, 17(6):555-558, 2010
- [8] Christine Toumoulin, Jorge Brieva, Jean-Jacques Bellanger, and Huazhong Shu String Matching Techniques for High-Level Primitive Formation in 2-D Vascular Imaging. *IEEE Trans Inf Technol Biomed*, 7(4):291-301, 2003
- [9] Lim SW, Cheung N, Wang JJ, Donaghue KC, Liew G, Islam FM, Jenkins AJ, Wong TY Retinal vascular fractal dimension and risk of early diabetic retinopathy: A prospective study of children and adolescents with type 1 diabetes. *Diabetes Care*, 32(11):2081-3, 2009
- [10] Barry R. Masters Fractal Analysis of the Vascular Tree in the Human Retina. *Annual Review of Biomedical Engineering*, 6: 427-452, 2004
- [11] Azemin M.Z.C., Kumar D.K., Wong T.Y., Kawasaki R, Mitchell P, Wang J.J. Robust Methodology for Fractal Analysis of the Retinal Vasculature. *Diabetes Care*, 30(2):243-250, 2011
- [12] Cheung N, Liew G, Lindley RI, Liu EY, Wang JJ, Hand P, Baker M, Mitchell P, Wong TY Retinal fractals and acute lacunar stroke. *Ann Neurol*, 68(1):107-11., 2010
- [13] Huajun Ying, Ming Zhang, Jyh-Charn Liu Fractal-based Automatic

- Localization and Segmentation of Optic Disc in Retinal Images. *IEEE EMBS*, 4139-41., 2007
- [14] Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. *IEEE Trans Neural Netw.*, 10(5):1055-64, 1999
- [15] Burges Christopher J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.*, 2(2):121-167, 1998
- [16] Vladimir N. Vapnik Statistical Learning Theory. *Wiley-Interscience*, 736 pages, 1998
- [17] Bernhard Schölkopf, Christopher J. C. Burges, Alexander J. Smola Advances in Kernel Methods: Support Vector Learning. *The MIT Press*, 386 pages, 1998
- [18] Aage R. Moller Sensory Systems: Anatomy and Physiology. *Academic Press*, 469 pages, 2002
- [19] Pinz A., Bernogger S., Datlinger P., Kruger A. Mapping the human retina. *Engineering in Medicine and Biology Society*, 17(4):606-619 , 2002
- [20] STARE dataset. Available:<http://www.ces.clemson.edu/~ahoover/stare/>.
- [21] Staal J.J., Abramoff M.D., Niemeijer M., Viergever M.A., Ginneken B. Ridge based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23:501-509, 2004
- [22] Hoover A, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans Med Imaging*, 19(3):203-10, 2010
- [23] Persepolis Computation Group. Available:<http://vv811a.math.wisc.edu/>.

Common-Path Fourier-Domain Optical Coherence Tomography using Surface Tracking Algorithm

C. G. Song¹, K. S. Kim¹, M. H. Kim¹, S. H. Ryu¹, J. H. Seo²

¹ Div. of Electronics Eng., Chonbuk Natl. Univ., Korea

² Dept. of Rehabilitation Med., Chonbuk Natl. Univ. Hospital, Korea

Abstract - Conventional optical coherence tomography (OCT) systems generally have a limited imaging range within a depth of only 1-2 mm and suffers from unwanted noise such as speckle, ghost or mirror noise. To overcome these limitations, we developed a motorized-stage-based OCT system with an extended imaging range, using a common-path Fourier-domain optical coherence tomography (CP-FD-OCT) configuration. Using this OCT systems, OCT image was obtained from an onion, and their subsurface structure were observed. The result showed that the OCT images obtained using our motorized-stage-based system had a significantly extended imaging range due to its real-time accurate depth tracking. Consequently, the devised CP-FD-OCT systems and algorithms have good potential for the further development of endoscopic OCT for microsurgery.

Keywords: OCT, Common-path OCT, FD-OCT, Tracking algorithm

1 Introduction

Optical coherence tomography (OCT), which is one of the various optical imaging modalities, is a novel imaging technology that provides high-resolution, subsurface depth profiling, and cross-sectional imaging in vivo with relatively simple optical arrangements and an inexpensive light source in a non-invasive manner. The concept of OCT and its application were first introduced by Fujimoto et al. in 1991 [1]. OCT imaging is somewhat analogous to the B-scan imaging technique based on ultrasound, except that it uses light instead of sound.

This technique typically makes use of a Michelson interferometer and allows a depth-profile to be obtained by measuring the optical pathlength or phase difference between the reflected or backscattered light beam and a reference one, when near infra-red light (wavelength: 600-

1,300 nm) is illuminated onto the sample. First, the light generated from the light source is divided into two arms, viz. the sample arm used for exploring the sample and the reference arm which is usually obtained using a mirror, via a beam splitter or optical coupler. Next, the combination of the reflected or backscattered light from the sample arm and reference light gives rise to an interference pattern. The more light that is reflected back from the reflective layer within the sample, the greater the intensity of the interference fringe. This reflectivity profile (A-scan) contains information about the spatial dimensions and location of the structures within the sample of interest and, thus, OCT images can be obtained by measuring the optical delay according to the depth within the sample, i.e. different reflective loci at different depths in the sample. Finally, a cross-sectional image (B-scan) may be achieved by laterally combining a series of these axial depth profiles.

The OCT technique has several benefits for the non-invasive, high-resolution and fast-acquisition tomography of the internal microstructure in biological systems and materials. First of all, it can provide much higher-resolution images (2-10 μm) than conventional imaging techniques, such as ultrasound (over 500 μm), MRI and CT (over 100 μm), although its depth information is limited to a range of approximately 2-3 mm in turbid tissue [2]. Also, OCT has a faster scanning speed for acquisition and relatively wider dynamic range [3]. OCT can image with an acquisition rate of up to 20 frames per second, which should allow this technology to image surgical procedures in near real time. Moreover, the entire system is simple and portable and, thus, has the potential to enable OCT catheters to be incorporated into endoscopic instruments or bedside devices. Finally, since OCT is based on optics, it can be combined with other spectroscopic techniques to assess the optical and biochemical aspects of the tissue being imaged.

Common-path OCT (CP-OCT) was proposed by Vakhtin et al. in 2003 [4]. In the CP-OCT configuration, the beam paths, which the sample signals backscattered from the sample and reference signals reflected from the reference plane follow, are commonly shared, thereby

This work was supported by the second stage of Brain Korea 21 Project in 2011, the National Research Foundation of Korea Grant funded by the Korean Government (MEST) (NRF-2010-0021864) and Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government Ministry of Knowledge Economy (No. 20104010100660)

eliminating the need for the reference arm in the interferometer. This modality can minimize the effect caused by the mismatch of the polarization and dispersion states between the optical elements in the interferometer and the sensitivity to vibration, and enhance the scanning speed, simplicity and system robustness. Consequently, this configuration has the potential to be used as a microsurgical tool. Some researchers have reported the feasibility of an endoscopic CP-OCT implementation based on the common-path modality [5-7].

However, unfortunately, most OCT systems generally suffer from a limited imaging depth range of only 1-3 mm, depending on the tissue type and, thus, this limitation restricts their clinical applications when the sample's topological variance is larger than the imaging depth range [8]. To overcome these limitations, some techniques such as the adaptive ranging technique based on depth tracking have been proposed in previous papers [9-10]. In these methods, the coherence gate offset and range on the reference arm are adaptively adjusted by means of an active tracker consisting of various optical lenses and a galvanometer. Also, alternative techniques using an auxiliary spectral domain partial coherence interferometer [11], tunable endoscopic MEMS probe consisting of a pneumatically-actuated micro-lens and a GRIN lens [12], or dual reference arms and a high-speed fiber optic switch [13] were attempted. However, these techniques require the supplementary alignment of the various optical lenses or components and synchronization control and, thus, the composition and control procedure of the OCT system might become more onerous and complicated. Also, they compensate for the topological variance and motion by adjusting the optical pathlength on the reference arm and, therefore, this strategy might be inappropriate for the CP-FD-OCT system constructed in this study, since CP-OCT uses the common beam path of the sample and reference signal instead of using the reference mirror used in the conventional OCT composition. Recently, Zhang et al. [14] reported a CP-FD-OCT system providing a surface

topology and motion compensation technique in the axial direction by means of a 1-D erosion-based edge-searching algorithm, which makes use of the relatively simple signal processing of the A-scan data instead of the alignment of complex optical components.

To assess the feasibility of the system described in this paper, an active compensation algorithm of the topological variance by means of a sample surface detection algorithm using a Savitzky-Golay smoothing filter and feedback control for adjusting continuously the position of the motorized stage was developed in the present study. This algorithm makes it possible to image a deeper range along the z-axis by keeping the distance between the end of the probe and the sample's surface constant, as compared to the conventional scanning strategies.

2 FD-CP-OCT system

2.1 Hardware Configuration

To obtain high-resolution OCT images with an extended range of imaging depths, a motorized-stage-based OCT system was developed. It consists of a high-resolution spectrometer, actively controllable motorized-stage, actuators and control modules, as well as basic compositions such as a light source, 50/50 coupler, and single mode fiber-optic probe. Figure 1 shows the block diagram of the developed CP-FD-OCT system. A superluminescent diode (SLD) (SLD-351, Superlum Diode Ltd., Ireland) with a central wavelength of 860 nm and spectral full-width at half maximum (FWHM) of ~60 nm was used as the light source. A 50/50 coupler (FC850-40-50-APC, Thorlabs Inc., U.S.) was used as the beam splitter, and only one branch on the right side was used as the common path for the signal and reference. The single mode fiber-optic probe constructed in this study was fixed on a standing vise, with A-scan (z-axis) and B-scan (x-axis). The

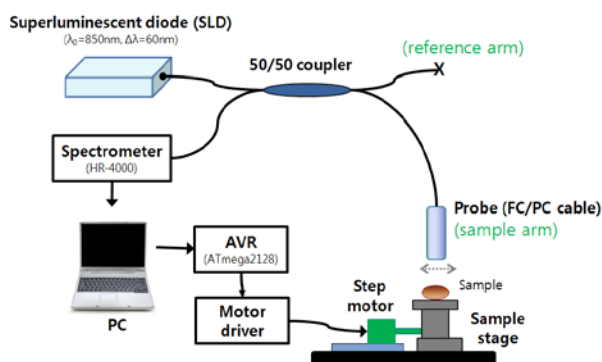


Fig 1. Block diagram of the developed CP-FD-OCT system.

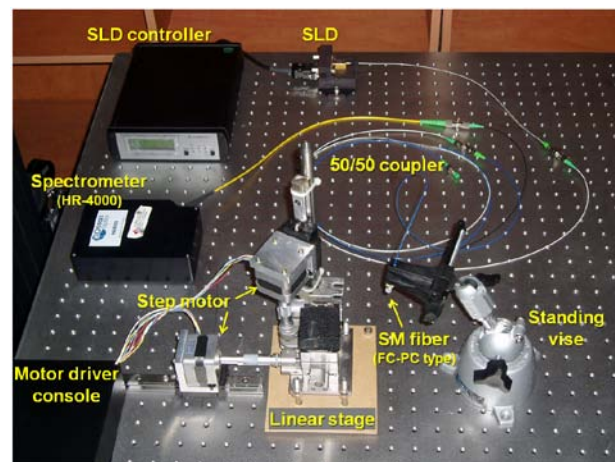


Fig 2. Photograph of the developed CP-FD-OCT system.

two axes of the x and z directions were driven by a motorized stage (M-561D-XYZ, Newport Corp., U.S.) with two separate step motors (SE-SM243, N.T.C., Korea) installed on its lateral side. The reference signal came from the Fresnel reflection at the fiber probe end and the sample signal and the reference were received by a high-speed spectrometer (HR-4000, Ocean Optics, U.S.) with a charge-coupled device detector array with 3648 pixels covering a range of 700-900 nm. This system make it possible to extend the imaging range, since the position of the probe can be adjusted actively and simultaneously according to the sample's topological variance, whereas the time needed for image acquisition is relatively longer. Figure 2 shows the photograph of the developed system.

2.2 Software configuration

To control the actuators and perform image processing in our motorized-stage-based CP-FD-OCT system, OCT acquisition software was developed using the LabVIEW language (ver. 8.6, National Instruments, U.S.) based on a graphic user interface (GUI) with buttons and graphs.

Figure 3 shows the developed software. The 'New Reference' button is used for measuring the reference signal in the CP-OCT configuration. The 'B-scan' button is used for obtaining the B-scan OCT image, while 'Stop' is used for holding the B-scanning. The 'Reset' button is used for initializing variables such as the rotation speed and period of the actuator per step, B-scanning range, and set-up distance between the probe and sample's surface and so on. Also, the 'Save' and 'Load' buttons are used for storing the measured OCT image on the host PC and loading the OCT image file from the PC, respectively. 'Exit' is used to terminate the software.

The lateral resolution, which means the lateral displacement per A-scan data, and total B-scan range are given by inputting the information into the combo ('Set step degree') and control boxes ('Distance') on the top left of the

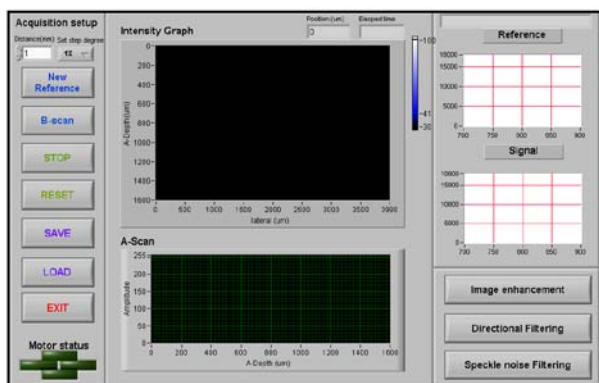


Fig 3. Photograph of the developed OCT acquisition software.

screen, respectively. Also, the position of the stage can be manually adjusted by repeatedly entering the keyboard keys, 'Δ' (upward), '▽' (downward), '▷' (forward), or '◁' (backward), and then the moving direction of the stage is instantly indicated by the corresponding LED at the bottom left of the screen.

The instant spectral distribution and A-scan data measured during B-scanning are simultaneously displayed in the corresponding graphs of 'Signal' in the middle right and 'A-scan' in the bottom center, respectively. At the end of B-scanning, the final 2D OCT image is displayed on the 'Intensity Graph' in the top center.

2.3 Active surface tracking algorithm

Figure 4 shows a flow chart of the active topological variance compensation algorithm during B-mode scanning in CP-FD-OCT, while the distance from the sample's surface exceeds the OCT imaging depth range or when the probe is too close to the sample.

In 'Step-1', the A-scan data, $a(z)$ is obtained from the probe (N is the total length of $a(z)$, as shown in Figure 5(a).

In 'Step-2', the distance (Dist) between the end of the probe and the sample's surface is determined, as follows; i) $a(z)$ is smoothed by a 3rd-order Savitzky-Golay filter (its

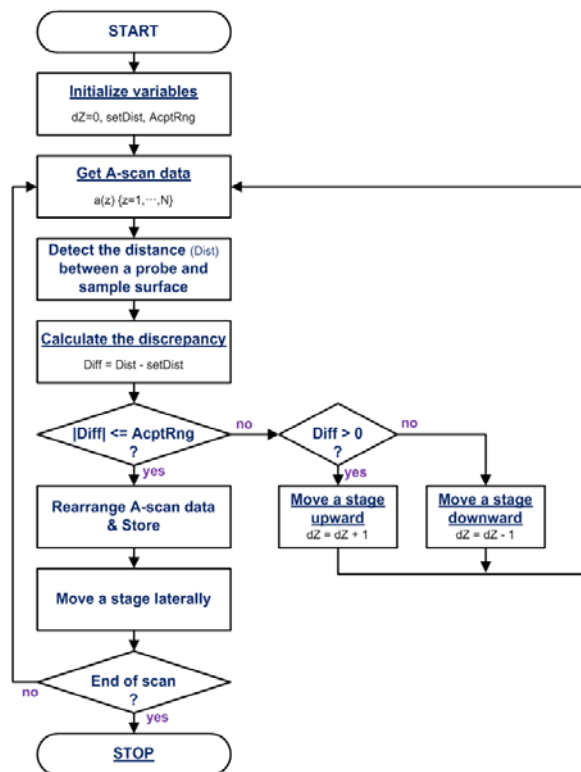


Fig 4. Flow chart for active surface tracking algorithm.

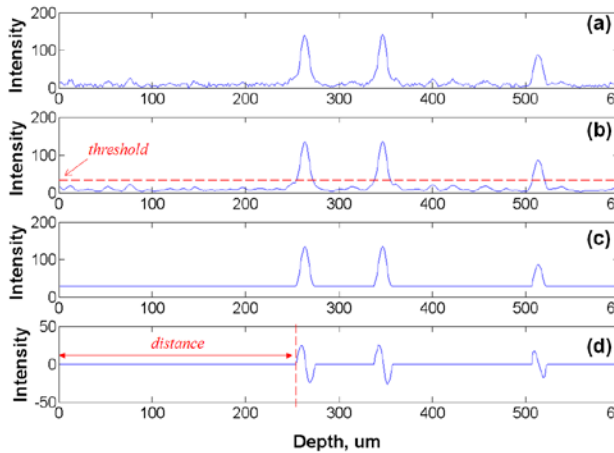


Fig 5. Active surface tracking algorithm. (a) Raw A-scan data, (b) A-scan data after Savitzky-Golay smoothing filter, (c) Thresholding of A-scan, and (d) First increment point detection for edge location.

window length is 9), as shown in Figure 5(b). The main advantage of the Savitzky-Golay filter used in this algorithm is that it can preserve the unique features of the distribution, such as the relative maxima, minima and width, which are usually flattened by other adjacent averaging techniques, such as a moving average or low-pass filter, as well as effectively reducing the unnecessary speckle noise [15], as shown in Figure 6. This attribute is quite useful for the more accurate detection of the edges from the A-scan data and, thus, over- or under-estimation of the distance can be effectively diminished compared to the other smoothing methods. ii) the smoothed A-scan data, $a_{sm}(z)$, is processed using a certain threshold level ($thre$) to avoid the noise effect, as follows (Figure 5(c));

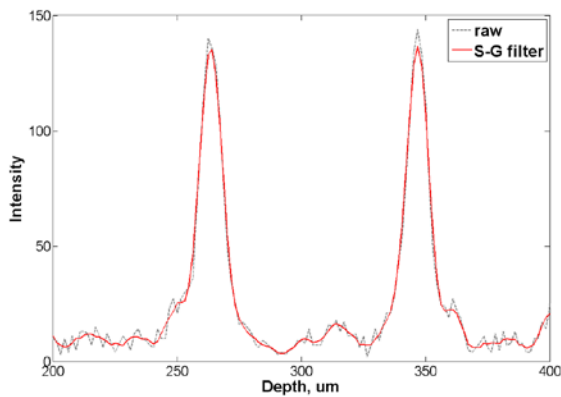
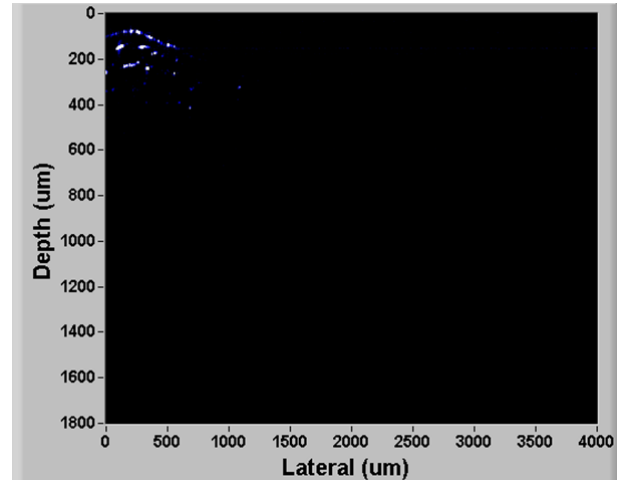
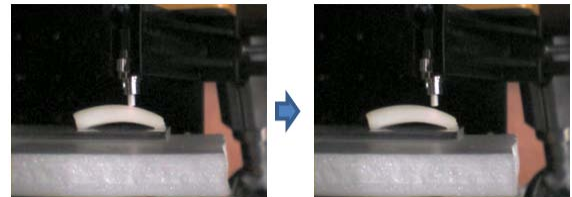


Fig 6. Comparison the raw A-scan data (dotted line) and smoothed one after Savitzky-Golay filter (solid line).



(a) OCT image



(b) Probe position at the start (left) and end times (right) of the lateral scan

Fig 7. Image of an onion sample obtained by the conventional static stage on the z-axis with limited imaging depth.

$$a_{thre} = \begin{cases} a_{sm}(z), & a_{sm}(z) > thre \\ thre, & others \end{cases} \quad (1)$$

iii) $Dist$ is given by the first increment point of the differential of the post-thresholding data, as shown in Figure 5(d).

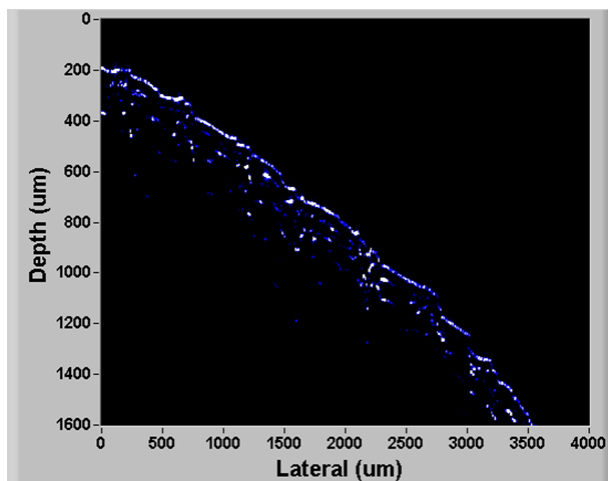
In 'Step-3', the discrepancy ($Diff$) between the preset ($setDist$) and measured ($Dist$) distances is calculated, as follows;

$$Diff = Dist - setDist \quad (2)$$

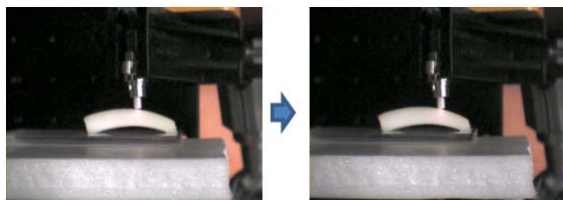
In 'Step-4', by using the $Diff$ value obtained in 'Step-3', the control system sends the feedback control signal to the motorized stage. If the absolute value of $Diff$ is outside of the preset acceptable range ($AcptRng$), the stage is moved either upward for a positive value of $Diff$ or downward for a negative value of $Diff$. Subsequently, 'Step-1' is performed again until $Diff$ is within $AcptRng$. On the other hand, if it is within $AcptRng$, the measured $a(z)$ is rearranged and stored in memory. During recording, the values of $a(z)$ are repeatedly obtained while maintaining a constant distance

between the end of the probe and the sample's surface and, thus, this $a(z)$ can be rearranged by considering the practically moved height of the stage. The variable, dZ , is used for counting the relative displacement at the current position compared to that at the start of the B-scan ($dZ=0$) on the z -axis. For example, a positive value of dZ indicates that the probe has moved closer to the sample, so it implies that the practical depth of the OCT image might be relatively larger than that of the measured $a(z)$, whereas a negative dZ means that the practical depth of the OCT image is relatively smaller than that of the measured $a(z)$.

In 'Step-5', the stage is moved laterally for one step, and 'Step-1' is performed again until the moved position of the stage is the end of the scan on the x -axis.



(a) OCT image



(b) the probe position at the start (*left*) and end times (*right*) of the lateral scan

Fig 8. Image of an onion sample by the active topological variance compensation algorithm with extended imaging depth. (a) OCT image, (b) the probe position at the start (*left*) and end times (*right*) of the lateral scan.

3 Results and Discussions

The performance of the active topology compensation algorithm was tested under static conditions using an onion sample with several layers of highly curved surfaces. At first, a B-scan 2-D OCT image was obtained by the conventional fixed-stage method, as shown in Figure 7(a).

The 860 nm CP-FD-OCT provided effective imaging in the range below 500 nm and the structure of some of the layers was very clear within this range. However, the OCT image fades away as the probe is moved further away from the sample's surface, due to the limited depth range, as shown in Figure 7 (b).

Figure 8 (a) shows an improved OCT image obtained using the active topological variance compensation algorithm. By using our algorithm, the probe could actively track the sample surface variance, as shown in Figure 8 (b) and, consequently, the effective imaging depth was extended to the probe's free-moving range. Also, the sub-layers of the sample could be monitored more clearly, even if the distance between the probe and sample's surface was outside of the limited imaging range.

4 CONCLUSION

We developed CP-FD-OCT systems with an active surface tracking algorithm to extend the image range of OCT scanning. Consequently, the OCT images obtained using the motorized-stage-based system showed a significantly extended imaging range through real-time accurate depth tracking. These results demonstrate that our OCT system and algorithms have good potential to resolve several of the limitations of conventional OCT systems.

5 REFERENCES

- [1] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, J. G. Fujimoto, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178-1181, 1991.
- [2] S. A. Boppart, B. E. Bouma, C. Pitris, J. F. Southern, M. E. Brezinski, J. G. Fujimoto, "In vivo cellular optical coherence tomography imaging," *Nat. Med.*, vol. 4, no. 7, pp. 861-865, 1998.
- [3] G. J. Tearney, M. E. Brezinski, B. E. Bouma, S. A. Boppart, C. Pitris, J. F. Southern, J. G. Fujimoto, "In vivo endoscopic optical biopsy with optical coherence tomography," *Science*, vol. 276, no. 5321, pp. 2037-2039, 1997.
- [4] B. Vakhtin, D. J. Kane, W. R. Wood, K. A. Peterson, "Common-path interferometer for frequency-domain optical coherence tomography," *Appl. Optics*, vol. 42, no. 34, pp. 6953-6958, 2003.
- [5] R. Tumlinson, J. K. Barton, B. Povazay, H. Sattman, A. Unterhuber, R. A. Leitgeb, W. Drexler, "Endoscope-tip interferometer for ultrahigh resolution frequency domain

optical coherence tomography in mouse colon," *Opt. Express*, vol. 14, no. 5, pp. 1878-1887, 2006.

[6] U. Sharma, J. U. Kang, "Common-path optical coherence tomography with side-viewing bare fiber probe for endoscopic optical coherence tomography," *Rev. Sci. Instrum.*, vol. 78, no. 11, pp. 113102, 2007.

[7] K. M. Tan, M. Mazilu, T. H. Chow, W. M. Lee, K. Taguchi, B. K. Ng, W. Sibbett, C. S. Herrington, C.T. A. Brown, K. Dholakia, "In-fiber common-path optical coherence tomography using a conical-tip fiber," *Opt. Express*, vol. 17, no. 4, pp. 2375-2384, 2009.

[8] Low, G. Tearney, B. Bouma, I. Jang, "Technology insight: Optical coherence tomography - Current status and future development," *Nat. Clin. Pract. Card.*, vol. 3, no. 3, pp. 154-162, 2006.

[9] N. Iftimia, B. Bouma, J. F. de Boer, B. Park, B. Cense, G. Tearney, "Adaptive ranging for optical coherence tomography," *Opt. Express*, vol. 12, no. 17, pp. 4025-4034, 2004.

[10] G. Maguluri, M. Mujat, B. Park, K. Kim, W. Sun, N. Iftimia, R. Ferguson, D. Hammer, T. Chen, J. Boer, "Three dimensional tracking for volumetric spectral-domain optical coherence tomography," *Opt. Express*, vol. 15, no. 25, pp. 16808-16817, 2007.

[11] M. Pircher, B. Baumann, E. Götzinger, H. Sattmann, C. K. Hitzenberger, "Simultaneous SLO/OCT imaging of the human retina with axial eye motion correction," *Opt. Express*, vol. 15, no. 25, pp. 16922-16932, 2007.

[12] K. Aljaseem, A. Werber, H. Zappe, "Tunable endoscopic MEMS-probe for optical coherence tomography," *Proc. 2007 IEEE/LEOS Int. Conf. Optical MEMS and Nanophotonics*, pp. 8-9, Hualien, Taiwan, 2007.

[13] [13] H. Wang, Y. Pan, A. M. Rollins, "Extending the effective imaging range of Fourier-domain optical coherence tomography using a fiber optic switch," *Opt. Lett.*, vol. 33, no. 22, pp. 2632-2634, 2008.

[14] K. Zhang, W. Wang, J. Han, J. U. Kang, "A surface topology and motion compensation system for microsurgery guidance and intervention based on common-path optical coherence tomography," *IEEE T. Biomed. Eng.*, vol. 56, no. 9, pp. 2318-2321, 2009.

[15] J. Luo, K. Ying, J. Bai, "Savitzky-Golay smoothing and differentiation filter for even number data," *Signal Process.*, vol. 85, no. 7, pp. 1429-1434, 2005.

Registration of Confocal Fluorescence Endomicroscopy Images Using Phase Correlation

Feng Zhao and T. M. McGinnity
Intelligent Systems Research Centre
University of Ulster, Magee Campus
Londonderry, UK, BT48 7JL

Abstract—*The emerging confocal fluorescence endomicroscope is capable of imaging living tissues in a non-invasive way using a probe to continuously scan the surface and sub-surface tissue structures. Due to possible tissue movement and tissue expansion/contraction, the acquired images contain various noises and distortions. It is necessary to align these images in order to obtain a better 3D reconstruction of the tissue's microstructure for clinicians. In this paper, we present an automatic image registration method using the phase correlation technique, which uses a fast frequency-domain approach to estimate the relative transformation parameters between two consecutive endomicroscopy images.*

Keywords: Confocal Fluorescence Endomicroscope, FFT, Phase Correlation, Image Registration

1. Introduction

The confocal fluorescence endomicroscope (Fig. 1) is a newly developed endoscopic tool that makes it possible to carry out *in vivo* microscopic observations of living subjects with about 1000-time magnification and subcellular resolution [1], [2], [3]. An endomicroscope is more powerful than a microscope or an endoscope which generally needs biopsy and carries the risk of causing bleeding, infection, perforation, or mechanical agitation that may lead to the spread of tumor cells through the blood and lymphatic vessels [4]. In addition, a microscope or an endoscope can only see the surface layer without depth resolution.

As illustrated in Fig. 2, the endomicroscope operates in a non-invasive way. By placing the probe on the surface of the target subject, it enables direct observation of molecular mechanisms by continuously scanning the surface and subsurface tissue structures without removing tissues from the body or sacrificing animals. The fluorescence imaging parameters are optimised for a wide range of tissues including brain, intestine, lungs, colon, kidneys, muscle, heart, liver, and pancreas. Thus, molecular imaging of different types of tissues and diseases is becoming feasible, and thus has the potential to facilitate early diagnosis of cancers. Compared with other multi-million pounds imaging instruments such as magnetic resonance imaging (MRI) scanners [5], [6] and X-ray computed tomography (CT)

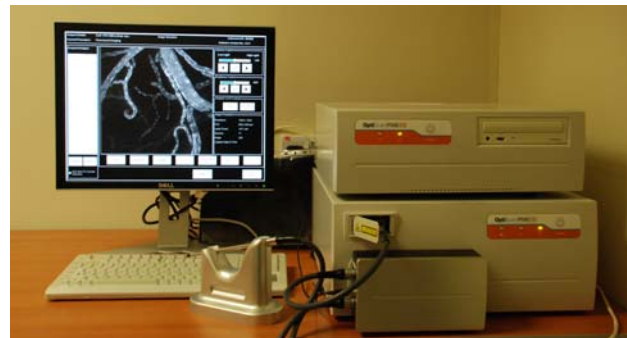


Fig. 1: The *in vivo* cellular imaging system using an endomicroscope.

scanners [7], [8], the endomicroscope is of much lower cost (£100K). While MRI and CT are widely used in disease diagnosis by acquiring global information from the scanned subjects, the new technique of endomicroscopy can provide complementary local information in detail for the clinicians, thus further improving disease diagnosis accuracy. This will provide profound health benefits to society. The technique will also enable further advancement in the field of basic cell biology, aid our understanding of the mechanism of disease progression, and allow monitoring of drug effects at the cellular level.

Before the imaging process is carried out, a fluorescence dye is injected into the target subject. After half an hour, the endomicroscope probe is placed on the subject's surface. The probe begins to scan an area of $475\mu\text{m} \times 475\mu\text{m}$ (field of view) from the surface layer. Once an image is acquired and saved, it continues to scan the subsurface layer ($4\mu\text{m}$ below the surface) by adjusting the laser illumination within the probe. This process continues until the laser light reaches the deepest layer ($250\mu\text{m}$ below the surface). Finally, a stack of 60 slice images is obtained with a resolution as high as 1024×1024 . Then the probe may be moved to a new site to capture another image stack.

The acquired endomicroscopy images are quite different from natural images in several aspects: (1) the images are molecular imaging of the living tissues across a $475\mu\text{m} \times 475\mu\text{m}$ area, (2) they are usually magnified by 1000 times by the microscopic probe, (3) the images are labelled with photosensitisers that selectively accumulate within the tissue,

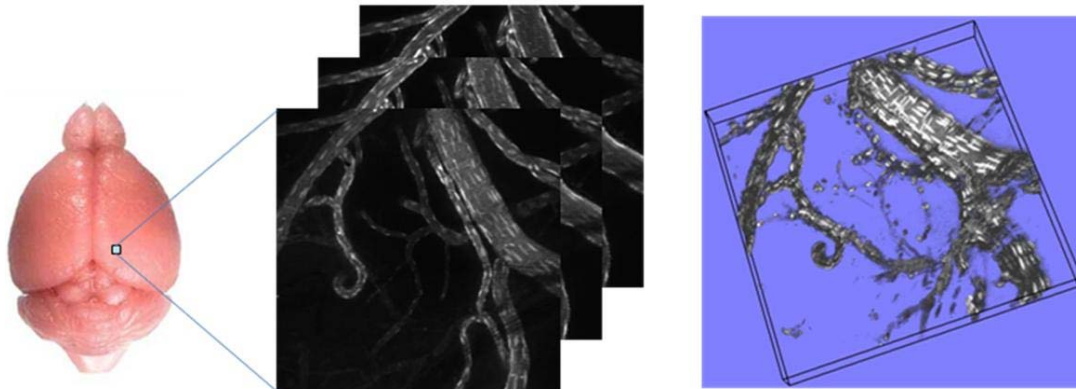


Fig. 2: The endomicroscopy images of mouse brain microvasculature at different z-depths and its 3D reconstruction.

(4) the fluorescence images are much noisier due to low signal to noise ratio, (5) the images are non-uniformly illuminated, and (6) the images are translation-, rotation-, and scale-variant. To reconstruct the 3D microstructure of the living tissue, we need to first register/align consecutive slice images. The challenges are twofold. On the one hand, each of the arbitrarily taken slice images suffers from various distortions due to possible tissue movement and tissue expansion/contraction. On the other hand, beyond a certain time frame, the 3D volumetric images may be different due to physiological changes.

Image registration or image alignment algorithms can be classified into two categories: spatial-domain methods and frequency-domain methods. One of the images is referred to as the reference and the second image is referred to as the target. In this work, we present a phase correlation-based image registration algorithm, which finds the transformation parameters while working in the frequency domain. Applying the phase correlation method to a pair of images produces a third image which contains a single peak. The location of this peak corresponds to the relative translation between the images. Compared with the spatial-domain algorithms such as intensity-based correlation methods [9], feature-based methods [10], and graph-theoretic methods [10], the phase correlation method is resilient to noise, occlusions, and other defects typically in the biomedical images. Additionally, the phase correlation uses the fast Fourier transform (FFT) to compute the cross-correlation between the two images, generally resulting in large performance gains. The method can be extended to determine rotation and scaling differences between two images by first converting the images to log-polar coordinates. Due to properties of the Fourier transform, the rotation and scaling parameters can be determined in a manner invariant to translation.

2. Theoretical Analysis

Assume two images $I_1(x, y)$ and $I_2(x, y)$ with a displacement (x_0, y_0) , i.e., $I_2(x, y) = I_1(x - x_0, y - y_0)$. Applying

the Fourier transform, we have,

$$I_2(u, v) = e^{-j2\pi(ux_0+vy_0)}I_1(u, v). \quad (1)$$

The cross-power spectrum of the two images is defined as,

$$\frac{I_1(u, v)I_2^*(u, v)}{|I_1(u, v)I_2^*(u, v)|} = e^{j2\pi(ux_0+vy_0)}, \quad (2)$$

where $I_2^*(u, v)$ is the complex conjugate of $I_2(u, v)$. The Fourier shift theorem guarantees that the phase of the cross-power spectrum is equivalent to the phase difference between the images.

By applying the inverse Fourier transform to the above phase difference, we have an impulse function $r(x, y) = \delta(x - x_0, y - y_0)$. The location of its peak value corresponds to the displacement that is needed to optimally register the two images. Fig. 3 shows the flowchart of the phase correlation technique.

The advantage of this method is that the discrete Fourier transform and its inverse can be performed using the fast Fourier transform, which is much faster than intensity-based correlation for large images. In practice, it is more likely that $I_2(x, y)$ will be a simple linear shift of $I_1(x, y)$, rather than a circular shift as required. In such cases, $r(x, y)$ may not be a simple delta function, which can possibly reduce the performance of the method. Therefore, a window function such as the Hamming window [11] should be employed during the Fourier transform to reduce edge effects, or the images should be zero padded so that the edge effects can be ignored. If the images consist of a flat background, with all detail situated away from the edges, then a linear shift will be equivalent to a circular shift, and the above derivation will hold exactly. For periodic images such as a chessboard, phase correlation may yield ambiguous results with several peaks in the resulting output.

The method can be extended to determine the rotation and scaling differences between two images by first converting the images to the log-polar coordinates. Assume $I_2(x, y)$ is a translated, rotated, and scaled replica of $I_1(x, y)$ with displacement (x_0, y_0) , rotation θ_0 , and scale s , according

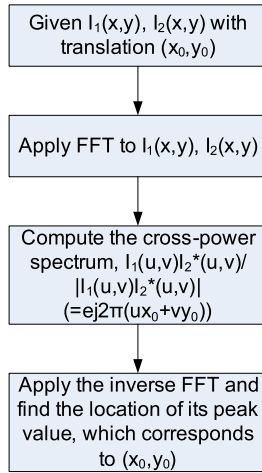


Fig. 3: The phase correlation technique.

to the Fourier translation, rotation, and scale properties, we have,

$$I_2(u, v) = e^{-j2\pi(u'x_0 + v'y_0)} I_1(su' \cos \theta_0 - sv' \sin \theta_0, su' \sin \theta_0 + sv' \cos \theta_0). \quad (3)$$

Assume $F_1(u', v') = |I_1(u', v')|$ and $F_2(u, v) = |I_2(u, v)|$ are their Fourier magnitude spectra, we have,

$$F_2(u, v) = F_1(su' \cos \theta_0 - sv' \sin \theta_0, su' \sin \theta_0 + sv' \cos \theta_0), \quad (4)$$

i.e.,

$$\begin{aligned} u &= s(u' \cos \theta_0 - v' \sin \theta_0), \\ v &= s(u' \sin \theta_0 + v' \cos \theta_0). \end{aligned} \quad (5)$$

In the polar coordinate system, we have,

$$\begin{aligned} u &= \rho \cos \theta, \\ v &= \rho \sin \theta, \end{aligned} \quad (6)$$

and

$$\begin{aligned} u' &= \rho' \cos \theta', \\ v' &= \rho' \sin \theta', \end{aligned} \quad (7)$$

By combining Eqs. (5)-(7), we have,

$$\begin{aligned} u &= s(\rho' \cos \theta' \cos \theta_0 - \rho' \sin \theta' \sin \theta_0), \\ &= s\rho' \cos(\theta' + \theta_0), \\ &= \rho \cos \theta, \\ v &= s(\rho' \cos \theta' \sin \theta_0 + \rho' \sin \theta' \cos \theta_0), \\ &= s\rho' \sin(\theta' + \theta_0), \\ &= \rho \sin \theta, \end{aligned} \quad (8)$$

i.e.,

$$\begin{aligned} \rho &= s\rho' \Rightarrow \rho' = \rho/s, \\ \theta &= \theta' + \theta_0 \Rightarrow \theta' = \theta - \theta_0. \end{aligned} \quad (10)$$

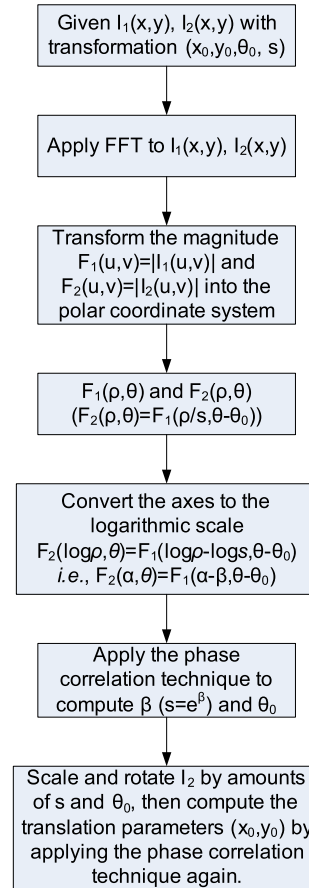


Fig. 4: The flowchart of the phase correlation-based image registration.

Thus, the Fourier magnitude spectra $F_1(u', v')$ and $F_2(u, v)$ in the polar representation are related by,

$$F_2(\rho, \theta) = F_1(\rho/s, \theta - \theta_0). \quad (11)$$

By converting the axes to logarithmic scale, we have,

$$F_2(\log \rho, \theta) = F_1(\log \rho - \log s, \theta - \theta_0), \quad (12)$$

i.e.,

$$F_2(\alpha, \theta) = F_1(\alpha - \beta, \theta - \theta_0), \quad (13)$$

where $\alpha = \log \rho$, $\beta = \log s$. Thus, the problem becomes one with relative translation only. Applying the phase correlation technique, we can find the scale $s = e^\beta$ and rotation θ_0 .

After scaling and rotating $I_2(x, y)$ by the amounts of s and θ_0 respectively, the translation parameters x_0 and y_0 can then be obtained using the phase correlation technique. Fig. 4 summarises the process of the phase correlation-based image registration approach.

3. Experimental Results

To evaluate the performance of the algorithms, we perform a series of experiments on several sets of confocal fluorescence endomicroscopy images. Note that all the images in

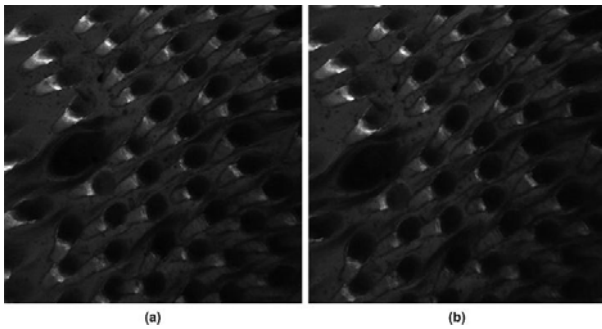


Fig. 5: Mouse tongue images (1024×1024) with displacement. (a) The original reference image and (b) the target image.

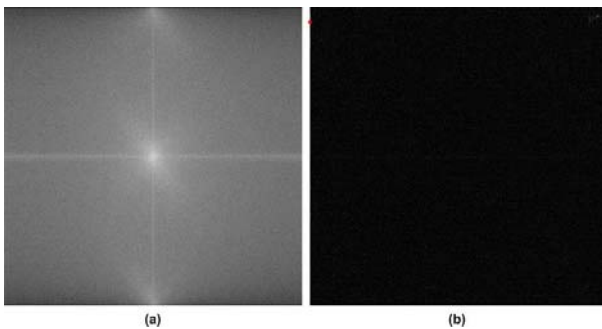


Fig. 6: The Fourier spectrum of the original reference image and the phase correlation image of the tongue images, where the translation parameters is estimated as: $x=1$, $y=49$.

the following figures are largely reduced for display purpose. First, we apply the phase correlation technique to a pair of mouse tongue images, as shown in Fig. 5. Fig. 6 shows the Fourier spectrum of the original reference image and the inverse Fourier transform of the cross-power spectrum of the tongue images. We can see a peak in the phase-correlation image approximately at (1, 49). Theoretically, the peak value should be equal to 1.0. However, the presence of dissimilar parts and the noise in images reduce the peak value. The aligned target image is illustrated in Fig. 7. Experimental results conducted on a pair of mouth images are shown in Fig. 8. From these results, we can see that the phase correlation technique does not work without preprocessing the endomicroscopy images.

In order to obtain a reasonable estimation of the translation parameters, we filter the original endomicroscopy images by a Laplacian filter to remove high-frequency components in the frequency domain. Fig. 9 shows the Fourier spectrum of the reference tongue and mouth images after applying the Laplacian filter and their aligned target images. The improved results demonstrate that by applying the Laplacian filter, the phase correlation is applicable to the endomicroscopy images with displacement.

Fig. 10 shows registration results on the mouse tongue (another pair) and mouse brain microvasculature images.

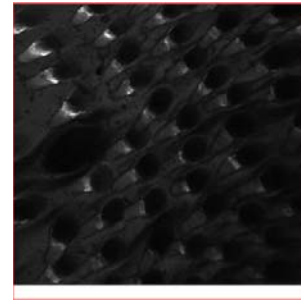


Fig. 7: The aligned mouse tongue images with white-pixel padding.

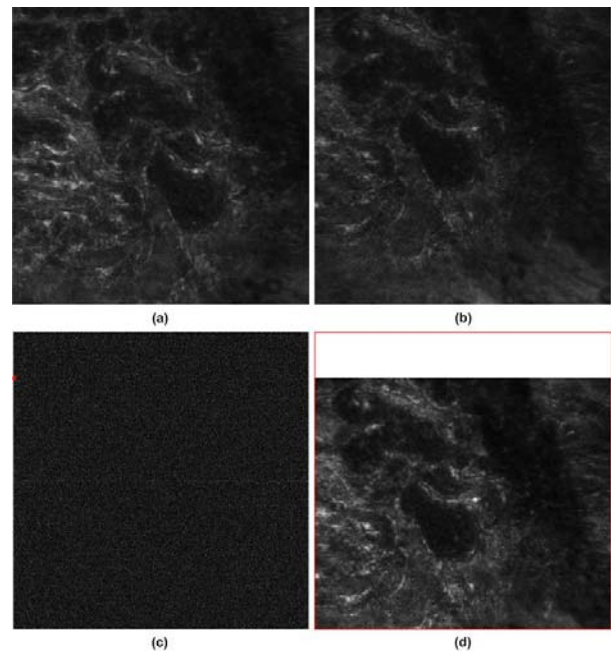


Fig. 8: The mouth image pair (1024×1024) with displacement. (a) The original reference image, (b) the target image, (c) the phase correlation image with the estimated translation parameters: $x=2$, $y=159$, and (d) the aligned tongue images with white-pixel padding.

From these experimental results, we can see that the phase correlation method is a robust approach for the estimation of the transformation parameters, leading to good image registration results.

4. Conclusions

In this work, we have developed a phase correlation-based registration approach for estimation of the relative transformations in the consecutive endomicroscopy images. The experimental results conducted on different sets of images reveal that the phase correlation-based alignment can be performed in real time and is robust to noise, occlusions, and other defects existing in the images. The good alignment between consecutive slice images will directly benefit the subsequent 3D reconstruction and visualisation of the

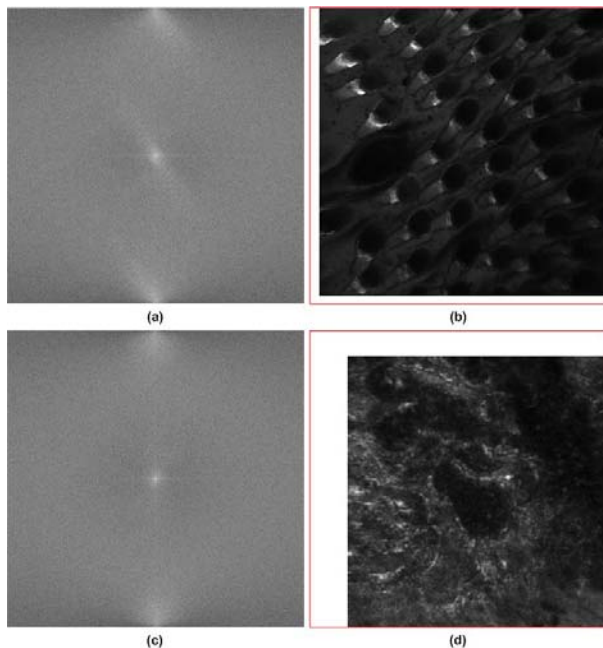


Fig. 9: The phase correlation results after applying the Laplacian filter. (a) The Fourier spectrum of the filtered reference tongue image, (b) the aligned tongue image using the estimated translation parameters ($x=36$, $y=27$) by phase correlation, (c) the Fourier spectrum of the filtered reference mouth image, and (d) the aligned mouth image using the estimated translation parameters ($x=133$, $y=89$).

living tissue's microstructure. It will enable clinicians to navigate within the living tissue freely, leading to a much more clinician-friendly imaging tool, and more definitive diagnostic results of various diseases including early-stage cancers, in a non-invasive way. This will provide profound health benefits to society.

Acknowledgment

This work is supported by the "Strengthening the All-Island Research Base" project, funded by the Northern Ireland Department of Education and Learning. The authors would like to thank M. Goetz, C. Schneider, et al. (University of Mainz, Germany), who provided the mouse brain microvasculature images for this study. We are also thankful to Prof. Hock Soon Seah, Dr. Feng Lin and Dr. Kemao Qian (Nanyang Technological University, Singapore), Prof. Soo Khee Chee, Dr. Malini Olivo and Dr. Patricia Thong (National Cancer Centre Singapore), for sharing the mouse tongue and mouth data with us.

References

- [1] M. Goetz, C. Fottner, E. Schirmacher, P. Delaney, S. Gregor, C. Schneider, D. Strand, S. Kanzler, B. Memadathil, E. Weyand, and M. Holtmann et al., "In-vivo confocal real-time mini-microscopy in animal models of human inflammatory and neoplastic diseases," *Endoscopy*, vol. 39, pp. 350–356, 2007.
- [2] M. Goetz, S. Thomas, A. Heimann, P. Delaney, C. Schneider, M. Relle, A. Schwarting, P. R. Galle, O. Kempfski, M. F. Neurath, and R. Kiesslich, "Dynamic imaging of microvasculature and perfusion by miniaturised confocal laser microscopy," *Eur. Surg. Res.*, vol. 41, pp. 290–297, 2008.
- [3] P. S.-P. Thong, M. Olivo, K.-W. Kho, W. Zheng, K. Mancer, M. Harris, and K.-C. Soo, "Laser confocal endomicroscopy as a novel technique for fluorescence diagnostic imaging of the oral cavity," *J. Biomedical Optics*, vol. 12, no. 014007, pp. 1–8, 2007.
- [4] F. Koenig, J. Knittel, and H. Stepp, "Diagnosing cancer in vivo," *Science*, vol. 292, pp. 1401–1403, 2001.
- [5] D. B. Parente, E. L. Gaspardo, L. C. H. da Cruz, R. C. Domingues, A. C. Baptista, A. C. P. Carvalho, and R. C. Domingues, "Potential role of diffusion tensor MRI in the differential diagnosis of mild cognitive impairment and alzheimer's disease," *American J. Roentgenology*, vol. 190, no. 5, pp. 1369–1374, 2008.
- [6] A. G. Filler, "The history, development, and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, DTL," *Internet J. Neurosurgery*, vol. 7, no. 1, 2010.
- [7] G. T. Herman, *Fundamentals of Computerized Tomography: Image Reconstruction from Projection*, 2nd ed. Springer, 2009.
- [8] R. Smith-Bindman, J. Lipson, R. Marcus, K.-P. Kim, M. Mahesh, R. Gould, A. B. de González, and D. L. Miglioretti, "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer," *Arch. Intern. Med.*, vol. 169, no. 22, pp. 2078–2086, 2009.
- [9] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients," *IEEE Trans. Medical Imaging*, vol. 23, no. 11, pp. 1430–1444, 2004.
- [10] A. A. Goshtasby, *2-D and 3-D Image Registration for Medical, Remote Sensing, and Industrial Applications*. Wiley Press, 2005.
- [11] Y. Song and X. Peng, "Spectra analysis of sampling and reconstructing continuous signal using hamming window function," in *Proc. IEEE Fourth Int'l Conf. Natural Computation*, 2008, pp. 48–52.

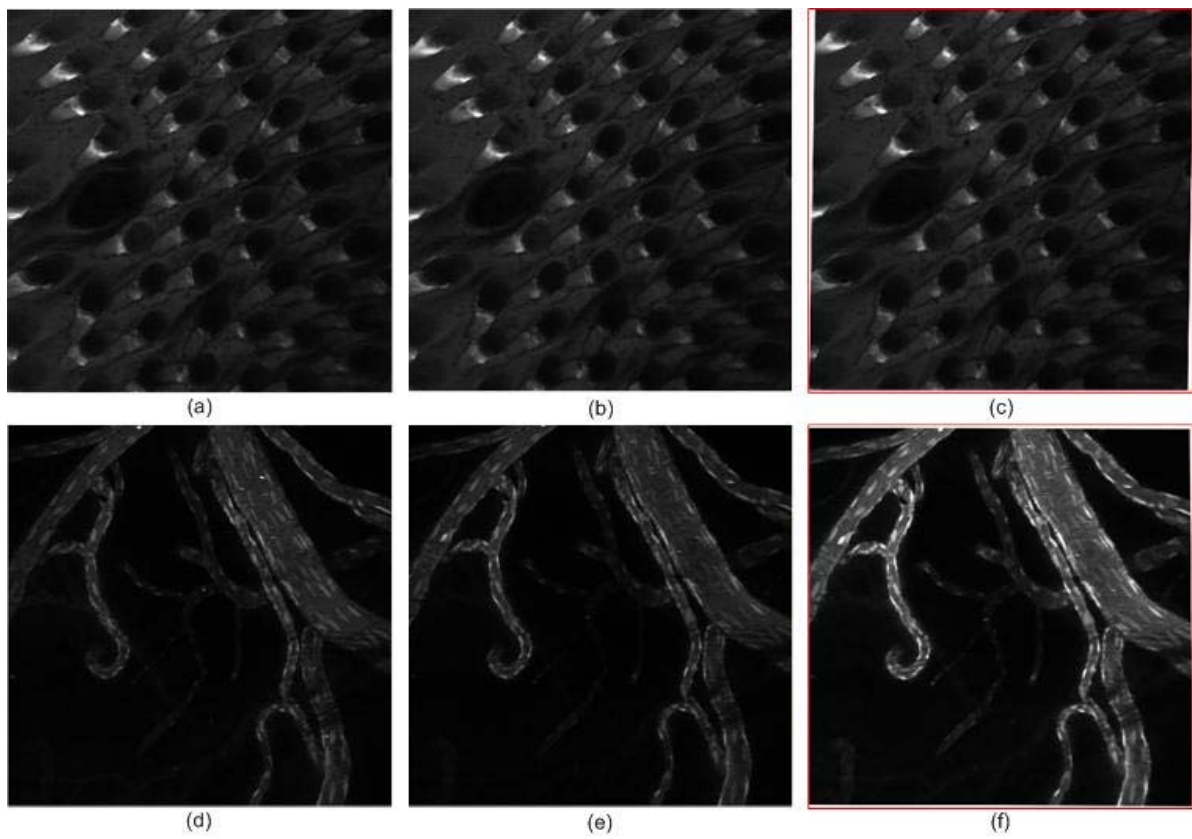


Fig. 10: The alignment results of mouse tongue and brain microvasculature images (1024×1024). (Left column) The original reference images, (middle column) the target images, and (right column) the aligned images with white-pixel padding.

Signal Processing Algorithm for Wireless ECG Monitoring Systems

Abishek T.K, Sahajhaksh Hariharan, and Dr.Maneesha V Ramesh

Amrita Center for Wireless Networks and Applications, Amrita University, Kollam, Kerala, India

Abstract - *The Electrocardiogram (ECG) is a graphical recording of the electrical signals generated by the heart. The signals are generated when the cardiac muscles depolarize in response to electrical impulses generated by the pacemaker. In this work, we propose an efficient method to monitor and classify the ECG signals. The initial task carried out was to eliminate the noise, which involved extracting the required cardiac components by rejecting the background noise. The second task was to perform R peak detection, which was achieved by using the Windowed Short Time Fourier Transform (STFT). The Heart Rate Variability (HRV) was also found by calculating the difference between two simultaneous R-Peaks. The simulations were carried out in the MATLAB environment. The experiments were carried out using data from the MIT-BIH Database. This paper proposes an algorithm to monitor cardiac atrial fibrillation, which is an essential precursor to myocardial infarction.*

Keywords: Short Time Fourier Transform, Atrial Fibrillation, Teager Energy Operator (TEO), Empirical Mode Décomposition, Electrocardiogram

1 Introduction

Bioelectrical signals express the electrical functionality of different organs in the human body. The ECG is an important signal amongst all bioelectrical signals. It reflects the performance and the properties of the human heart and conveys very important hidden information in its structure. This information has to be extracted and analyzed before any useful and meaningful interpretations can be made. Extracting or decoding this information or features from the ECG signal are found to be very helpful in explaining and identifying various pathological conditions. The second phase of this work comprises of extracting the features, which is accomplished in a straightforward manner by analyzing the ECG visually on paper or on screen. [1] However, the complexity and the duration of ECG signals are often quite considerable, making manual analysis a very time-consuming and limited solution. [3]. In addition, manual feature extraction is always prone to errors. Therefore, ECG signal processing has become an indispensable tool for extracting clinically significant information from ECG signals, thereby reducing the

subjectivity of manual ECG signal analysis. The proposed system is a Wireless ECG Monitoring System which incorporates a Signal Processing Algorithm for pre-processing and peak-detection of the ECG signal. Being a wireless system, it overcomes the mobility and environment problem. The system also gives out warning signals to the doctor about possible cardio-vascular disorders in patients who could be remotely located. Section II describes the related work. Section III explains the architecture and design of the proposed wireless ECG Monitoring System. Section IV describes the proposed algorithm. Section V describes the algorithm used to extract the features of the ECG using Daubechies 4-tap algorithm. Section VI describes the Implementation and Section VII describes the Conclusion.

2 Related Work

In [1], the authors described the difference between the original and the filtered ECG signal pattern. There was also a study conducted by the authors about the convergence time, the execution time and the relative statistics in time and frequency domain for the ECG signal.

In [2], the authors described how wavelets could be used in combination with Neural Networks to model ECG signal. In this paper the authors make use of the multi-resolution nature of wavelets and the adaptive learning ability of Artificial Neural Networks which is trained by an algorithm that includes the Particle Swarm Optimization (or the PSO).

In [3], the authors describe about foveation which modulates the coefficients of the Discrete Wavelet Transform (DWT) of an ECG record. This process is mainly used to select the major portions of interest in an ECG record by using a mask in the spatial domain. Also, they say foveation can be used for denoising and coefficient quantization.

In [4], the authors described the usage of Hidden Markov models to classify the ECG waveform. The classification is done after the ECG is decomposed into three levels of decomposition using Wavelet Transforms. There are three types of classifications described based on the number of beats. They are Normal (N), Atrial Flutter (AF) which often acts as a precursor to myocardial infarction, and Normal Sinus Rhythm (NSR).

In [5], the authors describe about a technique called Phase-rectified signal averaging which is a method recently introduced in the field of signal processing to process quasi-

periodic signals. Herein the authors use this approach to detect Atrial Fibrillation from the surface ECG components. The fibrillation components are highly contaminated ventricular complexes, and the cancellation of these components is never perfect. Hence, this method was adopted to cancel out these artifacts.

In [6], the authors detect the QRS complexes using an operator known as Teager energy operator. This operator operates only on three adjacent samples of the ECG and requires only three arithmetic operations per time shift. This method adopted by the authors gives them 99.9% efficient results for the MIT-BIH database.

In [7], the authors formulate an algorithm for robust QRS onset and offset detection. This algorithm developed was more efficient when tested on MIT-BIH database. The algorithm produced good results for QRS offset and onset detection.

In [8], the authors have proposed an algorithm for ECG signal denoising using Hilbert-Huang transform. The authors use empirical decomposition method to decompose the noisy signal into Intrinsic Mode functions (IMF's). Spectral analysis was conducted on the successive IMF's to find out the boundary between the noise dominated IMF's and ECG signal dominated IMF's. The authors carry out simulation experiments and claim that this method is more efficient compared to wavelet denoising method.

In [9], the authors use EMD method to decompose an ECG signal. Hilbert transform was used for spectral analysis. The authors claim that decomposing the signal into IMF's is more suitable compared to wavelet denoising methods.

In [10], the authors use an approach to detect the pacemaker pulses from the ECG. In order to realize this they proposed a fully digital approach that uses a two step filtering strategy which was then followed by a thresholding mechanism. The results obtained after the simulations were carried out were very significant and they claim that it outperforms all the results that were obtained from a well known patented algorithm.

3 Architecture and design

The top overall architecture of the proposed system is depicted in Figure 1. The proposed system has a two tier architecture.

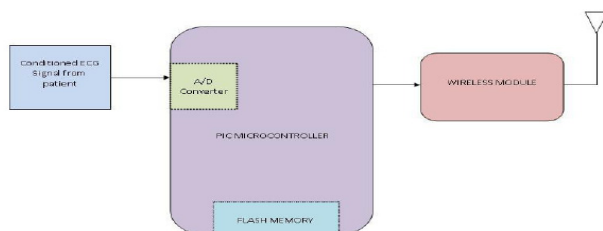


Fig 1 : ECG Transmitting Unit

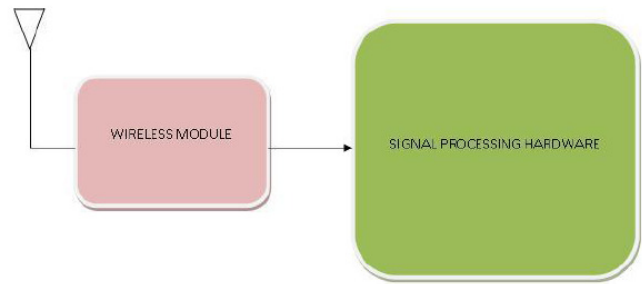


Fig 2 : ECG Receiving Unit

3.1 ECG Acquisition and Transmission Unit (EcgATU)

EcgATU is the module that has an interface with the patient. This module acquires the ECG signal from the patient, performs basic signal processing operations and wirelessly transmits to the ECG Receiving Unit otherwise known as (EcgRU).

The EcgATU and EcgRU are the two modules shown above. The conditioned ECG signal was extracted from the patient. It was then given as an input to the ADC of the microcontroller to obtain its digital equivalent. The microcontroller used was placed on the NI-ELVIS MX kit. The specification of the PIC used in this context was PIC16F877A which is a 32-bit pin microcontroller. The program was loaded onto the PIC using a software known as WINPIC-800. The language used in this context was Embedded C. Since the language used was Embedded C, we are intentionally converting it to an embedded microcontroller.

3.2 Procedure adopted to detect R-Peaks

The ECG signal we have is uneven, thus our first step is to straighten it. To say in mathematical language, we should remove the low frequency component. To achieve this we applied Fast Fourier Transform (FFT), which restores the low frequency components and restores the ECG with the help of Inverse Fast Fourier Transform (IFFT). In the next step, we found out the local maxima, we achieved this using the windowed filter; that sees only the maximum in its window and ignores all the other values. Window of default size was used in this case. Next step is to remove all small values and preserve the significant ones. For this purpose, we used a threshold filter. In this case, the result is good in general case. But we can't be sure that we have all the peaks. So the next step is to adjust the filter window size and repeat filtering. It is only after performing these operations we obtain the result. The signal processing algorithm used incorporates Short Time Fourier Transform (STFT) for spectral analysis. Below the continuous and discrete time versions of the Fourier Transform are shown in (1) and (2) [7]. The block diagram for the spectral detection is as shown below:

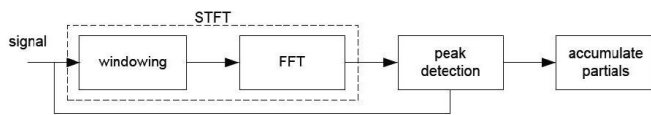


Fig 3 : Block diagram for STFT and Spectral Detection

$$X(\omega) = \int x(t)e^{-j\omega t} dt \quad (1)$$

$$X[n] = \sum x[n]e^{-jk2\pi n/N} \quad (2)$$

3.3 Heart rate calculator

Heart rate in general, can be calculated using several time domain and frequency domain methods. In our analysis, we used a signal length of N=64 for computing the heart rate.

The formula used to compute the Heart Rate Variability is as follows:

$$PNN25 = (NN25)/(N - 1)*100 \quad (3)$$

Where, PNN25 is the ratio of the number of successive difference of intervals which differ by more than 25ms to the total number of all RR intervals.

4 Proposed algorithm

Condition 1: The algorithm proposed here checks for the normal and abnormal functioning of the heart. It does this by accepting an ECG signal as an input signal. The time interval of the ECG signal is checked for assertion case i.e. width of the signal should be between 0.023s and 0.1s, if the above condition validates, then R-R interval is estimated to be between 0.6 and 1.1s. If these two conditions satisfy, then we can say that the person is not suffering from any cardiac disorder.

Condition 2: If the width of the signal is less than 0.023s and R-R is estimated to be greater than 1.1s, then too we can say that the person is not for any cardiac disorder.

Condition 3: If condition 1 and 2 do not validate, then we can corroborate that the heart is functioning in an abnormal manner.

Condition 4: If none of these conditions satisfy, then we perform the iteration from the beginning.[6]

The detection algorithm can further be elaborated by checking for Rough Peaks in the ECG signal. The method to be adopted is to first check for positive and negative slope threshold values to assist in the selection process.

The following conditions should help in detecting the abnormal R-Peaks:[7-9]

1. The slope must change polarity i.e., from positive to negative.
2. The magnitude difference between the peak candidate and the current bin's magnitude component must exceed the threshold component.
3. A new peak candidate search occurs only after there is a slope change from negative to positive, and when a threshold value is exceeds the normal threshold.

Next, we can look for prominent peaks following the Rough-Peak search method, it can be done in the following manner:

1. The R-Peak with the maximum value is found.
2. Relative to position of the R-Peak with maximum amplitude, peaks are analyzed moving towards the iso-electric line.
3. Local maxima or R-Peaks are picked out using an adaptive threshold value that is reflective of all prominent R-Peaks and neighboring R-Peaks.

The flowchart of the algorithm is as shown in Figure 4.

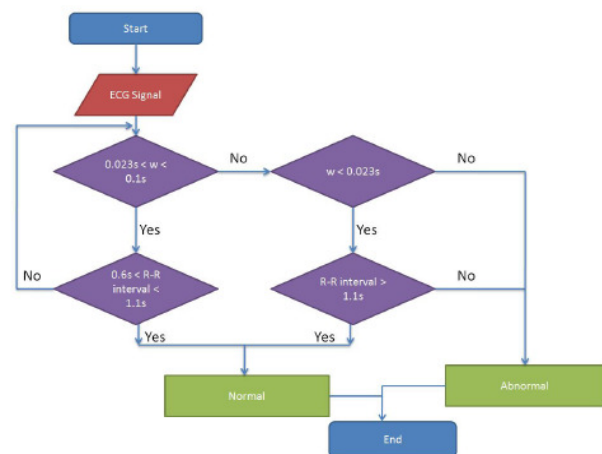


Fig 4 : Flowchart for atrial fibrillation

5 Implementation

As a first phase to the development and implementation of the system, the ECG sample signals were obtained from the MIT-BIH Database and was further used as mat files. These mat files were used as input signals to the developed algorithm.

The algorithm developed was for ECG-R peak detection in MATLAB environment. Two mat files were used and a comparison was drawn between two subjects for finding out

the Heart Rate Variability (HRV). The results obtained are as shown for subject 1 and subject 2. The HRV obtained for the

two samples were 52.13 bpm and 56.24 bpm respectively.

Also, a Java Data Base Connectivity (JDBC) program was written to read the values of the ECG signal from the database which was subsequently plotted. The results indicated that the R-peaks exceeded the normal threshold.

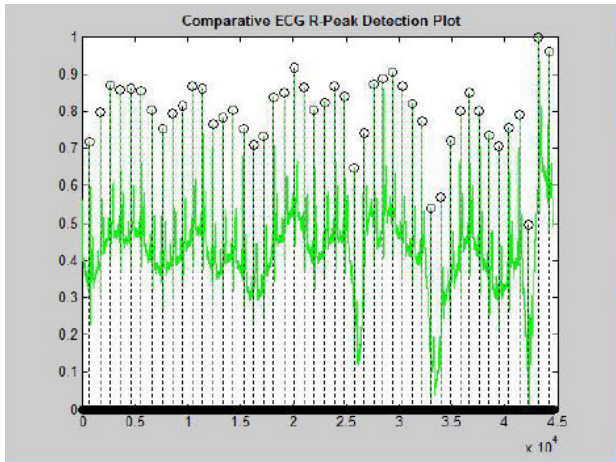


Fig 5 : Comparative R-Peak detection for Sample 1

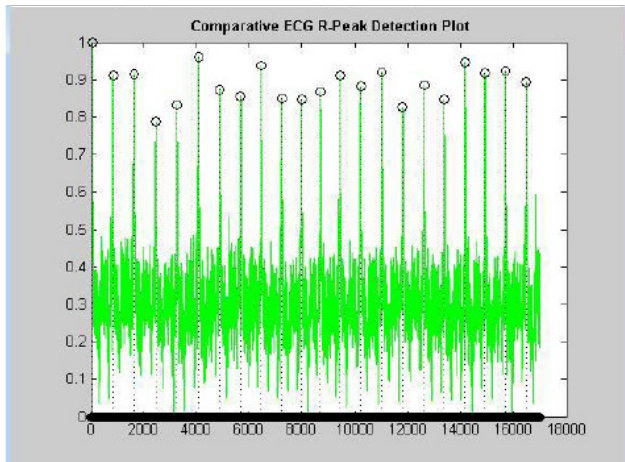


Fig 6 : Comparative R-Peak detection for Sample 2

6 Conclusion

This paper proposes a design for monitoring and detection of the ECG. The main advantage of this system is that remote monitoring and diagnosis is made easier. The system can further be enhanced in performing feature extraction and classification using neural networks. Also, it provides good feasibility and good performance if the objective is to analyze and interpret the ECG in an efficient manner. The algorithm proposed in this paper for atrial fibrillation is less complex compared to the algorithms previously proposed. It reduces the complexity from $O(n^2)$ to $O(n \log n)$.

7 References

- [1] FC Chang, CK Chang, KY Chi, YD Lin, Evaluation Measures for Adaptive PLI Filters in ECG Signal Processing, IEEE Journal for Signal Processing, Year 2007.
- [2] Suranai Pongpouri, Xiao-Hua Yu, ECG Signal Modeling and Noise Reduction using Wavelet Neural Networks, Proceedings of the IEEE International Conference on Automation and Logistics, Shanghai, China Aug 2009.
- [3] Iulian B. Ciocoiu, ECG Signal Compression using Wavelet Foveation, IEEE, 2009
- [4] Pedro R. Gomes, Filomena O. Soares, J. H. Correia, C. S. Lima, Members, IEEE, 32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina, August 31 - September 4, 2010
- [5] Mathieu Lemay, Yann Prudat, Vincent Jacquemet, and Jean-Marc Vesin, Member, IEEE, IEEE Transactions on Biomedical Engineering, Vol. 55, No. 11, November 2008
- [6] ZHANG Aihua, CHAI Long, DONG Hongsheng, QRS Complex Detection of ECG Signal by Using Teager Energy Operator, IEEE, 2008
- [7] A Illanes-Manriquez, Q Zhang, An Algorithm for Robust Detection of QRS Onset and Offset in ECG Signals, Computers in Cardiology 2008, 35:857-860
- [8] Changnian Zhang, Xia Li, Mengmeng Zhang, A novel ECG signal denoising method based on Hilbert-Huang Transform, International Conference on Computer and Communication Technologies in Agriculture Engineering 2010
- [9] B. NarsimhaJ, E. Suresh, K. Punnamchandrar and M Sanjeeva Reddy, IEEE, 2011
- [10] A. Polpetta, Student Member IEEE, P. Banelli, Member IEEE, 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008

Tracking of Active Cells Based on Kalman Filter in Time Lapse of Image Sequences of Neuron Stem Cells

Chunming Tang¹ Yanqing Wang Ying Cui¹

(1 Department of Information & Communication Engineering,
Harbin Engineering University, Harbin, Heilongjiang Province, China 150001)

Abstract - In multi-cells' tracking of time-lapse image sequences imaged by optical microscope, problem of tracking active cells correctly is still unsolved. It affects tracking accuracy and speed whether an algorithm can predict the position of the active cells or not. The tracking strategy is guided by Kalman forecast in image Cartesian coordinates systems, which may search target cells via minimizing their cost function of characteristics and updating their state equations. Prediction and tracking results from six active cells in three image sequences show that the algorithm can track segmented active cells accurately. And the errors between the tracking estimate values and the practical observation values are no more than 10 pixels.

Keywords: tracking; active cells; Kalman filter; neuron stem cells; image sequences

1 Introduction

Most of the cells in human body are unregenerate and suffer from various diseases. Some serious damage can't be repaired through natural processes. But with the development of the research in cytology, these damaged cells can be cured by cell therapy. So far, many researchers have made some achievements in the field. Whether to continue the experiments or not depends on correctly extracting some special stem cells for treatment. How to extract stem cells reliably? What internal mechanism may control stem cells growing and proliferating? What are the exact substances to induce their differentiation? How to find the more effective ways to solve these problems?

Along with the development of computer technology and automatic system, more researchers are inspired in the research field. With the digital microscopic image developing, the imaging system can obtain ideal image sequences for the cells living in vitro during a period. Computer vision can establish mathematical models whether the cells in 2D or 3D environments. In this way researchers may find the solution of the above problems.

To track cells in 2D image sequences, main methods are mostly based on overlapping^[2,6], where man-machine interaction is usually added to eliminate some tracking errors due to segmentation problem. Mean shift is a directly tracking which tracks cells in original image sequences^[1,3,4]. Active contour model^[5], auction algorithm^[7] and other improved algorithms have also been studied in cells' tracking. However, all these algorithms are not very effective in tracking active cells. Although active cells are very few in one sequence, they have some significance in cells movement analysis. So far no papers have discussed the methods in tracking active cells specially. As the active cells are the cells locomotive distance above one average diameter of cells in two successive images, mean shift^[1] which is based on fixed bandwidth is very difficult to find the active cells' centroids. Xiaobo Zhou^[10] has studied cancer cells' cycle progression via Kalman filter. However, on Kalman prediction in the first several frames, whether it is artificially operation or automatic recognition, he has not stated in detail. Kalman filter is applied which is focusing on the prediction of the active cells in this paper.

This paper is organized as follows: segmentation of level set combined with average gray threshold has been introduced firstly. In cells' tracking part, all cells have been classified into two categories: inactive cells and active cells. The former is tracked by overlapping firstly. The latter is tracked by Kalman filter. Tracking results and conclusion have been given in the final part.

2 Segmentation

Level set is a particular contour evolution approaching which has good performance in segmentation. In this paper, segmentation of image sequences of neuron stem cells which have been imaged by confocal microscopy is based on level-set combined with local gray threshold^[9]. The algorithm can not only solve the problem of focal shift but also separate adherent and clustered cells successfully as well as keeping cells' shape and position.

3 Tracking

3.1 Overlapping

Overlapping being used for the inactive cells' tracking is based on the overlap region between previous frame and current frame to ensure one cell having its own ID till last frame of the image sequence. We can track the inactive cells which moved distance is less than L between previous frame and current frame via overlapping, where L refer to the average diameter of all cells in the sequence.

3.2 Identification of active cells

To the active cells, the overlapping does not work. And in this paper if one of the following three criteria is satisfied, the cells is regarded as active cells.

1. The cell's moving distance is more than L in two adjacent frames;
2. The cell moves in from any edge of an image;
3. The cell emergence is due to over-segmentation.

3.3 Kalman filter tracking

Kalman filter^[8] is an optimal recursive data processing algorithm which is extensively applied in tracking and navigation.

In active cells tracking, position and velocity are parameters of motion state. Acceleration is the external input. Although cells' locomotion is chaotic, as sampling interval of image sequences is 10 minutes, the changes of motion state may be very small between two adjacent frames. It is reasonable that the cells motion is uniformly accelerated in two adjacent frames.

We denote the state vector

$X(k)=[x_{s_k} \ xv_k \ ys_k \ yv_k]^T$ and the system external input vector $U(k)=[xa_k \ ya_k]^T$. We assume xs, xv, xa, ys, yv, ya are cells' position, velocity and acceleration on x axis and y axis respectively. Let the interval time of two adjacent frames is T . In T , state transition matrix $A(k)$, system parameters $B(k)$, and observation matrix $H(k)$ are defined in equation (1)~(3).

$$A(k) = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

$$B(k) = \begin{bmatrix} 0.5T^2 & 0 \\ T & 0 \\ 0 & 0.5T^2 \\ 0 & T \end{bmatrix} \tag{2}$$

$$H(k) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{3}$$

If active cells, say C_2 , has been located in the $(t-1)^{th}$ frame, we track C_2 in the $t^{th}, (t+1)^{th}, (t+2)^{th}$ frames using the following method. The searching area is defined as five times of the diameter of C_2 . If the cells are considered as elliptic shape, its long axis, short axis, angle between long axis and x -axis, which is called azimuth, eccentricity and centroid can be regarded as six parameters to describe itself. The cell having the smallest parameters changes is the matching cell of C_2 in the next frame. After C_2 has been tracked in the $t^{th}, (t+1)^{th}, (t+2)^{th}$ frames, we can calculate the initial value of $P(0/0), \hat{X}(0/0)$ and $Z(1)$ in Kalman filter. After $T > (t+2)$, the matching area is estimated via Kalman filter firstly, then local searching is used to save searching time.

Kalman filter is to predict active cells' location form the $(t+3)^{th}$ frame. Searching area is square. Those cells falling into the square are called candidate cells. The best matching one can be found by association operation.

In association operation, we set up a cost function using the above six parameters to show the changes of cells' movement. The smaller the cost function is, the higher probability confidence the matching has.

We assume $\{B_j, j=1, 2, \dots, J\}$ as the candidate cells' centroids set in the t^{th} frame. If \tilde{A} is the best prediction of cell A , the cost function is in equation (4):

$$COST(A, B_j) = \alpha * D(A, B_j) + \beta * K(A, B_j) + \gamma * H(A, B_j) + \sigma * G(A, B_j) + \lambda * S(A, B_j) + \delta * Q(A, B_j) \tag{4}$$

Where, $D(A, B_j) = \frac{distance(\tilde{A}, B_j)}{\max(distance(\tilde{A}, B_j))}$, represents for

distance; $K(A, B_j) = \frac{Vc_j}{\max(Vc_j)}$, represents for

velocity, $V_j = distance(\tilde{A}, B_j) / T$, $V_{cj} = |V - V_j|$;

$H(A, B_j) = \frac{|Area_A - Area_{B_j}|}{\max(|Area_A - Area_{B_j}|)}$, represents for area;

$G(A, B_j) = \frac{|Dia_A - Dia_{B_j}|}{\max(|Dia_A - Dia_{B_j}|)}$, represents for diameter;

$$S(A, B_j) = \frac{|Ecc_A - Ecc_{B_j}|}{\max(|Ecc_A - Ecc_{B_j}|)}, \text{ represents for}$$

eccentricity; $Q(A, B_j) = \frac{|Azi_A - Azi_{B_j}|}{\max(|Azi_A - Azi_{B_j}|)}, \text{ represents for}$

azimuth.

$\alpha, \beta, \gamma, \sigma, \lambda$ and δ are the coefficients in cost function. And the sum of them equals to 1. Find the cell's centroid in candidate cells having the minimal cost function. Then the cell is considered as the best matching one of A in the next frame. Thus the observation vector and the external input vector can be obtained. Kalman filter can be iterated till target cells can not be found, it means the cell having moved out of observation vision or changing its locomotive characteristics.

4 Simulation

The presented algorithm has been tested in three image sequences. The image and cells' information are listed in Table 1. M_1 is the number of the cells which moving distance is greater than L between the adjacent image sequences. M_2 is the number of the cells which move to the edge of image. M_3 is the number of cells which moving distance is greater than L caused by over segmentation between the adjacent frames. M_4 is the number of frames that the cells' moving distance are

greater than L . M_5 is the number of the cells which moving distance are greater than L and frame number must be greater than 2. According to the identification conditions of active cells in 3.2, $M_6 = M_1 - M_2 - M_3 - M_4 + M_5$ is the number of the active cells in the three image sequences.

(1) From Table 1, we know that there are three active cells in sequence I. Because one of them exists only in five frames, and then moves out of the image sequences, we did not simulate it. We track the other two separately. In Fig.1 (a) ~ (e), it shows 1, 10, 20, 40 and 60 frames in the image sequence I respectively, and the corresponding segmentation images. For active cell 1, it has moved out of the image in the 45th frame. So we can track the active cells between the 1st frame and the 44th frame. And it costs 10.625 seconds. For the active cell 2, it is moved into the image in the 46th frame, so we can track the active cells between 46th frame and the 70th frame.

(2) In Fig. 2 (a) ~ (e), it shows 1, 10, 20, 40 and 60 frames in the image sequence II, and the corresponding segmentation images respectively. For the only one active cell, it has moved in the image in the 4th frame. Therefore, we can track it between the 4th frame and the 70th frame. It costs 13.985 seconds.

Table1 Numbers of active cells in the image sequences

Image sequences	Image size and number of frames /(pix * pix * frame)	M_1	M_2	M_3	M_4	M_5	M_6
I	127*127*70	60	34	18	7	2	3
II	184*169*70	36	12	16	8	1	1
III	256*256*50	152	20	119	5	2	10

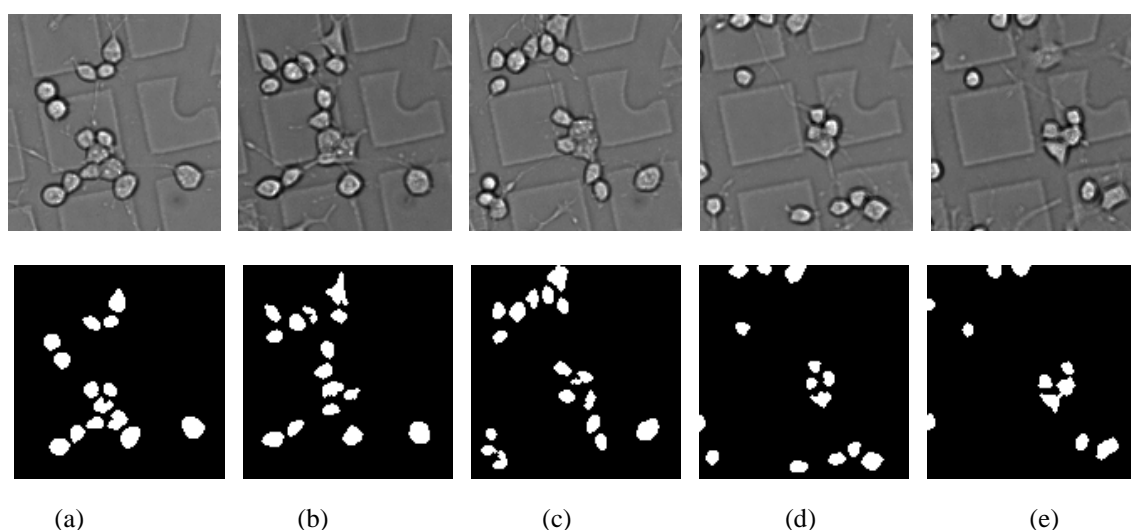


Fig.1 Segmentation of frame 1、10、20、40 and 60 in sequence I

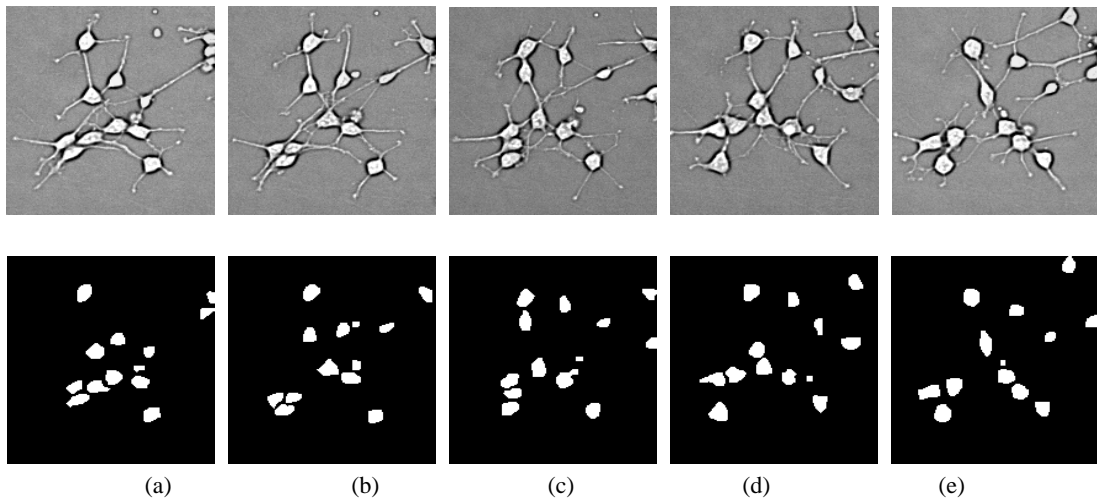


Fig. 2 Segmentation of frame 1、10、20、40 and 60 in sequence II

(3) There are clustered cells and under segmentation cells in image sequence III, which can lead to tracking error. From Table 1, we know that there are 10 active cells in it. As the majority of them results from under segmentation, 3 active cells should be tracked by Kalman filter actually.

In Fig. 3 (a) ~ (e), it shows 1, 10, 20, 40 and 60 frames in image sequence III respectively, and the corresponding segmentation images. To active cell 1, it has moved out of the image in the 31st frame. So we can track it from the 1st frame and the 30th frame. It costs 17.125 seconds. For the active cell 2, we can track it from the 1st frame and the 50th frame. It costs 27.016 seconds. For the active cell 3, we can track it from the 9th frame and the 50th frame. It costs 22.563 seconds.

In Fig. 4(a), it shows the trajectory of active cell 1 in image sequence I which is labeled in black "1" to represent its initial position in the first frame. "." represents its centroid. "-" represents its tracking trajectory. below is same. In Fig. 4 (b), it shows the trajectory of active cell 2 in the same sequence which is labeled in black "2" to represent its position in the 46th frame. In Fig. 4(c), it is the active cell's tracking trajectory in image sequence II. We use a white square to represent the cells' starting position in the first frame. In Fig. 4(d), it is the three active cells' trajectories in image sequence III. We use black "1", "2" and "3" to label the cells' starting positions in the 9th frame respectively.

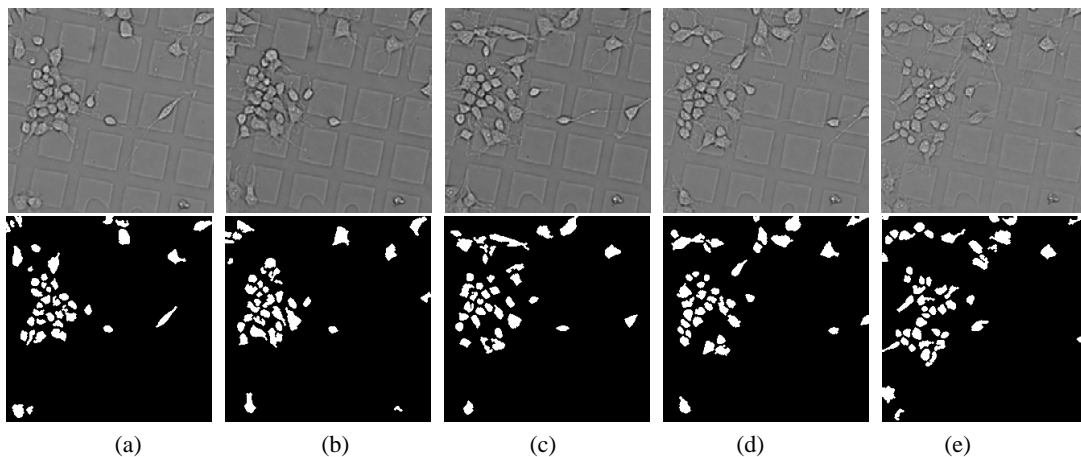


Fig. 3 Segmentation of frame 1、10、20、40 and 60 in sequence III

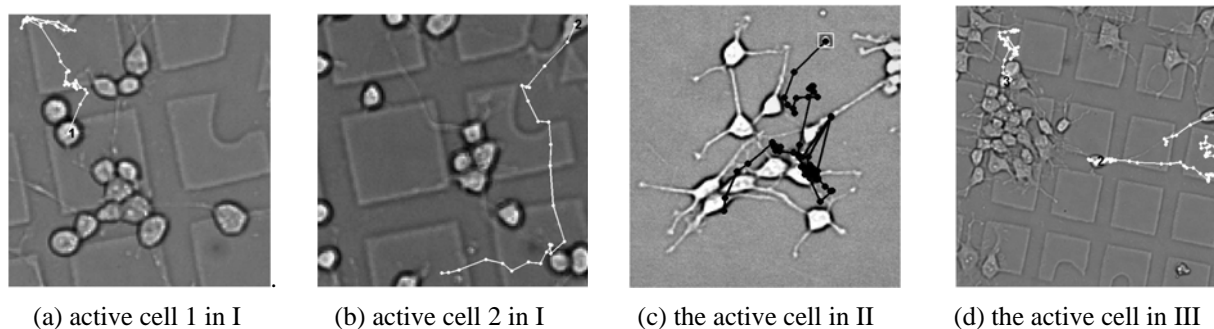


Fig. 4 Cells' trajectory in three image sequences

5 Conclusions

The curves shown in Fig. 5 are the best estimate in x and y directions of the active cells 1 and 2 in image sequence I via Kalman filters which are compared with the actual observation values of the two cells. Fig. 6 shows the two pairs of comparison curves of active cells in the image sequence II. Fig. 7 shows the three pairs of curves of the active cell 1, 2 and 3 in image sequence III. From Fig. 5-7, we can see that the differences

between the best estimate values and the actual observed values are less than 10 pixels both in x and y directions, which satisfies the local search criteria.

Aimed at tracking the active cells in confocal microscopy image sequences, this paper has proposed a tracking algorithm based on active cells' characteristics via Kalman filter. The cost function may reflect individual differences among the active cells. The results show that Kalman filter can track the active cells if we preset some appropriate parameters in cost function.

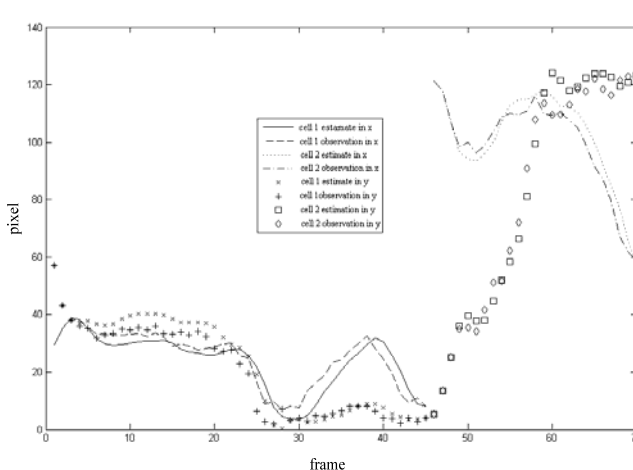


Fig. 5 Kalman filter estimating curve and the observing curve in image sequence I

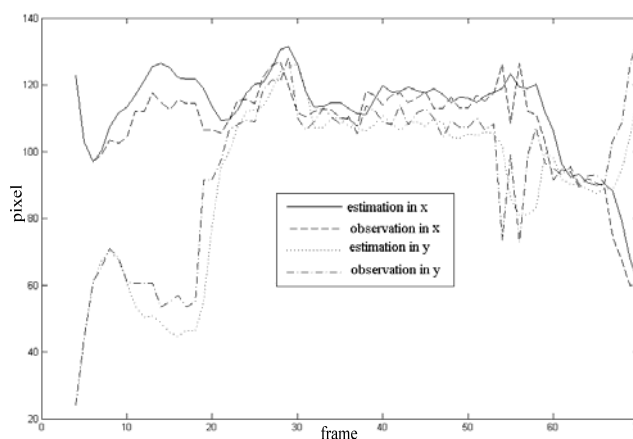


Fig. 6 Kalman filter estimating curve and the observing curve in image sequence II

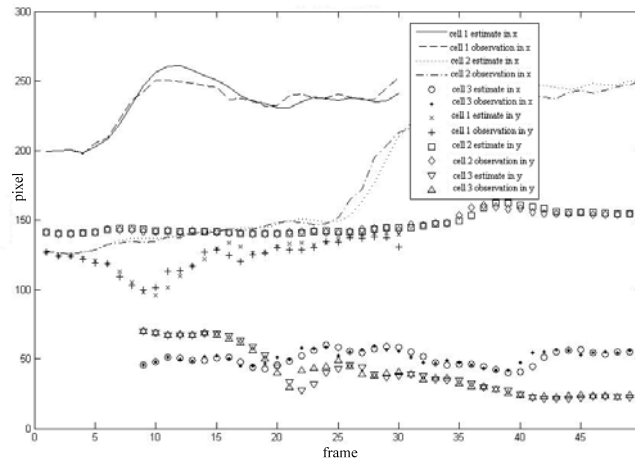


Fig. 7 Kalman filter estimating curve and the observing curve

6 Acknowledgement

This project is supported by Nature Science Foundation of China (NSFC). The grant number is 60875020. The authors would like to thank the DSS of Chalmers University of Technology for recording the stem cell image sequences.

7 References

- [1] Yuan Xiaohu, Cui Yan. Cell Tracking with Image Algorithm [J]. Chinese Journal of Biomedical Engineering, 2008, 27(3): 393-399
- [2] Tang Chunming, Zhao Chunhui. Automatic Segmentation and Tracking of Neural Stem Cells in Sequential Digital Images [J]. Chinese Journal of Biomedical Engineering, 2006, 25(1): 46-50
- [3] Yang X., Li H. and Zhou X. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy [J]. IEEE Transactions on Circuits and Systems, Nov 2006, 53(11), 2405-2414
- [4] Debeir O, Van-Ham P, Kiss R., et al. Tracking of migrating cells under phase-contrast video microscopy with combined Mean-shift processes [J]. IEEE Trans on Medical Imaging, 2005, 24(6): 697-712
- [5] Zimmer C, Labruyere, Meas-Yedid V, et al. Segmentation and tracking of migrating cells in video microscopy with parametric active contours: a tool for cell-based drug testing [J]. IEEE Trans on Med Image, 2002, 21(10): 1212-1221
- [6] Chunming Tang and Ewert Bengtsson. Segmentation and Tracking of Neural Stem Cell [J]. ICIC 2005, 851-859
- [7] Althoff, K., Degerman, J. and Gustavsson, T. Tracking Neural Stem Cells in Time-Lapse Microscopy Image Sequences [J]. SPIE 2005, 218, 1883-1891
- [8] Song Yingchun. Research on Kalman Filter in Kinematic Positioning [D]. ChangSha: Central South University, 2007
- [9] Chunming Tang, Dongbin Xu and Ling Ma. Segmentation of image sequences of neuron stem cells based on level-set algorithm combined with local gray threshold. [J]. Journal of Computer Aided design and Graphics. 2010, 22(8): 1279-1285
- [10] Donggang Yu, Tuan D. Pham, Xiaobo Zhou. Detection and Analysis of Cell Nuclear Phases. KES (1) 2008: 401-408

An Efficient Soft Graph Clustering Method for PPI Networks based on Purifying and Filtering the Coupling Matrix

Ying Liu¹, Amir Foroushani²

¹Department of Mathematics and Information Sciences, University of North Texas at Dallas, Dallas, TX

²Department of Molecular Biology and Biochemistry, 8888 University Drive, Simon Fraser University, Burnaby, British Columbia

Abstract - One of the most pressing problems of the post genomic era is identifying protein functions. Clustering Protein-Protein-Interaction networks is a systems biological approach to this problem. Traditional Graph Clustering Methods are crisp, and allow only membership of each node in at most one cluster. However, most real world networks contain overlapping clusters. Recently the need for scalable, accurate and efficient overlapping graph clustering methods has been recognized and various soft (overlapping) graph clustering methods have been proposed. In this paper, an efficient, novel, and fast overlapping clustering method is proposed based on purifying and filtering the coupling matrix (PFC). PFC is tested on PPI networks. The experimental results show that PFC method outperforms many existing methods by a few orders of magnitude in terms of average statistical (hypergeometrical) confidence regarding biological enrichment of the identified clusters.

Keywords: Protein-Protein Interaction networks; Graph Clustering; Overlapping functional modules; Coupling Matrix; Systems biology

1 Introduction

Homology based approaches have been the traditional bioinformatics approach to the problem of protein function identification. Variations of tools like BLAST [1] and Clustal [2] and concepts like COGs (Clusters of orthologous Groups) [3] have been applied to infer the function of a protein or the encoding gene from the known a closely related gene or protein in a closely related species. Although very useful, this approach has some serious limitations. For many proteins, no characterized homologs exist. Furthermore, form does not always determine function, and the closest hit returned by heuristic oriented sequence alignment tools is not always the closest relative or the best functional counterpart. Phenomena like Horizontal Gene Transfer complicate matters additionally. Last but not least, most biological Functions are achieved by collaboration of many different proteins and a proteins function is often context sensitive, depending on presence or absence of certain interaction partners.

A Systems Biology Approach to the problem aims at identifying functional modules (groups of closely cooperating and physically interacting cellular components that achieve a common biological function) or protein complexes by identifying network communities (groups of densely connected nodes in PPI networks). This involves clustering of PPI-networks as a main step. Once communities are detected, a hypergeometrical p-value is computed for each cluster and each biological function to evaluate the biological relevance of the clusters. Research on network clustering has focused for the most part on crisp clustering. However, many real world functional modules overlap. The present paper introduces a new simple soft clustering method for which the biological enrichment of the identified clusters seem to have in average somewhat better confidence values than current soft clustering methods.

2 Previous Work

Examples for crisp clustering methods include HCS [4], RNSC [5] and SPC [6]. More recently, soft or overlapping network clustering methods have evolved. The importance of soft clustering methods was first discussed in [7], the same group of authors also developed one of the first soft clustering algorithms for soft clustering, Clique Percolation Method or CPM [8]. An implementation of CPM, called CFinder [9] is available online. The CPM approach is basically based on the “defective cliques” idea and has received some much deserved attention. Another soft clustering tool is Chinese Whisper [10] with origins in Natural Language Processing. According to its author, CW can be seen as a special case of the Random Walks based method Markov-Chain-Clustering (MCL) [11] with an aggressive pruning strategy.

Recently, some authors [12, 13] have proposed and implemented betweenness based [14] Clustering (NG) method, which makes NG’s divisive hierarchical approach capable of identifying overlapping clusters. NG’s method finds communities by edge removal. The modifications involve node removal or node splitting. The decisions about which edges to remove and which nodes to split, are based on iterated all pair shortest path calculations.

In this paper, we present a new approach, called PFC, which is based on the notion of Coupling matrix (or common neighbors). In the rest of the paper, we first describe PFC and compare its results with the best results achieved by the aforementioned soft approaches. The second part of this work aims to illustrate the biological relevance of soft methods by giving several examples of how the biological functions of overlap nodes relate to biological functions of respective clusters.

3 PFC Method

The method introduced here is based on the purification and filtering of coupling matrix, PFC. PFC is a soft graph clustering method that involves only a few matrix multiplications/ manipulation. Our experimental results show that it outperforms the above mentioned methods in terms of the p-values for MIPS functional enrichment [15] of the identified clusters. The PPI net works we used in the paper are yeast PPI networks (4873 proteins and 17200 interactions).

3.1 Coupling Matrix

Bibliographical coupling is an idea from text classification: If two documents (for example two scientific papers) share a significant number of cited references, they are likely to deal with similar topics. A coupling matrix in a network describes the number of shared neighbors (or paths of length two) for each node pair. For undirected graphs like PPI networks, this matrix is symmetric and can be easily obtained from the original adjacency matrix A by: $B = A * A$. Notably, for second degree neighbors, the entry in coupling matrix is nonzero, even if there is no edge between the nodes. The importance of second degree neighbors in PPI networks has been emphasized before in the literature. For example: [16] note that "A substantial number of proteins are observed to share functions with level-2 neighbors but not with level-1 neighbors."

3.2 Purification of the Coupling Matrix

Adjacency matrices of biological networks are in general very sparse. The coupling matrix described above is slightly denser. However, not all nonzero-values are equally valuable. In the purification step, we determine the number of nonzero values (in unweighted graphs like PPI-Networks, this corresponds to the row sum), the maximum entry and the minimum non-zero value for each line of the coupling matrix. Rows in which the minimum nonzero entry and the maximum value are relatively close are considered homogenous and left unchanged. For other rows, we delete nonzero entries that don't make a significant contribution to the row sum. The Purification Process is summarized below:

FOREACH row i of the Coupling Matrix B

IF $\min(B(i,:)) < \lfloor \max(B(i,:)) * \alpha \rfloor$

THEN $B(i,:) = \lfloor B(i,:) ./ (\text{Bavg}(i) * \beta) \rfloor$

Where: $./$ is the Matlab cell wise division operator, $\lfloor \rfloor$ is the basic floor operation and α and β are values less than and greater than 1 respectively.

This purification step is robust in regard to choice of values for its parameters. In particular in our experiment with a yeast PPI network, the results for $\alpha = 0.8$ and $\beta = 1.2$ did not differ from those for $\alpha = 0.7$ and $\beta = 1.3$.

3.3 Filtering of the purified coupling matrix

The set of nonzero entries in each line of the Purified Coupling matrix can be considered as a candidate cluster. For a network of n nodes, this generally means n candidate clusters. However, not all rows are equally interesting. The set of nonzero entries (the information content) of many rows is likely to be very similar to, or contained largely within the sets of nonzero entries of other rows. This means that many rows are likely to represent spurious or redundant clusters. In the filtering step, we address this problem and try to select the most relevant and interesting rows of the purified coupling matrix. The set of nonzero entries in each of the selected lines of the purified coupling matrix represent our final clusters. The filtering step of PFC is a flexible step. Two alternative filtering approaches are discussed below.

3.4 Filtering by Simple, Local Criteria

The first Filtering approach is motivated by assumptions about the nature of the data and size of the target clusters. PPI data are for the most part results of high throughput experiments like yeast two hybrid and are known to contain many false positive and many false negative entries. For certain, more thoroughly studied parts of the network, additional data might be available from small scale, more accurate experiments. In PFC, the emphasis lies on common second degree neighbors and this can magnify the effects of noise. Under the assumption that Nodes with low degree belong in general to the less thoroughly examined parts of the network, it is conceivable that the current data for the graph around these low nodes contains many missing links. Missing links in these areas can have dramatic effects on the constellation of second degree neighbors. This means the Coupling data for low degree nodes is particularly unreliable. On the other hand, many extremely well connected nodes are known to be central hubs that in general help to connect many nodes of very different functionality with each other, hence, their second degree neighbors comprise huge sets that are less likely to be all functionally related. Additionally, it has been shown that most functional modules are meso-scale [6]. There are also some fundamental physical constraints on the size and shape of a protein complex that make very large modules unlikely. Taking these considerations into account, a filter is easily constructed by the following rules:

Discard all clusters (rows of purified coupling matrix) where the labeling node (the $_th$ node in the $_th$ row) has a particularly low (< 14) or particularly high (>30) degree. Discard all clusters where the module size is too small (<35) or particularly large (>65).

The selected minimum and maximum values for degree of labeling nodes and module size are heuristically motivated. The intervals can be easily changed to obtain or discard more clusters, but the enrichment results for these intervals seem reasonably good. The peak log value for the enrichment of selected clusters is at -91.00 and the average lies at -18.99 . Using this filter, by clustering yeast PPI networks, PFC yields 151 clusters from 52 different Functional categories. Figure 1 gives an example.

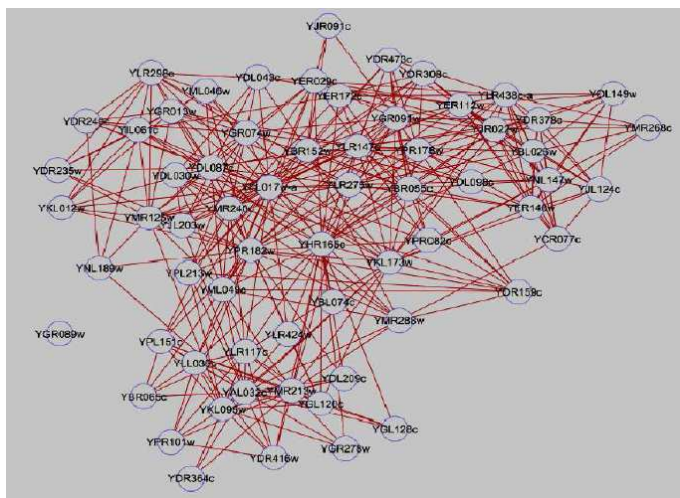


Figure 1 This Figure shows the community for the row labeled “YKL173w” in the purified coupling matrix of yeast PPI network. It is one of the clustered selected by PFC1. Out of the 63 proteins in this community, 58 belong to MIPS Funcat 11.04.03.01.

4 Experimental Results and Discussions

The results of the PFC are compared with results obtained by other soft clustering methods. A PPI network of yeast with 4873 Nodes and 17200 edges is used as the test data set. The other methods are an in-house implementation of Pinney and Westhead’s Betweenness Based proposal [12], Chinese Whisper [10], CPM as implemented in C-Finder [9]. Whenever other methods needed additional input parameters, we tried to choose parameters that gave the best values. The results from different methods are summarized in Table 1.

4.1 Biological Functions of Overlap Nodes

The hypergeometric evaluation of individual clusters is the main pillar in assessing the quality of crisp clustering

methods. For soft clustering methods, further interesting questions arise that deal with relationships between clusters. A possible conceptual disadvantage, production of widely overlapping, redundant clusters was addressed in previous sections. Figure 2 is a clustering results of the PFC. The result demonstrates an important *advantage* of soft methods against crisp ones: They show how soft clustering can adequately mirror the fact that many proteins have context dependent functions, and how in some cases overlap nodes can act as functional bridges between different modules.

Table 1 Comparison of results from different methods

Method	Cluster Count	Average Cluster Size	Average Enrichment	Network Coverage	Diversity
Betweenness based	20	302.70	-15.11	0.58	19/20
Chinese Whisper	38	23.45	-12.11	0.17	32/38
C Finder	68	14.50	-15.70	0.19	48/68
PFC	183	44.76	-19.35	0.31	55/183

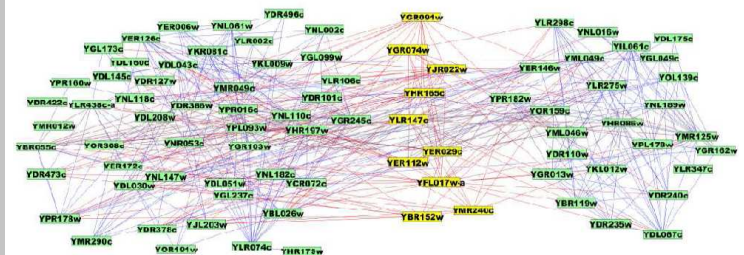


Figure 2. result #1: There is a relatively large overlap (yellow nodes). All 10 overlap nodes are involved in “nuclear mRNA splicing, via spliceosome-A”. The same is true for ca.25% (12 out of 45) of the green nodes to the left and 68% (17 out of 25) of the green nodes to the right of the overlap. Furthermore, two of the overlap nodes are also involved in spliceosome assembly the total number of such nodes in the entire network is 19.

5 Conclusions

This paper introduced PFC, a new clustering concept based on purification and filtering of a coupling (common neighbor) matrix. It discussed a very different filtering method. PFC consists of only a few matrix multiplications and manipulations and is therefore very efficient. The PFC outperforms current soft clustering methods on PPI networks by a few orders of magnitude in terms of average statistical confidence on biological enrichment of the identified clusters. The paper illustrated the importance of soft clustering methods in systems biology by giving a few concrete examples of how the biological function of the overlap nodes relates to the functions of the respective clusters.

6 References

- [1] Altschul, SF, et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic acids research* 25, no. 17: 3389, 1997.
- [2] Thompson, JD, DG Higgins, and TJ Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic acids research* 22, no. 22: 4673-4680, 1994
- [3] Tatusov, R. L., E. V. Koonin, and D. J. Lipman. "A genomic perspective on protein families". *Science* 278, no. 5338: 631, 1997.
- [4] Hartuv, E., R. Shamir. "A clustering algorithm based on graph connectivity". *Information processing letters* 76, no. 4-6: 175-181, 2000.
- [5] King, A. D., N. Przulj, and I. Jurisica. "Protein complex prediction via cost-based clustering". *Bioinformatics* 20,: 3013-3020, 2004.
- [6] Spirin, V., L. A. Mirny. "Protein complexes and functional modules in molecular networks". *Proceedings of the National Academy of Sciences* 100, no. 21: 12123-12128, 2003.
- [7] Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society". *Nature* 435, no. 7043 (Jun 9): 814-818, 2005.
- [8] Derenyi, I., et al. "Clique percolation in random networks". *Physical Review Letters* 94, no. 16: 160202, 2005.
- [9] Adamcsek, B., G. et al. "CFinder: locating cliques and overlapping modules in biological networks". *Bioinformatics* 22, no. 8: 1021-1023, 2006.
- [10] Biemann, C. "Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems". In *Proceedings of the HLT-NAACL-06 workshop on textgraphs-06*, new york, USA, 2006.
- [11] Van Dongen, S. "A cluster algorithm for graphs". *Report- Information systems* , no. 10: 1-40, 2000.
- [12] Pinney, J. W., D. R. Westhead. "Betweenness-based decomposition methods for social and biological networks". In *Interdisciplinary statistics and bioinformatics*. Edited by S. Barber, P. D. Baxter, K. V. Mardia and R. E. Walls. Leeds University Press, 2000.
- [13] Gregory, S. "An algorithm to find overlapping community structure in networks". *Lecture Notes in Computer Science* 4702: 91, 2007.
- [14] Girvan, M., M. E. Newman. "Community structure in social and biological networks". *PNAS* 99: 7821-7826, 2002.
- [16] Chua, H. N. et al. "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions". *Bioinformatics* 22: 1623-1630, 2006.
- [15] MIPS. The functional catalogue (FunCat). 2007. <<http://mips.gsf.de/projects/funecat>>.

BRATUMASS Antenna Positioning Optimization with Genetic Algorithm

Zhongling Han¹, Zhifu Tao¹, Meng Yao^{1*}, Yizhou Yao²,
Blair Fleet³, Erik D. Goodman³, and John R. Deller⁴

¹ Institute of information science and technology East China Normal University, Shanghai, China

² Weiyu high school, Shanghai, China

³ BEACON Michigan State University, East Lansing, MI

⁴ ECE Michigan State University, East Lansing, MI

*Corresponding Author, e-mail: myao@ee.ecnu.edu.cn

Abstract - Using the difference of dielectric constant between malignant tumor tissue and normal breast tissue, BRATUMASS (Breast tumor microwave sensor system [1]) can determine the detected target properties by analyzing the properties of target tissue back wave obtained by near-field microwave radiation. The practical experiments show that the target space and records of antenna corresponding position displacement when the antenna close contact with skin tissue will be changed, which might lower the quality of the inversion imaging of result. So, the target space characteristic data is introduced in order to eliminate the effect of the displacement. This paper presents a method of antenna relative position placement optimization using genetic algorithm and performs its feasibility with optimized examples.

Keywords: BRATUMASS, Genetic Algorithm, Pauta criterion, Antenna placing position

1 Introduction

A microwave reflecting interface will be formed between malignant and normal breast tissue for their different dielectric constants. BRATUMASS can use this property to locate the position of reflecting interface and the character of two tissues by analyzing the back wave[2]. During the process of BRATUMASS, displacement position of the antenna will directly affect the measurement results, thus antenna positioning will directly affect the location accuracy and characteristics of the target. We use a simple genetic algorithm to search the position of antenna, and give the optimizing result of example data with the consistency of statistical data.

2 BRATUMSS detecting principle and boundary adjusting

Detection target space of BRATUMSS is a special space. Detecting antenna distributes in the margin area of breast shape to capture the testing data. Breast shape is not fixed for

different individuals, persons have different breast boundaries. This will lead to biggish error in locating area of cancerous tissue.

Using typical radar correlation detection technology, BRATUMASS extracts frequency difference between back wave signal and transmitting signal to ensure the distance from reflecting interface to antenna. BRATUMASS signal is defined as:

$$V_1 = A_c \cos(2 \times \pi i \times fc + 2 \times Kc \times \int_i H(\tau) d\tau + \theta) \quad (1)$$

where, $H(\tau) = \text{sawtooth}(\tau)$ is the triangular pulse.

The different dielectric constants between cancerous tissue and normal tissue will form a dielectric constant mutation layer. The incident signals will produce backscattering in the layer. The backscattering signal received by BRATUMASS with transmitting signal directly seeks difference frequency. The intermediate frequency beat signal is the propagation delay from receiving antenna surface to different tissues interfaces. We can calculate the distance with propagation delay and the different dielectric constants^[3].

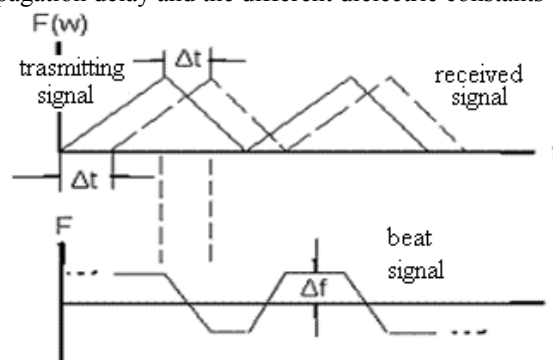


Figure 1. The schematic diagram of transmitting signal, received signal and beat signal

The propagation delay is relative to the position of the antenna. The delay can be converted into distance, which regards the position of the antenna instant position as the reference center. If the center moves, the whole test data will be changed, as shown in Figure 2.

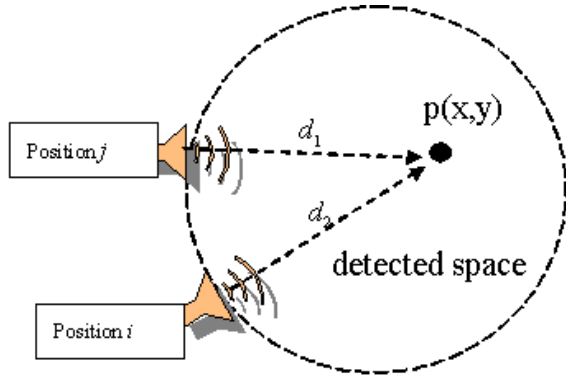


Figure 2. The BRATUMASS detecting schematic diagram

The point $P(x,y)$ is a target, the characteristic data^[4] $f_i(x,y)$ and $f_j(x,y)$ are obtained in the position i and position j , respectively. Suppose the characteristic value at point P is an invariant. A set of characteristic values at point P will be obtained from several measurements after sampling N times at the boundary. According to the pauta criterion of measurement error theory, repeated measurement data should satisfy:

$$|f_i(x,y) - \overline{f(x,y)}| < 3\sigma, (i=1\dots N) \quad (2)$$

Where, σ is the variance of measurement data.

When the change of the i -th antenna position make the whole i -th data exceeding the range of (2), the current antenna position should be adjusted to update the space target data. There must have one antenna position which makes all spatial units corresponding with (2) for the calculation over the whole space.

3 Antenna position optimization with Genetic algorithm

The traditional method, we usually scalarize the multiple objectives into a single objective by averaging the objectives with a certain weight vector in solving multi-objective optimization problem. In these cases, the obtained solution is highly sensitive to the weight vector used in the scalarization process. Moreover, the user should have knowledge about the underlying problem. Designers may be interested in a set of Pareto-optimal points instead of a single point. Since genetic algorithms work on a population of points, it is natural to be used in multi-objective optimization problems to capture a number of solutions simultaneously.

Let the initial position of each antenna is the initial value, the shape deformation of breast is 10 mm. The change of i -th point is denoted by δ_i , and the change of the corresponding coordinate (x_i, y_i) is $(x_i + \Delta x_i, y_i + \Delta y_i)$.

3.1 Encoding

Each change of coordinates is regarded as an encoded object whose code length is 10. The sampling number of cycle boundary is N , and the change of each coordinates are Δx_i and Δy_i which need to be encoded. So, the length of chromosome is $2N$.

$$p_k = \left[\left| \Delta x_{1k} \right| \left| \Delta y_{1k} \right| \left| \Delta x_{2k} \right| \left| \Delta y_{2k} \right| \dots \left| \Delta x_{ik} \right| \left| \Delta y_{ik} \right| \dots \left| \Delta x_{Nk} \right| \left| \Delta y_{Nk} \right| \right] \quad (3)$$

3.2 Selection

The judgment basis of coordinate position is (2). Let $g_k(x,y)$ be the objective function of the k -th chromosome corresponding to $P(x,y)$.

$$g(P) = \frac{\sigma(f_1(P), f_2(P), \dots, f_i(P), \dots, f_N(P))}{E(f_1(P), f_2(P), \dots, f_i(P), \dots, f_N(P))} \quad (4)$$

Where, $f_i(P)$ is the characteristic value at point P obtained from the antenna at the i -th position. The variance is $\sigma(\cdot)$ and the mean is $E(\cdot)$.

$$g_k(P) = \frac{\sigma(f_1(P + \delta_1^k), f_2(P + \delta_2^k), \dots, f_i(P + \delta_i^k), \dots, f_N(P + \delta_N^k))}{E(f_1(P + \delta_1^k), f_2(P + \delta_2^k), \dots, f_i(P + \delta_i^k), \dots, f_N(P + \delta_N^k))} \quad (5)$$

Where, δ_i^k is the change of antenna position at the i -th position which has $\delta_i^k = (\Delta x_i^k, \Delta y_i^k)$.

The objective function of the k -th chromosome corresponding to the whole space is $g_k(\Omega)$.

$$g_k(\Omega) = \iint_{\Omega} g_k(x,y) dx dy \quad (6)$$

The fitness function is $fit_k(\Omega) = 1/g_k(\Omega)$ when $g_k(\Omega)$ is minimum.

3.3 Mutation

First, randomly select chromosome's individual. Second, randomly select several bits from the $2N$ chromosomes. Then carry out mutation to these selected chromosome bit according to mutation probability

3.4 Program flow

① Population initialization: Initial population is composed of N chromosomes which are randomly generated according to (3).

② Initialization parameters: Let P_m be the mutation probability, fitness goals and the maximum number of iterations.

③ Solution space transform: Each chromosome represents the approximate solution, $[-1, 1]^{1 \times 2N}$ is mapped into the parameters space $[-10, 10]^{1 \times 2N}$.

④ Determine the fitness function and calculation. Calculate each chromosome's objective function according to (5) and then calculate fitness function.

⑤ Assigning fitness to each chromosome and progressing genetic operation

⑥ In accordance with the mutation probability, one chromosome are selected randomly from the population. Progress the selected chromosome mutation according to supposed probability.

⑦ Check the number of iteration. End the process if iteration exceed, else return to step ④.

4 Experiments and results

As shown in Figure 3(a), a rectangle plate, with size 0.2 mm*20 mm*40 mm, was placed into a piece of boneless pork with size 94 mm*80 mm*50 mm. The objective spatial characteristic at the beginning is illustrated in Figure 3(b). Figure 3(c) illustrates the objective spatial characteristic after 1000 times' genetic operation. Figure 3(d) demonstrates the trend of objective function in the 1000 times genetic process. In figure 3(a) (b) (c), the red mark points are the initial placement position of the antenna. And, in Figure 3(b), the black points are randomly assigned positions, in figure 3(c), the black points are searched positions of the antenna after GA optimization processing result.

For different materials which placed in the pork, we can gain the similar results which are shown in Figure 4.

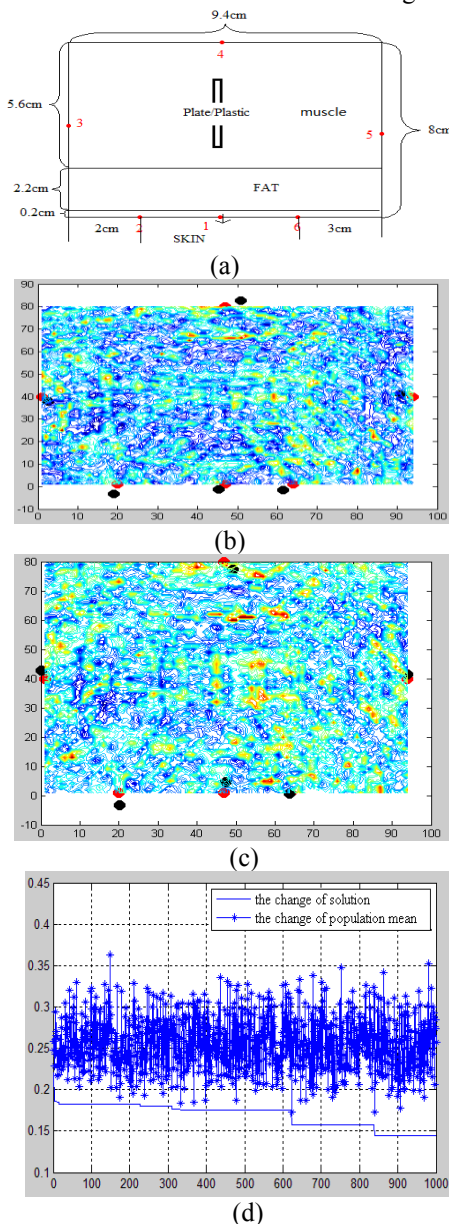


Figure3. (a) Experimental object diagram. (b)The objective spatial characteristic at the beginning is shown, Scales are in mm. The

objective spatial characteristic after 1000 times' genetic operation is shown in (c). Scales are in mm. The trend of objective function in the 1000 times' genetic process is shown in (d).

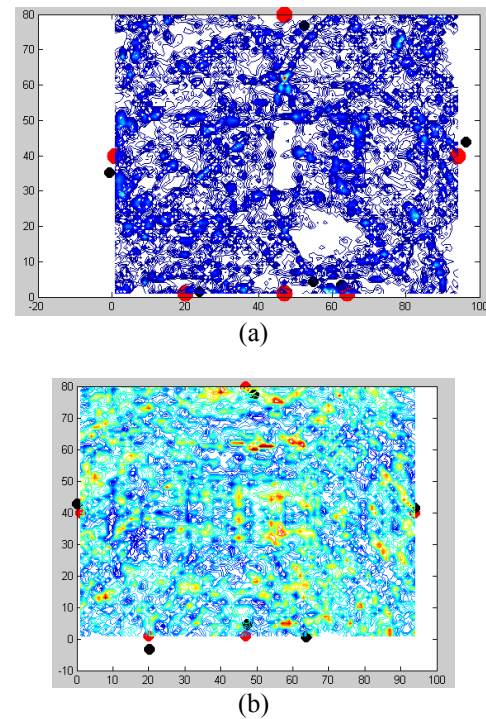


Figure4. It is the objective spatial characteristic placing different objects. Place a plastic plate (0.5mm*20 mm*40 mm) and a metal plate (0.2 mm*20 mm*40 mm) in (a) and (b) respectively. The placing position is shown in figure 3(a). Scales are in mm.

5 Conclusion

In general measurement, it is very difficult to process for the unknown antenna changing rules. Therefore, we use a simple genetic algorithm to reduce the impact of inversion imaging caused by uncertain of antenna's position. Experiment results showed that the simple genetic algorithm plays an important role in optimizing antenna's location in accurately and improve imaging effects. The simple genetic algorithm is time consuming to optimize the antenna position in real-time measurement process. Hence, it is employed to ensure the true antenna position offline, and it is essential for improving the imaging precision. Furthermore, the simple genetic algorithm can guide the placement region of the antenna position to obtain more information of imaging in practical detection process with BRATUMASS.

6 Acknowledgments

This work has been performed while Prof. M. Yao was a visiting scholar in Michigan State University, thanks to a visiting research program from Prof. Erik D. Goodman. M. Yao would also like to acknowledge the support of Shanghai Science and Technology Development Foundation under the project grant numbers 03JC14026 and 08JC1409200, as well

as the support of TI Co. Ltd through TI (China) Innovation Foundation.

7 Reference

- [1] Zhi-fu Tao, Xia-chen Dong, Meng Yao and Yi-zhou Yao. Biopsy Back Wave Preprocessing Research of BRATUMASS System based on Applications of Fractional Fourier Transform, Proceedings of The 2010 International Conference on Bioinformatics and Computational biology Vol.II pp.258--261
- [2] Meng Yao, Zhi-fu Tao and Qi-feng Pan. Application of Quantum Genetic Algorithm on Breast Tumor Detection with Microwave Imaging, GECCO 2009, 2685–2688
- [3] Zheng, S. 2006. Breast Tumor Imaging Method Investigation in UWB near-field microwave environment. Master Thesis, East China Normal University, 31-35
- [4] Zhifu Tao, Qifeng Pan, Meng Yao, and Ming Li. Reconstructing Microwave Near-Field Image Based on the Discrepancy of Radial Distribution of Dielectric Constant. ICCSA 2009, 717–728

Application of quarter Iteration of FRFT in BRATUMASS for Weak Signal Extraction

Zhongling Han¹, Zhifu Tao¹, Meng Yao^{1*}, Yizhou Yao²,
Blair Fleet³, Erik D. Goodman³, and John R. Deller⁴

¹ Institute of information science and technology East China Normal University, Shanghai, China

² Weiyu high school, Shanghai, China

³ BEACON Michigan State University, East Lansing, MI

⁴ ECE Michigan State University, East Lansing, MI

*Corresponding Author, e-mail: myao@ee.ecnu.edu.cn

Abstract - In this paper, we present a quarter iterative of FRFT algorithm to solve the problem, which is the extraction of weak signal from the back wave in BRATUMASS. [1] The energy of the same frequency in back wave add together and to discard the signal phase. The energy of the modulus as leading evolution function, the weak back signal in BRATUMASS can be separated out. Experiments showed that the quarter iterative of FRFT algorithm plays a bigger role in the extraction of the weak back signal.

Keywords: BRATUMASS, quarter iterative of FRFT, extraction of the weak back signal

1 Introduction

As an active microwave imaging system, BRATUMASS can image the breast tissues for their dielectric constants which have the obviously different between malignant tissue and normal tissue. The transceiver antenna is placed on the breast surface and transmits microwave to breast internal. Backscatter will be happened and the back wave signals will be produce when microwave signals meet different breast tissues. The properties of the detecting target can be obtained by analyzing the back wave. Thus, the main problem, for gaining the target characteristic and reconstructing of detecting space, lies in the extracting of the back wave. The antenna will receive two main kinds of signals: one is the back wave signal radiated from the antenna main lobe to the target; the other is radiated directly from the antenna side lobe to the receiver. In addition, the energy amplitude of the latter is larger than the former. That will make the BRATUMASS, a frequency related system, create abundant problems of fuzzy object and interference in receiver. Usually, clutter can be eliminated by filter; however, the filters inevitably reduce the useful information component, and also made a few small targets back wave loss. We propose a quarter iterative of FRFT to solve this problem. Experiments show that the method can effectively extract the small target from the weak back signal in BRATUMASS.

2 The signals of BRATUMASS

Detecting points are located in the surface of breast. The system uses transceiver antenna, which are shown in Figure 1D. Side lobe signal of the transceiver antenna is $f_p(t)$. The system noise is $N(t)$. The position of the detecting points and the structure of the antenna are shown in Figure 1^[2].

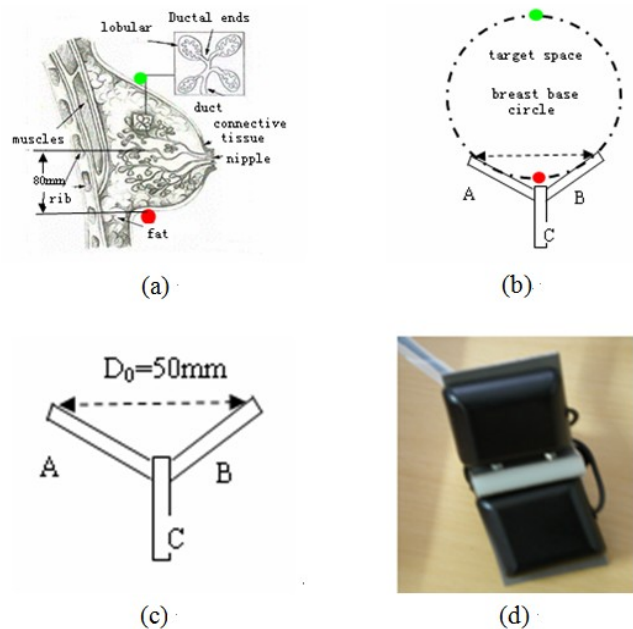


Figure 1. The position of the detecting points and the structure of the antenna, (a)The structure of breast. (b)The schematic of BRATUMASS detecting position, Red point is the position of antenna and green point is the position of metal slice in (a) and (b).(c) The schematic of transceiver antenna. A is Transmitting Antenna, B is Receiving Antenna and C is Center Clapboard. (d) The photo of transceiver antenna.

The transmitting signal of BRATUMASS is:

$$S(t) = \text{rect}\left(\frac{t}{T}\right) \exp\left\{j2\pi\left[f_0 t + \frac{1}{2}kt^2\right]\right\} \quad (1)$$

Where, $\text{rect}(t)$ is rectangular envelope; f_0 is initial frequency; T is time width; k is frequency modulation slope.

Side lobe signal of the transceiver antenna $f_p(t)$ is:

$$f_p(t) = \text{rect}\left(\frac{t-t_p}{T}\right) \exp\left\{j2\pi\left[f_0(t-t_p) + \frac{1}{2}k(t-t_p)^2\right]\right\} \quad (2)$$

Where, t_p is the transmission delay of side lobe signal.

The signal obtained by receiver antenna chiefly includes two parts: $N(t)$ and $f_p(t)$. Figure 2 illustrates the structure of frequency mixing. $S_f(t)$ is the output of mixer, $S_f(n)$ is obtained from A/D sampler.

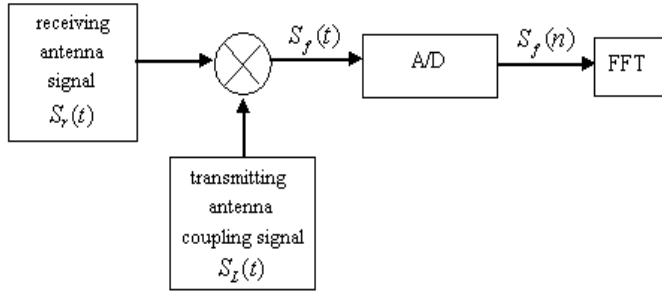


Figure 2. The structure of frequency mixing of BRATUMASS

The output of the mixer can be expressed as:

$$S_f(t) = S_r(t) \times S_L(t) \quad (3)$$

Where, $S_r(t)$ is receiving signal from transceiver antenna.

$S_L(t)$ is coupling signal from transmitting antenna which satisfy the requirements of frequency mixing of zero IF. Suppose γ is a coupling coefficient, τ_0 is coupling delay. The $S_L(t)$ is:

$$S_L(t) = S(t - \tau_0) \times \gamma \quad (4)$$

The state, which has no target in the detecting space, is S_0 , so the receiving signals include side lobe $f_p(t)$ and noise $N(t)$. The state, which only has one target in detecting space, is S_1 , so the receiving signals obtain side lobe $f_p(t)$, noise $N(t)$ and the back wave of the single target. By analogy, S_n represents the state which has N target in the detecting space.

In the state S_0 , equation (4) is substituted into equation (3):

$$S_f(t)|_{s_0} = \gamma \times S(t - \tau_0) \times (f_p(t) + N(t)) \quad (5)$$

$S_f(t)|_{s_0}$ is abbreviated to $S_f^0(t)$.

After Fourier transform, Equation (5) is changed to:

$$S_f^0(\omega) = A \times S(\omega) \otimes (F_p(\omega) + N(\omega)) \quad (6)$$

Where, $A = \frac{\gamma}{2\pi} \times \exp(j\omega\tau_0)$, $S(\omega)$, $F_p(\omega)$ and $N(\omega)$ is the

Fourier transform of $S(t)$, $f_p(t)$ and $N(t)$, respectively.

The sampling data of clinical case obtained by BRATUMASS is showed in Figure 3^[3]. The positions of detecting points are illustrated in Figure 1(b). The actual measure distance is about 150~170mm, for patient posture is not perpendicular between mental slice and base circle.

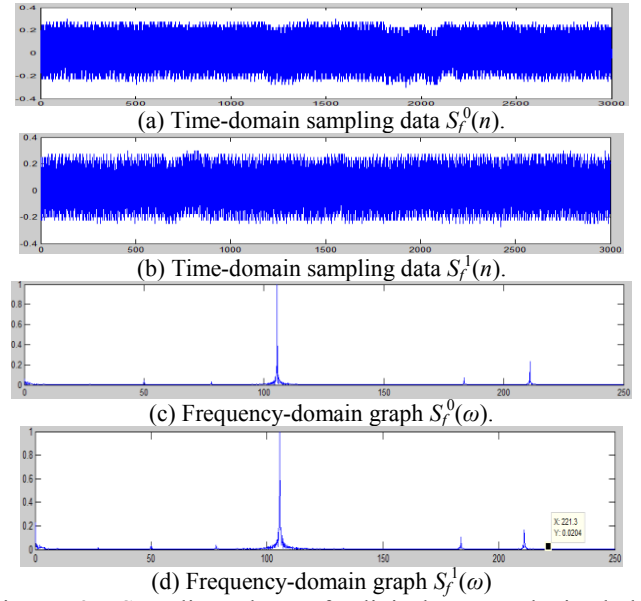


Figure 3. Sampling data of clinical case obtained by BRATUMASS. (a) The time-domain sampling data $S_f^0(n)$ when mental slice wasn't placed. (b) The time-domain sampling data $S_f^1(n)$ when breast surface placed a mental slice with radius 1cm. The abscissa is the sampling ordinal, and the ordinate is the voltage level in (a) and (b). (c) The frequency-domain graph $S_f^0(\omega)$ corresponding to $S_f^0(n)$. (d) The frequency-domain graph $S_f^1(\omega)$ corresponding to $S_f^1(n)$. The abscissa is the frequency, and the ordinate is the normalized amplitude in (c) and (d).

According to the above graphs, we can see that frequency spectrum $S_f^0(\omega)$ is similar to $S_f^1(\omega)$. In Figure 3(d), the position of mental slice is marked (x:221.3/y:0.0204), 221.3Hz corresponding to the distance 169.5 mm. The efficient information of the mental slice couldn't directly extract from the spectrum.

Theoretically, there are no any targets of back wave information in the state of S_0 . However, spectral lines about 105.7Hz and 210.6Hz always exist in the actual measurement. Their amplitude is higher than target of back wave signal and they also have the corresponding change with the change of objective and environment. Thus, elimination interference by filter is not properly. Consider the frequency characteristic of $S_f^0(\omega)$ in (6), quarter iteration of FRFT algorithm is presented to enlarge amplitude of object spectrum distribution in this paper.

3 Quarter of iteration of FRFT algorithm and the signal processing

3.1 Quarter of iteration of FRFT algorithm

$g(t)$ and $G(\omega)$ as a Fourier transform pair, the relationship can be written by

$$G(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} g(t) e^{-j\omega t} dt \quad (7)$$

$$g(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} G(\omega) e^{j\omega t} d\omega \quad (8)$$

Following, brief write down for $G = F(g(t))$.
 $F(F(g(t))) = F^2(g(t)) = g(-t)$;

$F^4(g(t)) = g(t)$. F^n indicates that operator is used N times.

$F(\omega)$ is the Fourier transform of $f(t)$, It has the following properties: $F(\omega) = \mathfrak{F}(f(t))$;

$\mathfrak{F}(F(\omega)) = \mathfrak{F}(\mathfrak{F}(f(t))) = f(-t)$. So repeat four times is periodic repeat.

From the perspective of nonlinear dynamics, the iteration of Fourier transform is an iterated function with period 4. Use the modular \mathfrak{R} as iterative evolution function, for accumulating the same frequency during iterative evolution and discard the phase of signal. Then,

$$|F(\omega)| = \mathfrak{R}(\mathfrak{F}(f(t))) \quad (9)$$

$$F^{(2)} = \mathfrak{R}(\mathfrak{F}(\mathfrak{R}(\mathfrak{F}(f(t)))))) \quad (10)$$

$$F^{(3)} = \mathfrak{R}(\mathfrak{F}(\mathfrak{R}(\mathfrak{F}(\mathfrak{R}(\mathfrak{F}(f(t))))))) \quad (11)$$

.....

$$F^{(n)} = \mathfrak{R}(\mathfrak{F}(\dots \mathfrak{R}(\mathfrak{F}(\mathfrak{R}(\mathfrak{F}(\mathfrak{R}(\mathfrak{F}(\mathfrak{F}(f(t)))))))))) \dots)) \quad (12)$$

After iterate N times, sorting frequency spectrum is obtained. Difference spectrum can be given as:

$$F_{di} = F^{(1)} - F^{(n)} \quad (13)$$

3.2 The influences to sinusoidal signal structure by quarter iteration of FRFT

Consider a signal $f(t) = A \sin(\omega t + \theta)$, where $\omega = 1.575\text{GHz}$, θ takes random value between $-\pi$ and π . A might as well be valued 100, sampling frequency is $10 \times \omega$ and iteration number $n = 4$.

Fig.4 shows the processing of quarter iteration of FRFT. The iterative result in $N=4$ is the high order spectrum of signal. Consequently, the location of high order frequency in the spectrum is given.

3.3 Processing results comparison between Sf0(n) and Sf1(n)

Compare fig.3 $S_f^0(\omega)$ and $S_f^1(\omega)$, Figure 5 demonstrates the processing result using the quarter iteration of FRFT algorithm. For the sake of convenience, the abscissa is transformed into distance which is corresponding to frequency, with unit mm. The ordinate is normalized amplitude. Figure 5(a) illustrates spectrum of the algorithm processing on signal S^0 . From (b), the back wave is saw clearly at 150~170 mm, which is corresponded to frequency 221.3Hz.

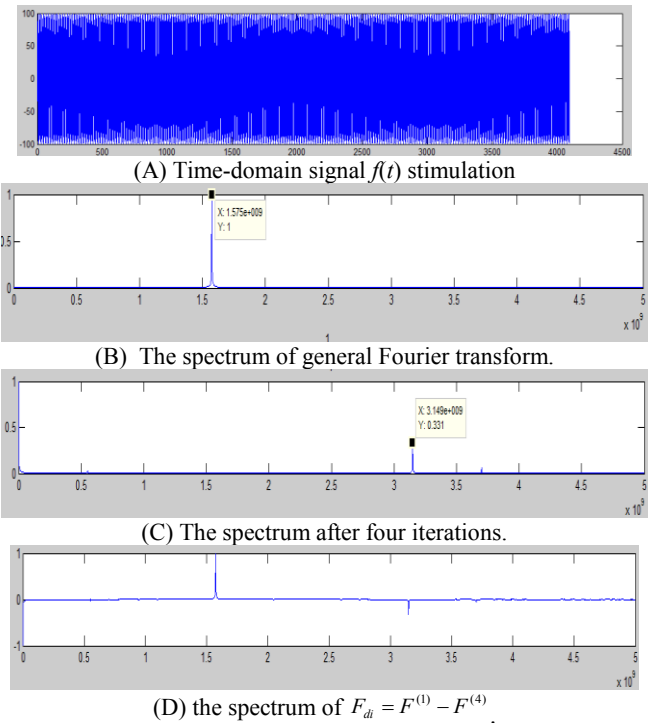


Figure4. Sinusoidal signal structure by quarter iteration of FRFT (a) The simulation signal $f(t)$. Its sampling depth is 4096. (b) The spectrum of normal Fourier transform, with $N=1$. (c) The spectrum after four iterations, with $N=4$. (d) The spectrum of $F_{di} = F^{(1)} - F^{(4)}$, the abscissa is frequency and the ordinate is normalized amplitude.

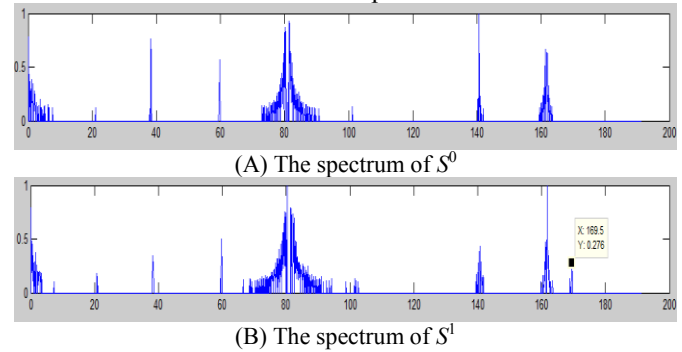


Figure5. The result using quarter iteration of FRFT algorithm (a) and (b) are the spectrums of S^0 and S^1 after separating by quarter iteration of FRFT algorithm, respectively.

4 Conclusions

The back wave spectrum amplitude is very small (see Figure 3(d)) in BRATUMASS. The amplitude of the back wave of the mental slice is only 0.0204, which is couldn't extract from the back wave. Nevertheless, invalid signal amplitude is high. (The system noise $N(t)$ and side lobe signal $f_p(t)$ are related to the frequency 105Hz and 210Hz, respectively. The amplitude of this two spectral lines is relatively large, and the amplitude of 105Hz spectral line even have verge on 1). According to signal characteristics, we proposed a method of weak signal separation in this paper. The amplitude of weak signal could be enhanced under the

condition of reducing the loss of frequency as much as possible. In Figure 5(b), the amplitude of metal slice back wave is change from 0.0204 to 0.207(10dB). This algorithm is only one of the methods of the weak signal extraction in the preprocessing on BRATUMASS. For determine breast tissue properties completely, energy information extraction from back wave have to be further researched. Furthermore, target tissue properties are acquired. The problem on how to get character distribution in detection space by combining information, namely space image inversion^[4] will be proposed in the following articles.

5 Acknowledgments

This work has been performed while Prof. M. Yao was a visiting scholar in Michigan State University, thanks to a visiting research program from Prof. Erik D. Goodman. M. Yao would also like to acknowledge the support of Shanghai Science and Technology Development Foundation under the project grant numbers 03JC14026 and 08JC1409200, as well as the support of TI Co. Ltd through TI (China) Innovation Foundation.

6 Reference

- 1.Zhi-fu Tao, Xia-chen Dong, Meng Yao and Yi-zhou Yao. Biopsy Back Wave Preprocessing Research of BRATUMASS System based on Applications of Fractional Fourier Transform, Proceedings of The 2010 International Conference on Bioinformatics and Computational biology Vol.II pp.258--261
- 2.Meng Yao, Zhi-fu Tao and Qi-feng Pan. Application of Quantum Genetic Algorithm on Breast Tumor Detection with Microwave Imaging, GECCO 2009, 2685–2688
- 3.Zheng, S. 2006. Breast Tumor Imaging Method Investigation in UWB near-field microwave environment. Master Thesis, East China Normal University, 34-35
4. Zhifu Tao, Qifeng Pan, Meng Yao, and Ming Li. Reconstructing Microwave Near-Field Image Based on the Discrepancy of Radial Distribution of Dielectric Constant. ICCSA 2009, 717–728

AN X-RAY ON METHODS AIMING AT ARRHYTHMIA CLASSIFICATION IN ECG SIGNALS

E. Luz and D. Menotti

Department of Computing - Universidade Federal de Ouro Preto
Campus Universitário, Ouro Preto, Brazil
{eduluz,menottid}@gmail.com

ABSTRACT

Arrhythmias (*i.e.*, irregular cardiac beat) classification in electrocardiogram (ECG) signals consists in an important issue for heart disease diagnosis due to the non-invasive nature of the ECG exam. In this paper, we present an X-ray, a generic view, on methods aiming at arrhythmia classification in ECG signals, which starts with signal preprocessing, and then segmentation of each heartbeat and so before classification, the feature extraction step. We also analyze and criticize the results of some arrhythmia classification methods present in the literature in terms of how the samples are chosen for train/test the classifier and the impact of this choice in their accuracies/sensitivities.

1. INTRODUCTION

The electrocardiogram (ECG) is the most widely used non-invasive technique in heart disease diagnoses. It can be described as a record of the electrical phenomena originated from cardiac activity. Fig. 1 shows a schematic record of a normal heartbeat. The ECG is frequently used to detect cardiac rhythm abnormalities, otherwise known as, arrhythmias. Arrhythmias can be defined in two ways: as a unique irregular cardiac beat or as a set of irregular beats. Arrhythmias can be rare and harmless, but may also result in serious cardiac issues.

There are several methods proposed in the literature for the purpose of automatic arrhythmia classification in ECG signals and a complete system for such an aim can be divided into four subsequent categories (as shown in Fig. 2): preprocessing, segmentation, feature extraction, and classification.

The most widely used database for evaluation of the accuracy/sensitivity/specificity (from now on performance) of arrhythmia classification systems is the MIT-BIH Arrhythmia Database [1]. This database was the first available for such a purpose and it has gone through several improvements over the years to encompass the broadest possible range of waveforms [2]. The Association for the Advancement of Medical Instrumentation (AAMI) also recommends the use of the MIT-BIH Arrhythmia Database for performance eval-

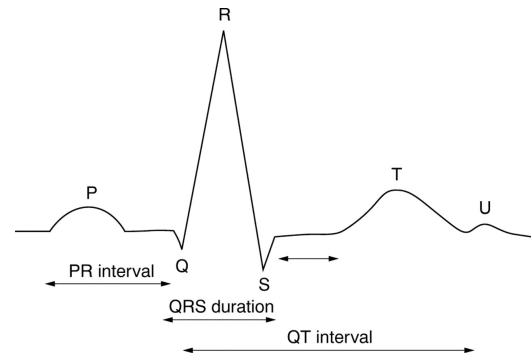


Fig. 1. A normal heartbeat ECG signal

uation of arrhythmia systems. The AAMI has developed a standard for testing and reporting performance results of algorithms aiming at arrhythmia classification (ANSI/AAMI EC57:1998/(R)2008). According to [3, 4] few researchers have used the AAMI recommendations and standards, leading to clinically unreliable results since several methods in the literature are favored by a biased dataset (*i.e.*, where heartbeats from the same patient are used for both training and testing the classifiers, which makes a fair comparison among methods difficult).

The aiming of this work is twofold. First to summarize recent techniques aiming at arrhythmia classification. And, second, analyzing the results obtained by different designs of automatic classification system using two ways for choosing samples for training/testing the performance of these systems - one following the AAMI recommendations and another one which disregard such recommendations.

The remainder of this work is organized as follows. In Section 2, we briefly describe each category of an arrhythmia classification system presenting the most relevant works, in our point of view, proposed so far. The methods used in our analysis are cited and grouped in Section 3. Finally, discussion of the results reported in those works and conclusions are pointed out in Section 4 and 5, respectively.

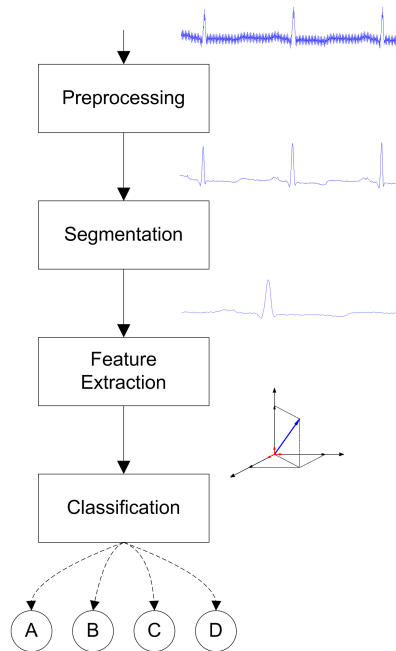


Fig. 2. A diagram of a classification system of arrhythmia

2. ARRHYTHMIA CLASSIFICATION SYSTEM

In this section, we present methods proposed for building a complete system for arrhythmia classification. The system can be divided into four subsequent categories, which starts with signal preprocessing, and then segmentation of each heartbeat and so before classification, the feature extraction step.

2.1. Preprocessing

The preprocessing consists mainly in detecting and attenuating frequencies of the ECG signal related to artifacts. Those artifacts can be from a biological source, like muscular activity, or can be originated from an external source, such as 50/60Hz from electrical network. It is also desired, in the preprocessing, to perform a signal normalization and complex QRS enhancement, in order to help the segmentation process.

Many methods have been proposed to reduce noise in the ECG signal. The most simple and fairly used is the implementation of digital filters [5]. Other architectures, like adaptive filters [6], have also been used to attenuate noise in ECG signal. Most sophisticated methods like adaptive filters based in neural network [7] have brought a significant improvement in noise attenuation process and then raised the effectiveness of segmentation and classification methods.

Statistics techniques, such as principal component analysis (PCA) [8] and independent component analysis (ICA) [9] are also powerful tools for noise attenuation in ECG signal, due to the fact that they allow one extraction of noises repre-

sented by frequencies very related or near to the ones of the ECG signal.

Nowadays, methods based in the wavelet transform are widely used. Due to a more accurate filtering process, they preserve the ECG signal, avoiding the loss of important physiological details [10, 11].

Other methods have also presented interesting results. In [12], a non-linear Bayesian filter is proposed to reduce the noise in ECG signal. In [13], a new algorithm based on extended Kalman filter structure incorporates the ECG dynamical model to attenuate noise and data compression of ECG signal. This approach has brought a significant improvement on noise suppression and overcome the most effective methods so far.

2.2. Segmentation

Regarding ECG signals analysis, segmentation consists in delimitating the part of the signal of more interest, the QRS complex, since it reflects the electrical activity of the heart (see Fig. 1). Once the segmentation of QRS complex is done one can obtain many physiological information, such as cardiac frequency, and so the techniques to extract features from the signal can be applied.

Several algorithms have been proposed in the literature for ECG beat segmentation. The problem faced by researches are many, since the ECG beat morphology can be vary for both inter- and intra-patient. A common approach for ECG signal segmentation, *i.e.*, the heartbeat detection, is based on digital filters for preprocessing, linear transformation for R peak enhancement, and adaptive thresholds for heartbeat recognition [14].

QRS detection methods have been proposed over three decades [14, 15, 16] and the evolution of those algorithms reflects the evolution of processing power of computers. Nowadays, more advanced methods are used and the most popular methods are based in neural network [17], genetic algorithm [15], wavelet transform [18], filter banks [19], and support vector machines [16].

2.3. Feature extraction

Feature extraction is the key point for the final classification performance. Features can be extracted directly from ECG wave morphology in time or frequency domain. More sophisticated methods have been used in order to find features less sensitive to noise, such as the autoregressive model coefficients, higher-order cumulant (higher order statistics) [20] and variations of wavelet transform. Researchers claim that wavelet transform is the most promising technique to extract features from the ECG signal [20, 21, 22]. However, in [23], the author argues that methods based on wavelet transform may have some limitations and its use should depend on the application.

Table 1. Mapping the MIT-BIH Arrhythmia types to the AAMI Classes

The AAMI heartbeat class	N	SVEB	VEB	F	Q
Description	Any heartbeat not in the S, V, F, or Q class	Supraventricular ectopic beat	Ventricular ectopic beat	Fusion beat	Unknown beat
	normal beat (N)	atrial premature beat (A)	premature ventricular contraction (V)	fusion of ventricular and normal beat (F)	paced beat (P)
	left bundle branch block beat (L)	aberrated atrial premature beat (a)	ventricular escape beat (E)		fusion of paced and normal beat (f)
MIT-BIH heartbeat types (code)	right bundle branch block beat (R)	nodal (junctional) premature beat (J)			unclassified beat (U)
	atrial escape beat (e)	supraventricular premature beat (S)			
	nodal (junctional) escape beat (j)				

The authors [24] claim that using techniques to reduce the dimension of feature space, such as PCA or linear discriminant analysis (LDA), can offer advantages such as reducing of time and amount of data required for training the classifier. According to them, the usage of techniques for reducing the feature space can worth the loss on accuracy. In [16], for a SVM classifier, the usage of LDA for reducing the feature dimension has shown greater accuracy than the usage of PCA. Moreover, those authors point out that the accuracy of the SVM classifier with reduced feature space using LDA is greater even than the accuracy with the original feature set.

2.4. Classifiers

In order to accurately detect cardiac frequency, it is necessary to consider sporadic arrhythmias occur. An accurate arrhythmia classification is also desirable to correctly diagnose cardiac issues and in some cases, the early detection can save lives. With that motivation in mind, researches keep the efforts to develop better and better methods.

Artificial neural networks (ANN) are widely used to arrhythmias classification in ECG signals [25], and the multi-layer perception (MLP), the most popular ANN, is often used for that purpose [20].

The conventional MLP has shown high accuracy in classification of arrhythmias. Nevertheless it suffers from slow local convergence, global minimum localization and random initial weights. These drawbacks could make it inappropriate to clinical usage [24]. To overcome this issues, hybrid systems, combining MLP with another ANN are normally indispensable [26]. In those kind of systems, the first level of networks are responsible to initially classify the heartbeats and also build models generating new feature inputs. The MLP completes the second task of multi-classification [27]. With

that approach, many weakness of MLP are surmounted.

In [27] and [20], a method based on higher order statistics to extract features, and a hybrid neuro-fuzzy method for classification [28], which uses type-2 fuzzy c-means algorithm to improve the accuracy of the neural network, have reached higher accuracies than conventional MLP methods.

SVM has also been widely used to classify arrhythmias. In [29], a comparison of different methods using SVM and ANN has shown that SVM methods should be choose when training time matters. Otherwise ANN methods have demonstrated better results. In [16], the authors have used linear discriminant analysis (LDA) in order to reduce the size of the feature space, and despite that fact a high accuracy has been shown.

In [30], it is proposed a method with fast learning rates and high accuracy (% 98.72), using morphology filtering, principal component analysis (PCA) and extreme learning machine (ELM). The algorithm is used to detect six types of arrhythmias and the results have shown that the method is faster than others like MLP and SVM.

3. METHODS

We chose eight methods to analyze their performances. Three of them, in our consideration, are state-of-the-art methods, since its authors have followed the AAMI recommendations [3, 4, 31]. In the remaining five methods, the authors did not follow the AAMI recommendations [32, 33, 25, 34, 35]. However they report performance in average near to 100% as shown in Table 2.

The MIT-BIH arrhythmia database contain 48 half-hour records, sampled at 360Hz, and eighteen types of heartbeats were classified and labeled. To comply with the AAMI recommendations, only 44 records of MIT-BIH arrhythmia

Table 2. Classification performance of methods using random selection of samples (heartbeats) - biased selection

Method	Accuracy	Sensitivities (%)														
		N	L	R	A	V	P	a	!	F	x	j	f	E	J	e
Ye <i>et al.</i> [32]	99.91	99.95	100	99.99	99.65	99.26	100	92.86	100	99.73	100	100	100	100	97.06	100
Yu & Chen [25]	99.65	99.97	99.33	99.54	99.76	99.04	100	-	-	-	-	-	-	-	-	-
Yu & Chou [33]	98.71	99.65	96.25	99.15	98.40	98.45	99.37	-	90.12	-	-	-	-	91.53	-	-
Korürek & Nizam [34]	-	95.49	-	97.56	86.78	93.33	-	-	-	74.51	-	-	84.06	-	-	-
Tsipouras <i>et al.</i> [35]	96.43	93.89	-	98.65	-	91.35	-	-	97.74	-	-	-	-	-	-	-

database should be used for evaluation of arrhythmia classification methods, excluding the 4 records that contain paced beats. The ANSI/AAMI EC57:1998/(R)2008 standards recommends to group those heartbeats into five classes: 1) normal beat; 2) ventricular ectopic beat (VEB); 3) supraventricular ectopic beat (SVEB); 4) fusion of a VEB and a normal beat; and 5) unknown beat type (see Table 1). Moreover, the AAMI standards also recommends to divide the recordings into two datasets, one for training and another for testing, such that heartbeats from one recording (patient) are not used simultaneously for both training and testing the classifier.

The methods which do not follow the AAMI standards for building the arrhythmia classifiers create randomly their datasets for training and testing, in such a manner that unavoidably heartbeats from one recording are present in both sets. This practice, *i.e.*, to put data from the same patient in both sets, should be avoided as already stated in [3].

There is also a lack of standard regarding classes of heartbeats to be analyzed. In some cases, the classifiers are design to classify a specific number of classes, *e.g.*, 2, 3, 10. In other cases, the authors present the performance of methods for non standard classes (*i.e.*, non beat annotation codes), such as Ventricular Flutter Wave (!) and Non-Conducted P-wave (x) [32, 33].

4. DISCUSSIONS

In order to analyze the classification performance, two measures are used, *i.e.* accuracy and sensitivity. Accuracy is defined as the ratio of total beats correctly classified and the number of total beats, *i.e.*,

$$Accuracy = \frac{\text{beats correctly classified}}{\text{number of total beats}}. \quad (1)$$

Sensitivity stands for the ratio of correctly classified beats of one class and the total beats classified as that class, including the miss classification beats, *i.e.*,

$$Sensitivity = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}. \quad (2)$$

Sensitivity¹ is the most important measure for our analysis, since the number of heartbeats for each class in the

¹together with specificity, which is not used in this study due to the lack of these data in the studied works.

Table 3. Classification performance of methods following the AAMI recommendations

Method	Accuracy	Sensitivities (%)				
		N	SVEB	VEB	F	Q
Chazal <i>et al.</i> [3]	85.9	86.86	75.93	77.73	89.43	0
Ince <i>et al.</i> [4]	93.6	97.04	62.11	88.39	61.36	0
Jiang & Kong [31]	94.5	98.73	50.58	86.61	35.78	0

Table 4. Classification performance of methods which do not follow the AAMI recommendations. The classes and method presented in Table 2 are grouped according to Table 1, to comply with the AAMI recommended classes

Method	Accuracy	Sensitivities (%)				
		N	SVEB	VEB	F	Q
Ye <i>et al.</i> [32]	99.91	99.96	98.48	99.83	99.21	99.96
Yu & Chen [25]	99.65	99.67	99.53	99.22	-	100
Yu & Chou [33]	98.71	99.81	98.50	97.74	-	100
Korürek & Nizam [34]	-	95.51	86.78	-	74.51	84.06
Tsipouras <i>et al.</i> [35]	96.43	93.90	-	91.35	-	-

MIT-BIH arrhythmia database is very imbalanced and a single class (*e.g.*, the normal beats) could represent most of the total accuracy, while the sensitivity and specificity directly depend on the number of samples for each class.

Comparing the results achieved by methods using the AAMI recommendations for designing the arrhythmia classification systems and the ones which do not follow them (Tables 3 and 4), respectively, we can observe a significant difference in terms of the sensitivities reported. This remark can be extended to the accuracy figures.

For both measures, the methods which do not follow the AAMI recommendations present higher values. It is noticeable that all methods analyzed in this work are consistent and use advanced techniques to solve the arrhythmia classification problem. Thus, we suggest that this significant difference in the performances are mostly related to datasets used for training and testing the classifiers. The use of a dataset for training a classifier and then testing it with samples (heartbeats) from the same patients helps the classifier to yield better classification results, since it is specialized in those data.

Besides the fact that heartbeats from same recording, used

both for training and testing, can favor the classifier, there is another practice that can lead to biased conclusions as well. Several methods do not use the complete data from the MIT-BIH arrhythmia database as done in [25] and [33], where only 23200 and 9800 heartbeats are used, respectively. In those approaches, the heartbeats were randomly chosen and the classifiers can be favored by eventually easily heartbeat patterns.

Moreover, according to [4], only a few of the methods presented in the literature have, in fact, used the AAMI standards. This statement suggest that the results of several methods in literature are unreliable and should not be taken into account clinically before a robust performance test can be performed.

5. CONCLUSIONS

In this work, we presented an X-ray, a generic view, on methods aiming at arrhythmia classification in ECG signals. Moreover, we showed that the challenges to properly classify arrhythmias in ECG signal are many.

Researchers have been working on improvements, and many of them have shown remarkable results. Nonetheless, few authors have considered the impact on the performance of the classifiers caused by the way the samples (heartbeats) were selected for building the dataset used for training and testing the classifiers. This work have cited methods that may use heartbeats from same patients for training and testing a classifier which could favor their results in terms of performance. However, those reported performances are not realistic, since those methods will classify “never seen” heartbeats (e.g., a new patient), and in these situations, the performance obtained by the method can be quite small.

Thus, the choice of unbiased dataset, such as recommended by the AAMI, should be used for arrhythmia classification methods in order to obtain more reliable results. Having this fact in mind, several methods in the literature can be re-run using unbiased datasets. These results should be used for report new prediction values for these methods, establishing a new state-of-the-art method in terms of performance.

6. REFERENCES

- [1] Massachusetts Institute of Technology, “MIT-BIH ECG database,” available at <http://ecg.mit.edu/>.
- [2] G. B. Moody and R. G. Mark, “The impact of the MIT-BIH arrhythmia database,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.
- [3] P. Chazal, M. O’Dwyer, and R. B. Reilly, “Automatic classification of heartbeats using ECG morphology and heartbeat interval features,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196–1206, 2004.
- [4] T. Ince, S. Kiranyaz, and M. Gabbouj, “A generic and robust system for automated patient-specific classification of ECG signals,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 5, pp. 1415–1427, 2009.
- [5] P. Lynn, “Recursive digital filters for biological signals,” *Medical and Biological Engineering and Computing*, vol. 9, no. 1, pp. 37–43, 1979.
- [6] M. Yelderian, B. Widrow, J. M. Cioffi, E. Hesler, and J. A. Leddy, “ECG enhancement by adaptive cancellation of electrosurgical interference,” *IEEE Transactions on Biomedical Engineering*, vol. 30, no. 7, pp. 392–398, 1983.
- [7] Q. Xue, Y. H. Hu, and W. J. Tompkins, “Neural-network-based adaptive matched filtering for QRS detection,” *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 4, pp. 317–329, 1992.
- [8] F. Castells, P. Laguna, L. Sörnmo, A. Bollmann, and J. M. Roig, “Principal component analysis in ECG signal processing,” *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 98–98, 2007.
- [9] T. He, G. Clifford, and L. Tarassenko, “Application of independent component analysis in removing artefacts from the electrocardiogram,” *Neural Computing and Applications*, vol. 15, no. 2, pp. 105–116, 2006.
- [10] B. N. Singh and A. K. Tiwari, “Optimal selection of wavelet basis function applied to ECG signal denoising,” *Digital Signal Processing*, vol. 16, no. 3, pp. 275–287, 2006.
- [11] O. Sayadi and M. B. Shamsollahi, “Multiadaptive bionic wavelet transform: Application to ECG denoising and baseline wandering reduction,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 14, pp. 1–11, 2007.
- [12] R. Sameni, M. B. Shamsollahi, C. Jutten, and G. D. Clifford, “A nonlinear Bayesian filtering framework for ECG denoising,” *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 12, pp. 2172–2185, 2007.
- [13] O. Sayadi and M. B. Shamsollahi, “ECG denoising and compression using a modified extended kalman filter structure,” *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 9, pp. 2240–2248, 2008.
- [14] J. Pan and W. J. Tompkins, “A real-time QRS detection algorithm,” *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, 1985.
- [15] R. Poli, S. Cagnoni, and G. Valli, “Genetic design of optimum linear and nonlinear QRS detectors,” *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 11, pp. 1137–1141, 1995.

- [16] M. H. Song, J. Lee, S. P. Cho, K. J. Lee, and S. K. Yoo, "Support vector machine based arrhythmia classification using reduced features," *International Journal of Control, Automation, and Systems*, vol. 3, no. 4, pp. 509–654, 2005.
- [17] I. Güler and E. D. Übeyli, "ECG beat classifier designed by combined neural network model," *Pattern Recognition*, vol. 38, no. 2, pp. 199–208, 2005.
- [18] Y. Jung and W. J. Tompkins, "Detecting and classifying life-threatening ECG ventricular arrhythmias using wavelet decomposition," in *IEEE International Conference on Engineering in Medicine and Biology Society*, 2003, vol. 3, pp. 2390–2393.
- [19] V. X. Afonso, W. J. Tompkins, T. Q. Nguyen, and S. Luo, "ECG beat detection using filter banks," *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 2, pp. 192–202, 1999.
- [20] E. Mehmet, "ECG beat classification using neuro-fuzzy network," *Pattern Recognition Letters*, vol. 25, no. 15, pp. 1715–1722, 2004.
- [21] I. Güler and E. D. Übeyli, "ECG beat classifier designed by combined neural network model," *Pattern Recognition*, vol. 38, no. 2, pp. 199–208, 2005.
- [22] C. Lin, Y. Du, and T. Chen, "Adaptive wavelet network for multiple cardiac arrhythmias recognition," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2601–2611, 2008.
- [23] Y. Özbay, "A new approach to detection of ECG arrhythmias: Complex discrete wavelet transform based complex valued artificial neural network," *Journal of Medical Systems*, vol. 33, no. 6, pp. 435–445, 2009.
- [24] R. Ceylan and Y. Özbay, "Comparison of FCM, PCA and WT techniques for classification ECG arrhythmias using artificial neural network," *Expert Systems with Applications*, vol. 33, no. 2, pp. 286–295, 2007.
- [25] S. Yu and Y. Chen, "Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1142–1150, 2007.
- [26] M. H. Fredric and H. Soowhan, "Classification of cardiac arrhythmias using fuzzy ARTMAP," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 4, pp. 425–429, 2002.
- [27] S. Osowski and T. H. Linh, "ECG beat recognition using fuzzy hybrid neural network," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1265–1271, 2001.
- [28] R. Ceylan, Y. Özbay, and B. Karlik, "A novel approach for classification of ECG arrhythmias: Type-2 fuzzy clustering neural network," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6721–6726, 2009.
- [29] M. Moavenian and H. Khorrami, "A qualitative comparison of artificial neural networks and support vector machines in ECG arrhythmias classification," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3088–3093, 2010.
- [30] J. Kim, H. S. Shin, J. Shin, and M. Lee, "Robust algorithm for arrhythmia classification in ECG using extreme learning machine," *BioMedical Engineering On-Line*, vol. 8, no. 1, pp. 1–12, 2009.
- [31] W. Jiang and G. S. Kong, "Block-based neural networks for personalized ECG signal classification," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 750–761, 2007.
- [32] C. Ye, M. T. Coimbra, and B. V. K. V. Kumar, "Arrhythmia detection and classification using morphological and dynamic features of ECG signals," in *IEEE International Conference on Engineering in Medicine and Biology Society (EMBC)*, 2010, pp. 1918–1921.
- [33] S. Yu and K. Chou, "Integration of independent component analysis and neural networks for ECG beat classification," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2841–2846, 2008.
- [34] M. Korürek and A. Nizam, "A new arrhythmia clustering technique based on ant colony optimization," *Journal of Biomedical Informatics*, vol. 41, pp. 874–881, 2008.
- [35] M. G. Tsipouras, C. Voglis, and D. I. Fotiadis, "A framework for fuzzy expert system creation-application to cardiovascular diseases," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 11, pp. 2089–2105, 2007.

The Application of the Genetic Algorithm Tool Box at Target Enhancement Processing in Breast Cancer Microwave Imaging

Zhongling Han¹, Zhifu Tao¹, Meng Yao^{1*}, Yizhou Yao²,
Blair Fleet³, Erik D. Goodman³, and John R. Deller⁴

¹ Institute of information science and technology East China Normal University, Shanghai, China

² Weiyu high school, Shanghai, China

³ BEACON Michigan State University, East Lansing, MI

⁴ ECE Michigan State University, East Lansing, MI

*Corresponding Author, e-mail: myao@ee.ecnu.edu.cn

Abstract - Genetic algorithms, as kind of fast, simply and highly fault-tolerant algorithms for near-field microwave detection of breast cancer; are useful processing method. For the image has not been dealt with high demands, which can quickly detect the interested area, and by bypassing the non-line Inversion of the problem brought about difficulties. The article mechanism to the imaging reflecting from analyzing a microwave starts off; analyses principle and process applying the Genetic algorithm detecting targets.

Keywords: Microwave reflection; genetic algorithm; Image inversion

1 Introduction

Microwave imaging sounding is an emerging breast cancer detection research method. This method might present a safe, convenient and cost-effective supplement to traditional BC imaging diagnoses methods. Different tissues can be imaged, which is based on a very high dielectric contrast between normal tissue and malignant tissue. It is known that the dielectric properties of tissues with high (tumor) and low (normal tissue) water content are significantly different. Reflection and refraction will occur and consequently the electromagnetic wave propagation path will be change, when the electromagnetic wave encounters all sorts of medium layer in process of propagation, such as skin, fat, breast lobules and tissue etc. Based on this property, we can detect and locate the breast tumor by detecting the microwaves reflect signal energy in different dielectric contrast mediums interface. The reflect signal energy also can convert into the transmitting waves propagation distance. In this paper, we start from analyzing the mechanism of microwave reflection imaging, and then using the genetic algorithm [1, 2] to optimize the inversion image. It has been found that this method can increase the detectable rate of breast cancer.

2 Breast Cancer Detection Microwave Imaging

Assumed microwave reflectivity function of the cross section of the imaged object is $f(x, y)$. The coordinate is fixed on the imaged object, and the non-directional transceiver antenna is located in the point (x_0, y_0) . The antenna can only be rattling around the imaged object in a circle of radius R , whose moving can change the azimuth angle ϕ between the microwave beam and the imaged object. The reflect signal wave $p(\rho, \phi)$ received by the transceiver antenna represents line integral of $f(r, \theta)$ along a concentric circle arc s which the center is (x_0, y_0) (Shown in Figure 1).

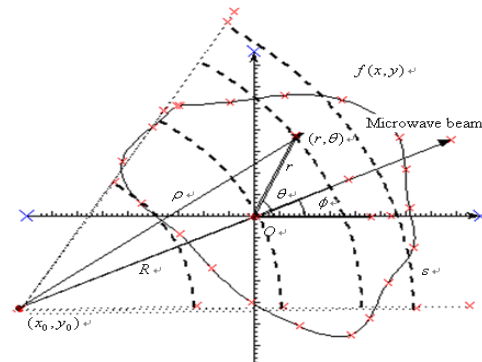


Figure 1. It gives the geometric scheme of microwave imaging. The microwave reflectivity function is $f(x, y)$. The transceiver is located in point (x_0, y_0) , and it can only be rattling around the imaged object in a circle of radius R . the tumor is located in point (r, θ) . The angle between the microwave beam and the imaged object is ϕ .

The projection is expressed as follow.

$$p(\rho, \phi) = \int_s f(x, y) ds = \int_0^{2\pi} \int_0^{\infty} f(r, \theta) \delta(\sqrt{r^2 + R^2 - 2rR \cos \alpha} - \rho) r dr d\theta \quad (1)$$

$$\sqrt{r^2 + R^2 - 2rR \cos \alpha} = \sqrt{(x - x_0)^2 + (y - y_0)^2},$$

$$\alpha = \pi - (\theta - \phi) \quad (2)$$

Where $\delta(d-\rho)$ is δ -function. We can get the equation (3) by the Fourier transform of projection $S(\omega, \phi)$ and paraxial approximation.

$$S(\omega, \phi) = e^{-j\omega R} [F(u, v) - j\omega F^{(1)}(u, v) - \omega^2 F^{(2)}(u, v) + \dots] \quad (3)$$

Where, $S(\omega, \phi) = \int_{-\infty}^{+\infty} p(\rho, \phi) e^{-j\omega\rho} d\rho$

$$F(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) e^{-j(ux+vy)} dx dy$$

$$F^{(1)}(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G(x, y) f(x, y) e^{-j(ux+vy)} dx dy$$

$$F^{(2)}(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G^2(x, y) f(x, y) e^{-j(ux+vy)} dx dy$$

...

$$F^{(n)}(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G^n(x, y) f(x, y) e^{-j(ux+vy)} dx dy$$

$$G(x, y) = \frac{(x \sin \phi - y \cos \phi)^2}{2R}$$

$$u = \omega \cos \phi, \quad v = \omega \sin \phi$$

We can use Fourier transform to rebuild the profile distribution $f(x, y)$ of the target by detecting the backscattering field (projection) $p(\rho, \phi)$. It can be ignored such as $F(1)(u, v)$ and $F(2)(u, v)$, when R is very big, and $f(x, y)$ is limited band function. Then equation (1) can be seen as Fourier slice theory of CT imaging. The theory shows that the value can be given by the one-dimensional Fourier transform of the signal $p(\rho, \phi)$, which is the two-dimensional Fourier transform of the function $f(x, y)$ in spatial frequency. u equals to $\omega \cos \phi$, v equals to $\omega \sin \phi$. We could get the function $f(x, y)$ of detection space after inverting above process.

3 principle of breast cancer detection

The function $f(x, y)$, which is the distribution function of reflection coefficient for each point in the detection space can be obtained from inversion. According to microwave propagation principle, $f(x, y)$ can be calculated by equation (4).

$$f(x, y) = \frac{\sqrt{\varepsilon_1(x, y)} - \sqrt{\varepsilon_2(x, y)}}{\sqrt{\varepsilon_1(x, y)} + \sqrt{\varepsilon_2(x, y)}} \quad (4)$$

Where $\varepsilon_1(x, y)$ and $\varepsilon_2(x, y)$ are the dielectric constant of two sides of the interface, respectively. Through calculation, the echo coefficient of cancerous tissue is greater than 0.49, the echo coefficient of lobules and gland tissue is between 0.2018 and 0.49 and the echo coefficient of blood vessels tissue is between 0.15 and 0.2018. Detecting breast cancer can be transformed into detecting some particular value distribution of $f(x, y)$ in reflection space. So, cancerous tissue target is located in those points which echo coefficient is greater than 0.49.

4 Genetic algorithm process and results

The goal of the microwave breast tumor detection is to separate cancerous target from inversion image $f(x, y)$. So the function $f(x, y)$ can be divided into two categories: one belongs to the cancerous target distribution C_1 which the image intensity is greater than M ; Another kind does not belong to the cancerous target distribution C_2 which the image intensity is less than or equal to M . Considering actual inversion image grey levels of 256, the threshold value of the gray-level image coding for an 8 bits binary code string. Fitness function is shown in equation (5).

$$f(M) = w_1(M) * w_2(M) * [u_1(M) - u_2(M)]^2 \quad (5)$$

Where $w_1(M)$ is pixel number of C_1 , $w_2(M)$ is pixel number of C_2 , $u_1(M)$ is the average gray value in C_1 , $u_2(M)$ is the average gray value in C_2 .

Following is the genetic algorithm processing function [3,4] of MATLAB code, which used genetic algorithm tool box developed by the University of Sheffield.

```
function C= Segmentation_GA (X, downth, upth);
% input: X is waiting process image matrix,
% downth and upth is the threshold value of detection
target
% output: C is the marked image
NIND=40; % Individual number
NVAR=1; % variable number
MAXGEN=50; % Maximum number of generations
PRECI=8; % bit number of variable
GGAP=0.9; % generation gap
FieldID=[8*NVAR;downth;upth;1;0;1;1]; % Establishment of
regional described device
Chrom=crtbp(NIND,PRECI*NVAR); % Establish initial
population
gen=0;
phen=bs2rv(Chrom,FieldID); % The initial population decimal
conversion
ObjV=target(X,phen); % Calculate population fitting
function
while gen<MAXGEN
FitnV=ranking(ObjV); % Distribution fitness value
SelCh=select('sus',Chrom,FitnV,GGAP); % Select
SelCh=recombin('xovsp',SelCh,0.7); % Recombine
SelCh=mut(SelCh); % Mutation
phenSel=bs2rv(SelCh,FieldID); % Offspring decimal
conversion
ObjVSel=target(X,phenSel);
[Chrom,ObjV]=reins(Chrom,SelCh,1,1,ObjV,ObjVSel); %
Reinsert
gen=gen+1;
end
[Y,I]=max(ObjV);
M=bs2rv(Chrom(I,:),FieldID); % Estimate threshold
[m,n]=size(X);
for i=1:m
for j=1:n
if X(i,j)>M
```

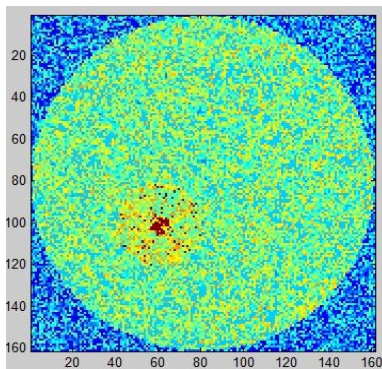
```

    X(i,j)=upth; % Using maximum mark
end
end
end
C=X; % Output marked image

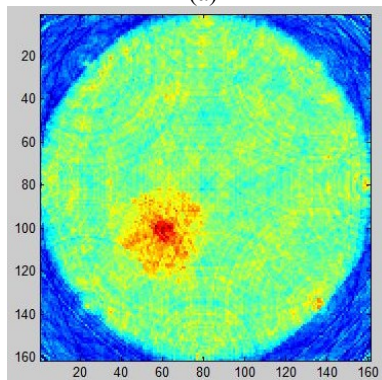
```

5 Results

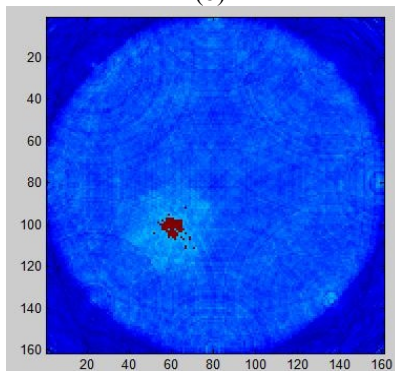
With the genetic algorithm method, we have succeeded in target enhancement. Figure 2 is simulation results of the data inversion when the microwave transceiver is located in circumference sampling uniformity of eight sampling test points. Form figure 2, we can see that the breast cancer is obviously enhanced after fifty iterations, and the background is smoother than ahead two imaging.



(a)



(b)



(c)

Figure2. It is inversion images of eight sampling test points. The distribution of the original detection space, the inversion detection space and the detection space processed after fifty

iterations by the genetic algorithm is shown in (a), (b) and (c), respectively.

6 Conclusion

In this study, we analyzed the microwave imaging theory and the microwave breast cancer detection theory. Then we use the simple genetic algorithm into microwave breast cancer detection to enhance tumor emerging area. As is a rapid, easy and fault-tolerance strong robust algorithm, the simple genetic algorithm can very good solve the nonlinear inversion problem in detection. A great deal of further research effort is needed to elucidate certain aspects of the genetic algorithm application to stability and target detection error.

7 Acknowledgments

This work has been performed while Prof. M. Yao was a visiting scholar in Michigan State University, thanks to a visiting research program from Prof. Erik D. Goodman. M. Yao would also like to acknowledge the support of Shanghai Science and Technology Development Foundation under the project grant numbers 03JC14026 and 08JC1409200, as well as the support of TI Co. Ltd through TI (China) Innovation Foundation.

8 Reference

- [1] Ming Zhou. Principle and application of the genetic algorithm Defense industry press 1999.6, pp. 36–37
- [2] Rajarshi Das, Melanie Mitchell, and James P. Crutch_eld, *Parallel Problem Solving from Nature---PPSN III*. Berlin: Springer-Verlag. 1994
- [3] Das, R. and Whitley, L.D. The Only Challenging Problems Are Deceptive: Global Search by Solving Order-1 Hyperplanes. Proceedings of ICGA 1991, pp.166-173.
- [4] Christopher R. Houck, Jeffery A.Joines, and Michael G. Kay A Genetic Algorithm for Function Optimization- A Matlab Implementation Technical Report NCSU-IE-TR-95-09, North Carolina State University, Raleigh, NC (1995)

Vitreous imaging system

New method for medical diagnosis

Dr. Boucherit Taieb,
Privet Laboratory, 12 impasse de Venise, Oran, Algeria
Privet Laboratory, Dr. Boucherit Taieb, Oran, Algeria

Abstract - The Importance of “vitreous imagery system” it is a discovery in the field of the medical imagery and of the diagnosis, the “vitreous imagery system” gives a completely remarkable new approach for the medical diagnosis it is precise, without passing by the traditional way which is, long, tiring method and sometimes dangerous for the patient, “the vitreous imaging system” puts all the capacities of the computer at the service of the patient.

- Environment slightly enlightened without important source of light
- The “*vitreous imagery system*” makes it possible to visualize the images of the patient’s organs in the vitreous humor , these images are laid out in bulk, with sometimes the repetition more than one organ,same organ with different view.
- We resize each image of organs obtained in the humor vitreous to isolate each image.

1 Introduction

I always thought that the process to make a traditional medical diagnosis is a very long and complicated process, Hard & tiring for the patient, but especially unreliable.

Taking this into consideration, part of my research work is concerned with trying to find a means of making the diagnosis quickly and accurate

With this new method “vitreous imaging system”. I discovered that we can make a quickly diagnosis without going through the classical process, the “vitreous imagery system” visualizes in images the pathological organs, these images contain an infinite data thus enabling us to have anatomical, histological, anatomopathologic, microscopic and ultra microscopic information.

I show you these images obtained by "the vitreous imagery system", you can see by yourself the quality and the precision of these images

2 Material & Methods

2.1 Materials

The material is very simple; it consists of a camera & computer,

2.2 Methods

- Photo of the eye.
- Front view photo of the eye
- Camera without flash

2.3 Theory & explanation

The images is formed on the retina, which converted it into nerve impulse and transmits it to the brain and since each eye receives an image a little different from the observed object, the brain compares the information coming from each eye and reconstitutes the image in three dimensions.

The human eye is a window open on the outside world, it receives the images from the outside environment and transmits to the brain to be analyzed and treated according to the corresponding answers. The image is formed on the retina, which converted it into nerve impulse and transmits it to the brain and since each eye receives an image a little different from the observed object, the brain compares the information coming from each eye and reconstitutes the image in three dimensions.

2.4 Anatomical composition of the eye

The eyeball of a grow- up measures 2.5 cm this little volume regroups nervous cells, muscles and transparent surroundings.

Muscles: these are the ciliary bodies, which modify the curvature of crystalline lens during accommodation

Cornea: is a transparent membrane made up of several layers which are directly in contact with the ambient air.

Aqueous humor: is a transparent watery fluid that is permanently, filtered and renewed in order to keep the eyeball in proper and good condition.

Iris: is a diaphragm that regulates the amount of light that the enters through the pupil.

Crystalline lens: is a simple convergent lens, that is held by ligaments which are tied to muscles (ciliary bodies) they modify in this way the curvature of the crystalline lens and make possible focusing.

Vitreous humor: is a transparent gelatinous and translucent substance whose function is to keep the retina against the inner lining of the eye it defines the form the eye and represents 90% of its volume.

Retina: is a nervous membrane forming the inner lining of the posterior wall of the eye, it is a few tens millimeters thick with a global surface of 2.5cm x 2.5cm, it consists of 130 million nerve cells (125 million retinal rods and 5 million retinal cones). It transforms light into electric signals which are conveyed to the brain.

Sclerotic: is the firm that forms the outer covering of the eyeball, its anterior covering is the cornea; the sclera is perceptible from the outside and constitutes the whites of the eyes.

This is the classical theory of today.

this theory is inadequate for it cannot explain the complexity of the eye : you notice that the major part of the eye (90%) is the vitreous humor whose only function is to maintain the shape of the eye , while the retina, a membrane of 2.5 cm² and few ten millimeters thick consists of 130 million nerve cells, each one , has a very precise function, a plant that is so complex and fitted with such technology that only 10% of its volume works, while 90% are for aesthetical reasons.

I looked into the problem and realized that the function of the vitreous humor is actually much more important than it seems. Chemical composition of the vitreous humor: 99,6% of water, vitamin C, glucose, lactic acid, NA ,CL, hyluronic acid, complete absence of vascularization.

My research enables me to prove that the images are materialized in the vitreous humor; it is its chief function. The functions of maintenance, the nutritious function are of a minor importance.

2.5 The eye has two functions

- **An open window on the outside world**
- **An open window on the inside body**

it is a movie, camera which is recording in both way, the image is recorded and formed in “energetic image” in the vitreous humor, the retina that consists of cone cells and rod cells digitizes the image and transmits it to the brain through the optic nerve, the digitization is carried out in “energetic language”

The numeric language uses a mathematical algorithm whose basis is 0 and1.

Energetic language uses energetic algorithm whose source are colors and shapes.

Numeric image is the resultant between the observed image and the approximate image in the data bank of the computer; actually, numeric image is not a real one.

Energetic image is real, it is itself image data bank, that is to say it contains endless images, that is why the image is multidimensional and not three-dimensional wich is transmitted to the brain.

I am doing my research on this particular field and I noticed that when an organ or several ones are affected, the image of that organ appears in the vitreous humor these enlarged and processed images of organs offer anatomical, histological images with an accuracy defying any radiological, scanographical,or microscopic equipment.

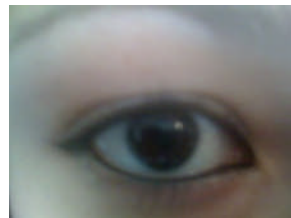
2.6 Example patient

- **Female patient 20 years age:**

Painful joints : wrist, elbow, pelvis , leg patient hospitalized in different hospitals of France and Belgium .

Diagnosis changing according to the hospital : Still disease, lupus...

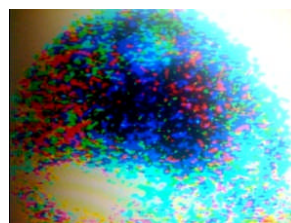
Vitreous imaging system diagnosis:



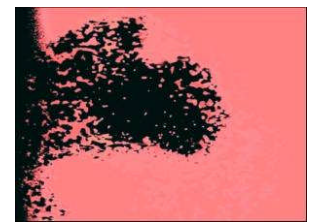
img 1



img 2



img 3



img 4

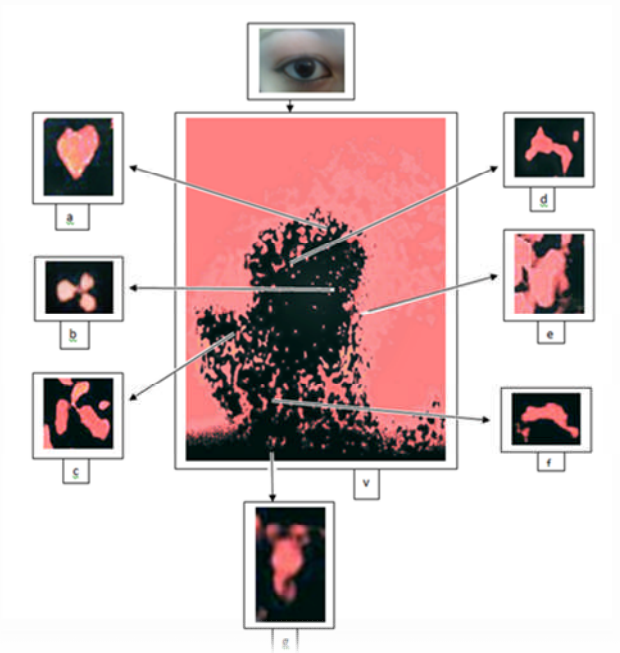


img 5

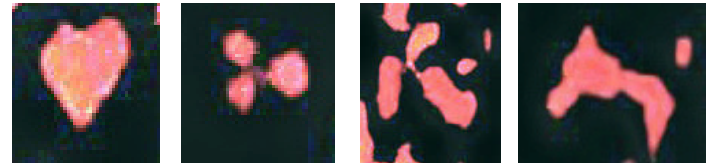


Vitreous humor

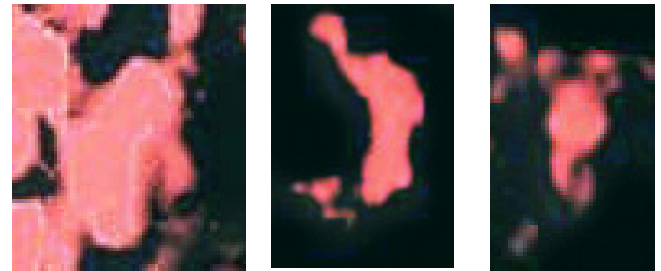
- Img 1** : right eye photo of the patient
- Img 2** : resize up image eye of the patient
- Img 3, Img 4, Img 5** , : images organs in the vitreous humor



2.6.1 The affected organs appears in the vitreous humor

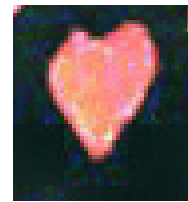


img(a) img(b) img(c) img(d)

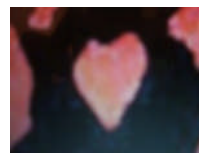


img(e) img(f) img(g)

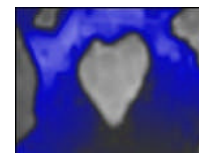
2.6.2 Images processing by computer:



img(a)



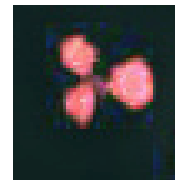
img(a)



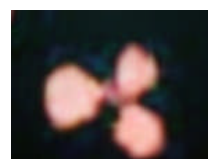
img(a1)



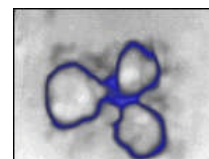
img(a2)



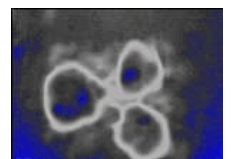
img(b)



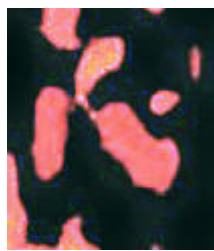
img(b)



img(b1)



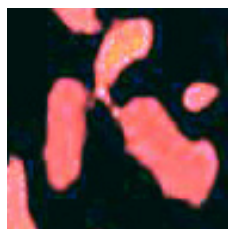
img(b2)



img(c)



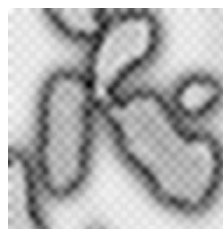
img(f)



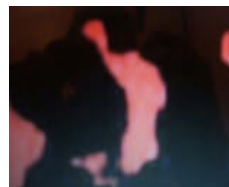
img(c)



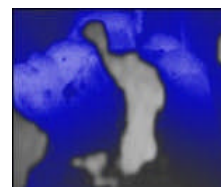
img(c1)



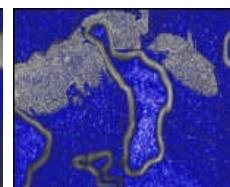
img(c2)



img(f)



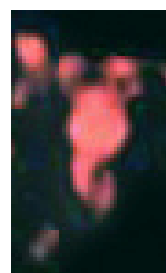
img(f1)



img(f2)



img(d)



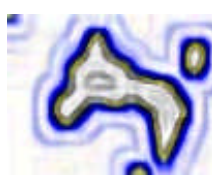
img(g)



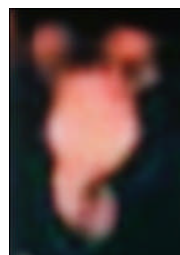
img(d)



img(d1)



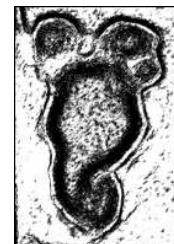
img(d2)



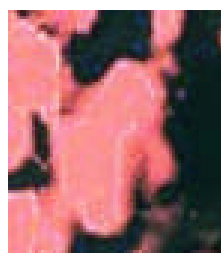
img(g)



img(g1)



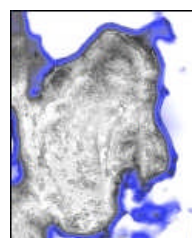
img(g2)



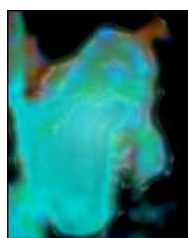
img(e)



img(e)



img(e1)

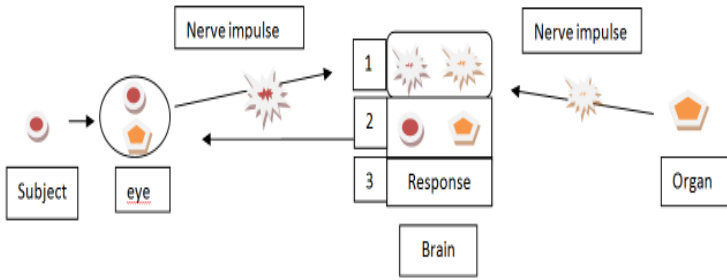


img(e2)

After processing images by computer we result:

- a. heart image (a , a1, a2)
- b. cervical vertebra image view face (b, b1, b2,)
- c. lung image (c, c1, c3, c4)
- d. cervical vertebra image view profile (d, d1, d2)
- e. kidney image (e, e1, e2)
- f. iliac bones image (f, f1, f2, f3, f4,f5, f6))
- g. apparatus genital (uterus, ovary and vagina) (g, g1, g2,g3, g4)

2.6.3 Diagram



EXPLANATION OF the MECHANISM OF RECEPTION IMAGES BY the BRAIN AND INTERCONNECTION WITH the EYE

2.6.4 The mechanism explanation

- Reception zone of encrypted image
- zone of decoding
- zone of Response

– **Outside the eye towards the brain**

The image is received by the eye is digitized and transmitted in nerve impulse by the optical nerve to the brain (occipital zone or surface number 8 in charge of the vision). The visual surface is dividing in three zones, the first zone is the zone of the reception of the nerve impulse, the second zone is responsible for the analysis of received information and the third is responsible for the response.

The image on the eye is digitized by the cells in cones and sticks in the form of electrical signal and is transmitted to the brain, in the zone number 1 receives coding information which is transmitted to the zone number 2 that converts this signal in real image.

The language of the brain is an energy language the brain receives a signal in the form of colors and shapes, which are decoded into real and comprehensible image.

– **Interior of the body towards the brain**

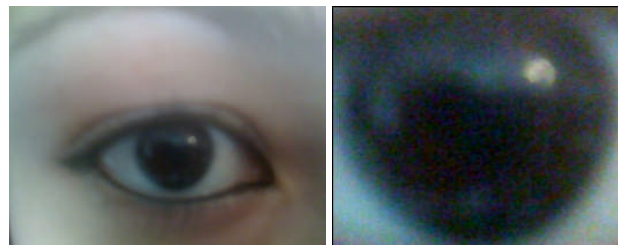
The affected organ sends its image by nerve impulse to the brain on the zone of reception of information, in codified “colors and shapes”. This information goes into zone number two in order to be converted into real images

But as the image is in contact with the equipment which transmits the image of opposite towards the interior, the image will be returned in the other direction, go towards and it is visible in the humor vitreous thanks to the effect of opposite mirrors.

2.7 The germ identification

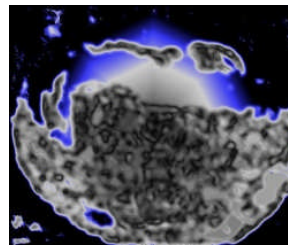
The traditional identification is made by: “direct examination” by “culture”, “search for antigen”... but, the identification by “vitreous imaging system” is done by imaging, we directly have the images of the germ. (h; photo of the eye, h1:h2, h3, h4, h5, h6)

not number any pages in your paper and do not reference page numbers in the text.

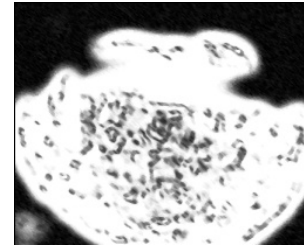


h1

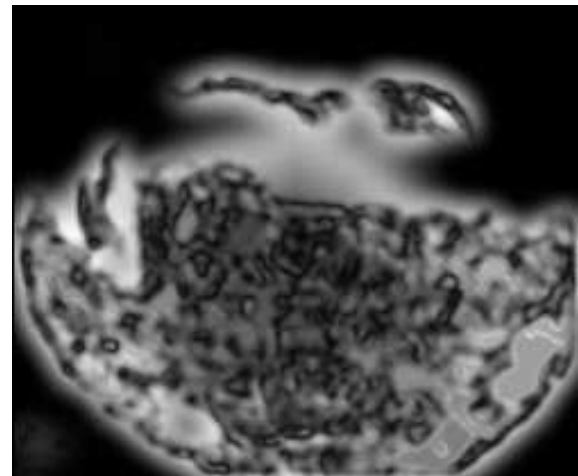
h2



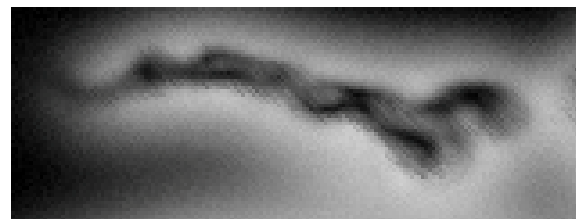
h3



h4



h5

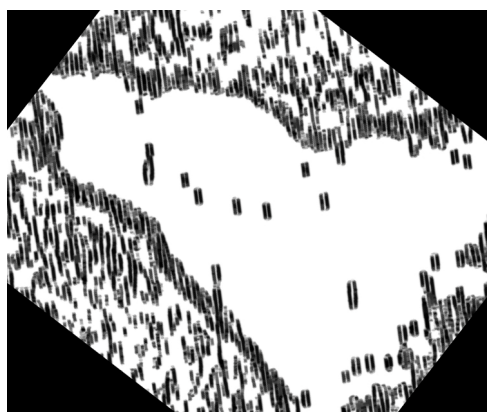


h6: Borrelia-germ

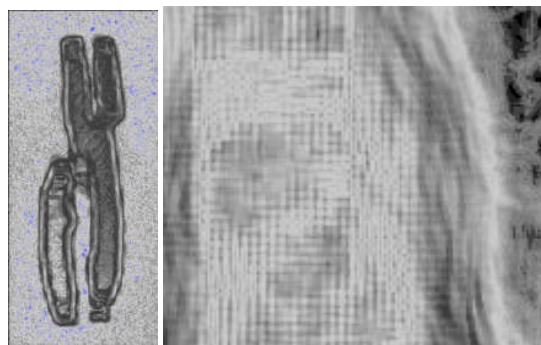
The identification of this germ is **Borrelia** germ, so the disease is **Lyme disease**.

2.8 The Chromosome & DNA identification

By the “*vitreous imaging system*” we can obtain the images of chromosom directly.



img (m)



img(p)

img(p1)

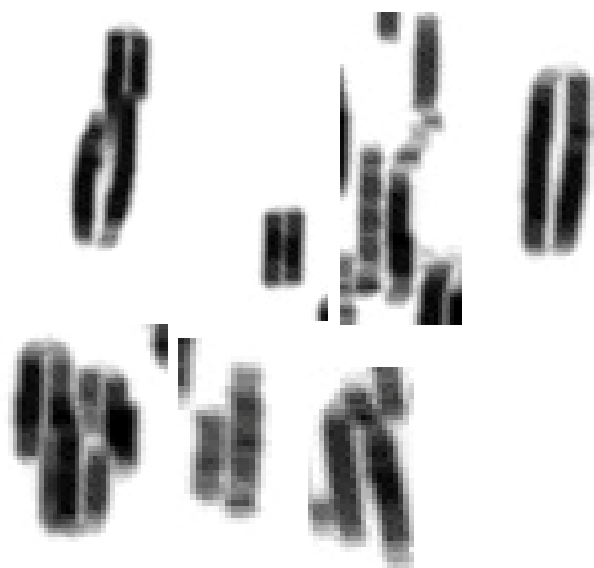


img(q)

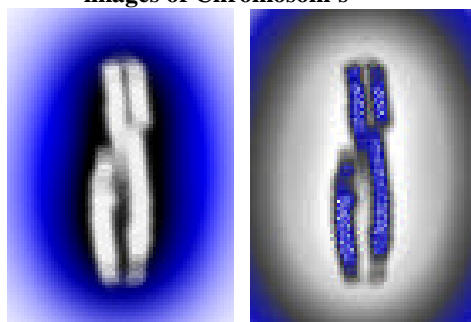
img(q1)

img(l)

- img(m) : visualization of chromosomal
- img(n) : image of the chromosomal X
- img(p) : chromosomal X
- img(p1), img(q); img(q1): part of chromosomal.
- img(l) : DNA



images of Chromosom's



imag(n) Chromosom X

3 Conclusions

The “*Vitreous imaging system*” is the most **precise** technique and most complete to take a diagnosis without mistake, with a maximum of safety for the patient.

The “*Vitreous imaging system*” will bring a giant jump in the field of the medical imaging and diagnosis namely:

- The great effectiveness for diagnosis
- Duration of examination (few minutes)
- Diagnosis in imaging in a few minutes
- Great safety for the patient and doctors
- No product of contrast or radiation in order to obtain the images.
- Exact localization of the pathological lesion, the images of diseased organs appear automatically in the «vitreous humor”.

The “*Vitreous imaging system*” makes it possible to take a diagnosis, to follow the evolution of the disease, the evolution of the treatment.

We explained that the images of the “vitreous imaging system” give an infinity of information's , the spectrum investigation is very large , and offers to us information's in others fields namely : genetics , cartography , and others disciplines .

SESSION

SOFTWARE PACKAGES AND OTHER COMPUTATIONAL TOPICS IN BIOINFORMATICS + RELATED DISCUSSIONS

Chair(s)

TBA

A Network of Hidden Markov Models and Its Analysis

Liqing Zhang¹, Layne T. Watson², and Lenwood S. Heath¹

¹Departments of Computer Science, Virginia Tech, Blacksburg, VA, USA

²Departments of Computer Science and Mathematics, Virginia Tech, Blacksburg, VA, USA

Abstract—*The Structural Classification of Proteins (SCOP) database uses a large number of hidden Markov models (HMMs) to represent families and superfamilies composed of proteins that presumably share the same evolutionary origin. However, how the HMMs are related to one another has not been examined before. In this work, taking into account the processes used to build the HMMs, we propose a working hypothesis to examine the relationships between HMMs and the families and superfamilies that they represent. Specifically, we perform an all-against-all HMM comparison using the HHsearch program and construct a network where the nodes are HMMs and the edges connect similar HMMs. We hypothesize that the HMMs in a connected component belong to the same family or superfamily more often than expected under a random network connection model. Results show a pattern consistent with this working hypothesis. Moreover, the HMM network possesses features distinctly different from previously documented biological networks, exemplified by the exceptionally high clustering coefficient and the large number of connected components. The current finding may provide guidance in devising computational methods to reduce the degree of overlaps between the HMMs representing the same superfamilies, which may in turn enable more efficient large-scale sequence searches against the database of HMMs.*

Keywords: hidden Markov models, network, centrality, clustering coefficient, tree

1. Introduction

The Structural Classification of Proteins (SCOP) database is a comprehensive protein database that organizes and classifies proteins based on their evolutionary and structural relationships [1], [7], [8]. It is organized into four hierarchical levels: family, superfamily, fold, and classes. At the lowest level (family), individual proteins are clustered into families based on some criteria that may indicate their common evolutionary origin, such as having a pairwise sequence similarity of more than 30% or lower sequence similarity but similar functions and structures. A good example of the latter is seen in globin proteins whose pairwise sequence similarities are much lower than 30% but which have similar protein functions. Next, families are grouped into superfamilies if their structures and/or function features indicate a possible common evolutionary origin. Then superfamilies are clustered into folds if superfamilies share

major secondary structures with the same topological arrangements. Finally, different folds are grouped into classes based on their secondary structural compositions. Unlike the other levels, a class might not necessarily imply common evolutionary origins and exists more for convenience than for actual biological implications.

Apart from the hierarchical classification and organization of proteins, the SCOP database employs hidden Markov models (HMMs) to represent superfamilies [4], [5]. The basic procedure of building an HMM for a particular superfamily starts with a seed protein and performs sequence search in a database to obtain other proteins that have sequence similarities above a set threshold. The newly obtained sequences are used to iterate the search for some number of times to obtain additional proteins. Finally, all sequences are aligned and an HMM is constructed for the multiple sequence alignment [4], [5]. It has been shown that different seed proteins might produce HMMs that cover different members of the superfamily [4], [5]. Thus, in order to represent the full set of proteins in a superfamily, multiple HMMs are built for the superfamily using multiple seed proteins. For example, the beta-beta-alpha zinc fingers superfamily has altogether 91 HMMs representing it, and the P-loop containing nucleoside triphosphate hydrolases superfamily has 406 HMMs representing it.

Because each superfamily might be represented by multiple HMMs, there may be a high degree of overlap and redundancy among the models. However, there have not been any studies examining this issue systematically. To understand how the HMMs in the SCOP database are related to one another and the degree of overlap or redundancy among HMMs from either the same or different superfamilies, we perform a detailed analysis of the HMMs in SCOP for their similarity and relationships using a network approach. Specifically, we perform an all-against-all HHsearch for the library of HMMs in the SCOP database. HHsearch is similar to BLAST, except that instead of matching a sequence against a database of sequences, it uses a query HMM or sequence to match against a database of HMMs and identifies the HMMs significantly homologous to the query HMM or sequence [10]. We then construct a network of HMMs, where the link between two HMMs is based on their similarity, and examine some commonly evaluated network properties. We compare the current network with previously documented networks and outline some questions for future research.

2. Methods

The SCOP library of HMMs was downloaded from the SCP website (<http://scop.mrc-lmb.cam.ac.uk>), where the SCOP version was filtered to 70% maximum pairwise sequence identity. The library contains a total of 13,730 HMMs, from seven classes *a, b, c, d, e, f, g*, where class *a* contains only α (i.e., α helix) proteins, class *b* contains only β (i.e., β sheet) proteins, class *c* contains α and β proteins (mainly parallel β sheets (*beta-alpha-beta* units)), class *d* contains α and β proteins (mainly antiparallel β sheets, i.e., segregated α and β regions), class *e* contains multi-domain proteins (i.e., folds consisting of two or more domains belonging to different classes), class *f* contains membrane and cell surface proteins, and class *g* contains small proteins. It is useful to mention that the SCOP domain classification ID specifies the entire hierarchy, e.g. c.1.1.1, the first field is for the class *c*, second for the fold, third for the superfamily, and the last for the family.

HHsearch [10] was performed for all-against-all HMMs with the default parameters. HHsearch, similar to BLAST, uses a query that can be either a protein sequence or an HMM to search a database of sequences or HMMs and identify homology between the query and sequences and HMM models in the databases that is above a given threshold. In the current study, the e-value, a measurement of homology similar to BLAST's e-value, was set to 0.001. This e-value cutoff has also been used by Pfam to identify a Pfam clan [2], which is essentially equivalent to the superfamily hierarchy. A total of 13,547 HMMs have matches that met the criterion, with 1,618 having no other matches except themselves. Thus, 11,929 HMMs were used for the subsequent network analysis.

To study the relationship of the HMMs, an undirected network $G = (V, E)$ was constructed, where the vertices V are HMMs, and there is an edge in E between two HMMs if their e-value is below the threshold. General network statistics were computed, and a quadratic function was fitted to the log-log degree distribution. Three common vertex centrality measurements, degree centrality, betweenness centrality, and closeness centrality, were computed to evaluate the importance of vertices in the network. The degree of a vertex a is the number of edges incident on a . Betweenness for a vertex a ,

$$b(a) = \sum_{\substack{s, t \in V \\ s \neq a \\ t \neq a}} \frac{\sigma(s, t | a)}{\sigma(s, t)}, \quad (1)$$

introduced in Freeman [3], measures roughly the number of shortest paths going through a . $\sigma(s, t)$ is the number of shortest paths between vertices s and t , and $\sigma(s, t | a)$ is the number of shortest paths between vertices s and t that go through a . Thus, the higher the betweenness of a vertex, the more central/important the vertex is. In a fully connected

network, the betweenness of all vertices is 0.

The closeness centrality measures the number of steps required to access every other vertex from a given vertex, specifically, the closeness of a vertex a , $c(a)$, is computed by

$$c(a) = \frac{|V| - 1}{\sum_{\substack{i \in V \\ i \neq a}} d_{a,i}}, \quad (2)$$

where $d_{a,i}$ is the length of the shortest path between vertex a and vertex i . Closeness ranges from 0 (does not reach 0) to 1; the higher it is for a vertex, the more "central" the vertex is. These centrality measurements have different motivations and show different aspects for the importance of vertices in a network.

The network clustering coefficient, C , also known as transitivity, measured by the ratio between the number of triangles and the number of connected triplets, was computed for the entire network. The number of connected components that are trees, where there are N vertices but only $N - 1$ edges between the vertices, was computed for the entire network as well.

To systematically study the consistency between the e-value cutoffs for the prediction of whether or not HMMs belong to the same hierarchical level and classification of the SCOP database, we examined the Receiver Operating Characteristic (ROC) curves for the prediction of the hierarchical categories of two HMMs provided by different e-value cutoffs. The ROC curve shows how the true positive rate changes with the false positive rate for a classification. Specifically, for example, at the family level, if a sample of two HMMs were classified to the same family by the SCOP database, the prediction based on a specific e-value cutoff is considered to be a false negative (FN) if the e-value similarity of the two HMMs is worse/higher than the e-value cutoff, a true positive (TP) if the e-value is better (i.e., lower) than the cutoff; if the two HMMs were not classified to the same family by the SCOP database, the prediction based on the specific e-value cutoff is considered to be a true negative (TN) if the e-value similarity of the two HMMs is worse/higher than the e-value cutoff, a false positive (FP) if their e-value is better (i.e., lower) than the cutoff. Similar rules were applied to classify each pair of HMMs into the four categories (TP, FP, FN, and TN), for the four hierarchies, class, fold, superfamily, and family. True positive rate (i.e., sensitivity) was calculated as

$$TPR = \frac{TP}{TP + FN}, \quad (3)$$

and false positive rate (i.e., $1 - \text{specificity}$) as

$$FPR = \frac{FP}{FP + TN}. \quad (4)$$

An ROC curve was plotted for the four levels (i.e., class, fold, superfamily, and family) with different e-value cutoffs ranging from 10^{-20} to 10^{-3} .

Table 1: The general statistics of the HMMs

Class	Number of HMMs	Number of folds	Number of superfamilies	Number of families
a	1975	157	262	506
b	2590	109	231	485
c	3391	120	194	686
d	2932	223	328	683
e	199	34	34	51
f	145	29	44	50
g	697	49	70	112
All	11929	721	1163	2573

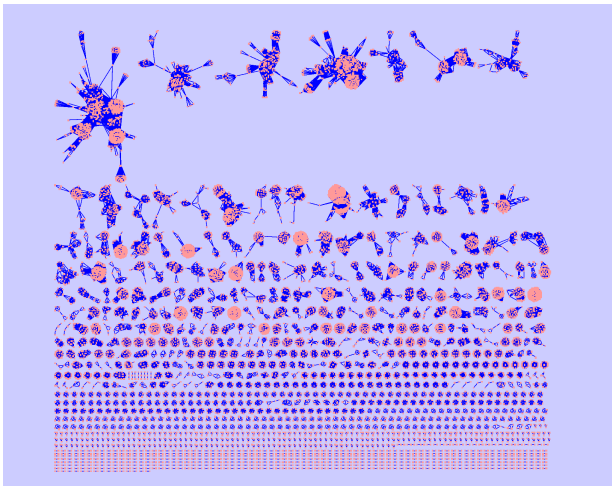


Fig. 1: The HMM network

3. Results

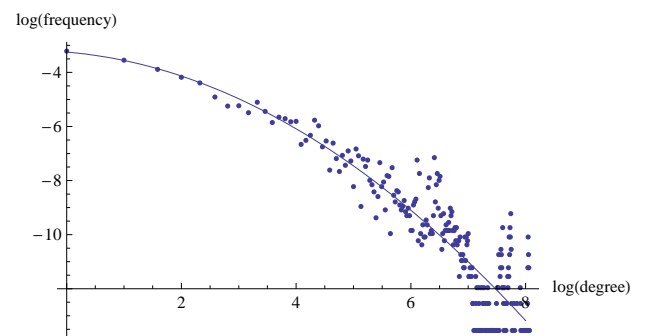
The working hypothesis. Taking into account the processes that built the HMMs and the hierarchical classification of the HMMs in the SCOP database, we hypothesize that the network should reflect this process, i.e., *the HMMs in a connected component belong to the same family or superfamily more often than expected under a random network connection model.*

General statistics of the HMMs and their network. A general description of the HMMs used to construct the network is shown in Table 1. There are seven classes in the collection of HMMs, falling into 721 folds, 1163 superfamilies, and 2573 families. Class *c* has the highest number of HMMs (3391) and class *f* the fewest (145).

The entire HMM network is shown in Figure 1, where the e-value cutoff is 0.001. There are altogether 151,461 edges for the 11,929 vertices. A significant property shown in Figure 1 is that the entire network is highly disconnected, with many much smaller connected components. In fact, there are altogether 1524 connected components (CCs). The smallest CC contains two vertices, the largest 590 vertices, $566/1524 = 37\%$ contain only two vertices and about 73% contain five or fewer vertices. The median CC size is 3 and the mean 7.8. The top 20 largest CCs are listed in Table 2.

Table 2: The 20 largest CCs and their densities

Size rank	Number of vertices	Density
1	590	0.12
2	349	0.21
3	277	0.65
4	155	0.15
5	141	0.38
6	121	0.33
7	120	0.19
8	106	0.72
9	99	0.84
10	90	0.95
11	86	0.99
12	85	0.89
13	81	0.32
14	80	0.83
15	74	0.66
16	73	0.65
17	72	0.16
18	70	1.00
19	69	0.97
20	66	0.40
All	11929	0.002

Fig. 2: Log-log degree distribution. The base is 2. The best fitting quadratic curve is $3.2481 - 0.176557x - 0.133088x^2$.

Degree distribution. The degree of the HMM network ranges from 1 to 268, with the average of 26 and median of 10. The log-log degree distribution is shown in Figure 2. It is evident that a power law distribution does not fit the data. The best fitting quadratic curve is also plotted with the data. It provides a relatively good fit for the smaller values of $\log(\text{degree})$, and then towards the larger degrees, the fit is not so good.

Network Density. Density, computed as the number of edges over the number of all possible edges (in a fully connected graph), provides some quantitative evaluation on the connectivity of a network. The density of the entire network is low, only $0.002 = 151461 / \binom{11929}{2}$. In contrast, individual CCs tend to have high densities, with more than 82.5% of CCs having density greater than 0.95. 1236 CCs are fully connected, i.e., cliques, with the largest clique of size 70.

Thus, individual CCs tend to have very high connectivity, whereas the entire network is not well connected. The

Table 3: The 20 HMMs with highest degree

Rank	HMM ID	SCOP ID	Degree
1	d1n26a1	b.1.1.4	268
2	d1f2qa1	b.1.1.4	265
3	d1qz1a3	b.1.1.4	265
4	d1biha1	b.1.1.4	264
5	d1rhfa1	b.1.1.1	263
6	d1tmna_	b.1.1.4	263
7	d2aw2a1	b.1.1.1	262
8	d1nbqa1	b.1.1.1	262
9	d1x44a1	b.1.1.4	262
10	d1biha3	b.1.1.4	262
11	d1cs6a3	b.1.1.4	262
12	d1f2qa2	b.1.1.4	261
13	d2avga1	b.1.1.4	261
14	d1epfa1	b.1.1.4	261
15	d3b5ha1	b.1.1.4	261
16	d1cs6a2	b.1.1.4	261
17	d1f97a2	b.1.1.4	261
18	d1epfa2	b.1.1.4	260
19	d2dava1	b.1.1.4	260
20	d1f97a1	b.1.1.1	260

Table 4: The 20 HMMs with largest betweenness

Rank	HMM ID	SCOP ID	Betweenness
1	d1bg6a2	c.2.1.6	14915.8
2	d1o8ca2	c.2.1.1	14665.7
3	d1e5qa1	c.2.1.3	14504.0
4	d2bzga1	c.66.1.36	9557.9
5	d3bswa1	b.81.1.8	9168.0
6	d1vj0a2	c.2.1.1	8211.0
7	d1ks9a2	c.2.1.6	7469.9
8	d2bmfa2	c.37.1.14	7439.8
9	d2dt5a2	c.2.1.12	7410.7
10	d1pjca1	c.2.1.4	7325.1
11	d1gtea4	c.4.1.1	7165.3
12	d1gu7a1	b.35.1.2	6768.0
13	d1tt7a1	b.35.1.2	6768.0
14	d2f1ka2	c.2.1.6	5985.2
15	d1ebfa1	c.2.1.3	5959.8
16	d1jqba2	c.2.1.1	5313.1
17	d1gr0a1	c.2.1.3	5220.0
18	d1ye8a1	c.37.1.11	5207.7
19	d1piwa2	c.2.1.1	4556.8
20	d1hdoa_	c.2.1.2	4403.8

density of the 20 largest CCs is shown in Table 2. The largest CC with 590 vertices has the lowest density, and the 18th largest CC with 70 vertices has a density of 1, and is therefore a fully connected component. There is a significant negative correlation between CC size and density (Kendall's rank correlation $\tau = -0.43$, p -value $< 2.2 \cdot 10^{-16}$ for CC size > 2).

Vertex centrality. Two centrality metrics, degree and betweenness, were computed for the vertices in the entire HMM network. Table 3 shows the top 20 HMMs that have the highest degrees. These 20 HMMs all belong to the same superfamily, b.1.1, Immunoglobulin, and also to the third largest CC that has 277 vertices. Thus, these 20 HMMs are connected with almost all other HMMs in the third CC. The HMM d1n26a1 (SCOP ID b.1.1.4, (A:1-93)) has the highest degree, 268, belonging to the Interleukin-6 receptor alpha chain, N-terminal domain (Homo sapiens).

Table 4 shows the top 20 HMMs that have the highest betweenness. Thirteen of the 20 HMMs belong to the superfamily c.2.1 (NAD(P)-binding Rossmann-fold domains), two to the superfamily b35.1.2, and two to the superfamily c.37.1. Eighteen of the 20 HMMs belong to the largest CC and the two remaining (c.37.1.14 and c.37.1.11) to the second largest. The HMM d1bg6a2 (SCOP ID c.2.1.6, (A:4-187)) has the highest betweenness, 14916, belonging to N-(1-D-carboxylethyl)-L-norvaline dehydrogenase (Arthrobacter, strain 1c). Interestingly, there is no overlap of HMMs that have the highest of both degree and betweenness.

Network diameter. The diameter of the largest CC (containing 590 vertices) is 9. The average distance between the vertices is 2.94. We also measured the diameters of all the CCs to see how they change as a function of CC size. Figure 3 shows that larger CCs tend to have larger diameters. However, smaller CCs can have large diameters as well. For

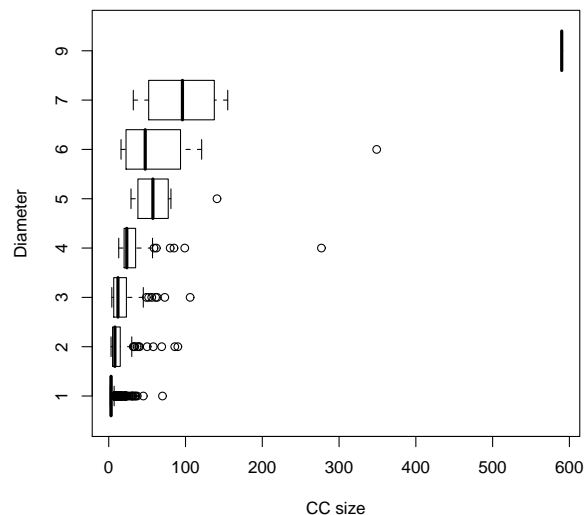


Fig. 3: Boxplot for the diameter of CCs as a function of CC size. The box marks the lower and upper quartile of CC sizes with the same diameter, the dark line marks the median, the whiskers mark the border of lower and upper outliers with the dots outside denoting the outliers.

example, a CC of size 32 has diameter seven, the same as a CC of size 155; a CC size of 16 has diameter six, the same as a CC of size 121. There are 1236 CCs with diameter 1, corresponding to the number of cliques.

CCs and hierarchy. Within the CCs, we examined whether the HMM members are from the same family, superfamily, fold, or class. There are altogether 1178 CCs whose members have the same SCOP domain classification

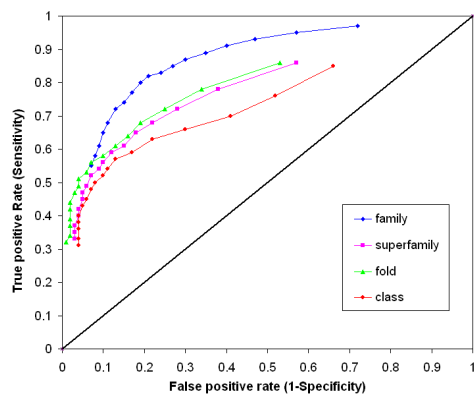


Fig. 4: The ROC curves for family, superfamily, fold, and class with different e-value cutoffs. For each curve, the data points from left to right correspond to the FPR and TPR for the e-value cutoffs from 10^{-20} to 10^{-3} .

(conserved at all hierarchical levels), 271 CCs whose HMMs belong to the same superfamily but to different families, 24 whose members belong to the same fold, but to different superfamilies, 18 whose members belong to the same class but have different folds, and the remaining 33 whose members are from different classes.

The consistency between the prediction of HMM memberships at different hierarchical levels in the SCOP database based on the e-value cutoffs and the classification of the SCOP database was evaluated by ROC curves, shown in Figure 4. We make several observations. First, for all four levels of the hierarchy, the higher the e-value cutoff, the higher the sensitivity (true positive rate), so is the false positive rate, which is expected because higher e-value means a less stringent prediction criterion that in turn leads to a higher number of true positive predictions, and also a higher number of false positive predictions. Meanwhile, the rate of increase in sensitivity outpaces the rate of increase in the false negative rate as the e-value becomes more stringent, suggesting that beyond a certain e-value cutoff, the HMMs belonging to the same hierarchical levels also tend to have high similarity, which make them robust to the e-value cutoff change. Second, the curves for the prediction of fold and superfamily are very similar to each other, indicating that for the same e-value cutoff, the predictions for whether two HMMs belong to the same fold or superfamily are similar. In fact, for the same e-value cutoff, the difference in true positive rate (sensitivity) between the fold and superfamily ROC curves is either 0 or 0.01, and the difference in false positive rate (1-specificity) falls within the narrow range [0.01 – 0.04]. Third, the prediction quality is the worst for class as compared to the other three levels, with worst sensitivity and specificity for the same e-value cutoffs. This might not be so surprising as classification at the class level

is more for convenience than for biological reasons.

4. Discussion

The important HMMs. In this work, we used three centrality measurements to evaluate the importance of an HMM. The results show that from the entire network, the vertices with the highest degrees do not necessarily have the highest betweenness, and vice versa. Degree measures how many immediate neighbors one HMM has, and therefore, the more it has, the more central it is. The vertices with the 20 largest degrees are all from the third largest CC, and are connected to about 94% of its vertices. The vertices with the 20 largest betweenness values are from either the largest CC or the second largest CC. Since betweenness reflects how essential one vertex is to the connection of any other two vertices in the graph, in the case of HMMs, it may reflect the possibility that one HMM is the *hybrid* of two HMMs, that is, between the two HMMs, there is no significant similarity, but through the one HMM, the HMMs can be linked. Biologically, this idea seems to reflect hybrid or mosaic proteins where one protein contains domains from multiple proteins. To our knowledge, the idea of hybrid HMMs has not been discussed previously and deserves more research attention. Moreover, we hypothesize that the HMMs with high centrality measurements may be better able to pick up the sequences that belong to the superfamily than the more peripheral HMMs. Future studies can be directed to test this hypothesis.

Comparison with other networks. The largest CC (590 vertices) of the current network has a diameter of 9 and the average distance between its vertices is 2.94. This bears some similarity to the protein interaction network [6], whose largest CC (containing 5,128 vertices) also has the same diameter of 9, but a larger average distance of 3.68. Thus, the protein interaction network seems to have more vertices that are a bit more spread out, which contributes to a larger average distance. To this point, it is very interesting that despite the big difference in the sizes of the two CCs of the two networks, the diameters are the same.

It is evident that the HMM network is highly clustered. In fact, its clustering coefficient is 0.85, which, to our knowledge, is the highest among the biological networks that have been studied so far. As shown by Newman [9], the undirected networks that tend to have high clustering coefficients are social networks. For example, the film directors network has a clustering coefficient of 0.20 and coauthorship networks for math, physics, and biology disciplines are 0.15, 0.45, and 0.088, respectively, whereas biological networks such as metabolic network and protein interaction network have only a clustering coefficient of 0.09 and 0.07, respectively. The comparison indicates that the current network has distinct features from the previously characterized real-world networks. Also, consistent with its high clustering coefficient, the network has altogether 585 trees (i.e., the CCs of size n

with $n - 1$ edges), most of which (566) are of size 2, 15 of size 3, and four of size 4.

Testing the working hypothesis. The results show strong evidence that HMMs in a connected component tend to represent the same family or superfamily. Among the total 1524 CCs, more than 77% have only members from the same family; more than 95% have only members from the same superfamily. Thus, there is overwhelming evidence supporting our working hypothesis that HMMs belonging to the same family or superfamilies tend to cluster together in the network. However, to formally evaluate this and provide some statistical support, we also simulated 10000 random networks while preserving the degree distribution and the number of connected components. Among the 10000 simulated random networks, the highest proportions of CCs having only members from the same family and superfamily are as low as 0.5% and 0.7%. This shows that in the observed network, the HMMs from the same family or superfamily do have a strong tendency to cluster, agreeing with our working hypothesis.

5. Conclusion

In this paper, we examined the properties of the network constructed for HMM models in the SCOP protein structural classification database. A number of questions remain to be addressed in future research. For example, can we devise a computational method to measure or evaluate the degree of redundancy or overlap between HMM models that are used to represent the same superfamily? This research is meaningful given the ever increasing number of large-scale genomic sequences (thereof more protein sequences). Given that we can measure the redundancy of the HMMs of a superfamily, the logical question becomes, can we computationally reduce the redundancy of the HMM library, e.g., possibly by constructing super-HMMs, each of which represents a collection of redundant HMMs, so that a protein sequence is scanned against a reduced set of HMMs (super-HMMs) rather than the entire set of HMMs that have overlaps and redundancies? Finally, because the HMM network shows distinct properties from many documented networks as discussed above, can we propose a theoretical model to better account for the observations in the current network? Moreover, as our HMM network is also weighted, with edges quantifying the similarity between two HMMs, future proposed models can also consider the incorporation of weighted edges into the network.

Acknowledgment

The authors thank T Murali for discussion. The work was partially supported by NSF Grant IIS-0710945 and AFRL Grant FA8650-09-2-3938 and AFOSR Grant FA9550-09-1-0153.

References

- [1] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res*, 36(Database issue):D419–25, 2008.
- [2] R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. Pfam: Clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–51, 2006.
- [3] L.C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.
- [4] J. Gough and C. Chothia. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res*, 30(1):268–72, 2002.
- [5] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol*, 313(4):903–19, 2001.
- [6] E.D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.
- [7] L. Lo Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Res*, 30(1):264–7, 2002.
- [8] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.
- [9] M. E. Newman. *Networks: An Introduction*. Oxford University Press. Inc., New York, NY, USA, 2010.
- [10] J. Soding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–60, 2005.

Correlation of Patristic Distance with Nominal Specimen Collection Date in Influenza A/H1N1 Neuraminidase-Encoding Segments

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA

Abstract

Neuraminidases are viral coat glycoproteins that facilitate the transmission of influenza from cell to cell. Characterizing the evolution of the neuraminidases is essential to effective development and deployment of neuraminidase-inhibitor therapeutics. Here, I describe a linear regression of patristic distance in Influenza A/H1N1 neuraminidase-encoding segments on the nominal specimen-collection date contained in the label field of the neuraminidase genomic sequence descriptors; the regression predicts an average mutation rate of ~1 bp/year (implying, on average, ~0.1 mutations in the neuraminidase active site per year).

Keywords: Influenza, H1N1, neuraminidase

1.0 Introduction

The most widely used anti-influenza therapeutic, oseltamivir (Tamiflu™, [4]), a neuraminidase inhibitor, was decreasingly effective against the dominant influenza strain (an Influenza A/H1N1 mutant) in the US in the 2009 "Spring/Fall" pandemic ([10]). Characterizing the evolution of the neuraminidases is essential to effective development and deployment of neuraminidase-inhibitor therapeutics.

Influenza type A is divided into nine sero-subtypes, and these subtypes correspond at least roughly to differences in the active-site structures of the corresponding neuraminidases. The subtypes fall into two groups ([3]): group-1 contains

the subtypes N1, N4, N5 and N8, whereas group-2 contains the subtypes N2, N3, N6, N7 and N9. Oseltamivir was designed to target the group-2 neuraminidases.

The known molecular structures of the neuraminidases are broadly consistent with this sero-taxonomic characterization. The available crystal structures of the group-1 N1, N4 and N8 neuraminidases ([1]) reveal that the active sites of these enzymes have a very different three-dimensional structure from that of group-2 enzymes. The differences lie in a loop of amino acids known as the "150-loop", which in the group-1 neuraminidases has a conformation that opens a cavity not present in the group-2 neuraminidases. The 150-loop contains an amino acid designated Asp 151; the side chain of this amino acid has a carboxylic acid that, in group-1 enzymes, points away from the active site as a result of the 'open' conformation of the 150-loop. The side chain of another active-site amino acid, Glu 119, also has a different conformation in group-1 enzymes compared with the group-2 neuraminidases ([13]).

The Asp 151 and Glu 119 amino-acid side chains form critical interactions with neuraminidase inhibitors. For neuraminidase subtypes with the "open conformation" 150-loop, the side chains of these amino acids might not have the precise alignment required to bind inhibitors tightly ([13]).

The difference in the active-site conformations of the two groups of

neuraminidases may also be caused by differences in amino acids that lie outside the active site. This means that an enzyme inhibitor for one target will not necessarily have the same activity against another with the same active-site amino acids and the same overall three-dimensional structure ([17]).

A first-principles theory of neuraminidase evolution is highly desirable but currently beyond the state of the art. First-principles computational methods such as molecular dynamics could provide insight into relevant drug-site free-energetics, but such methods are often computationally expensive and in the case of the neuraminidases, would require an initial, realistic specification of the *in situ* environment. Relatively few H1N1 neuraminidase structures are available at present, and none address the effect of the molecules' environment on their active sites. In contrast, phylogenetic comparisons of the genomic encoding of the neuraminidases might, by translational proxy, provide insight; some phylogenetic methods, furthermore, are computationally inexpensive. ~6800 neuraminidase-

encoding (NA) segments of the viral genomes are available for A/H1N1 ([7]).

2.0 Method

The general method of this study has four steps: downloading H1N1 NA segment descriptors, aligning the descriptors, computing the patristic distances among the segments, and analyzing the correlation of segment patristic distance with segment collection-date. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment, connected by a 1.5 Mbit/s DSL link to the Internet.

Because typical influenza neuraminidases are ~400-mers, ~1200 base-pairs (bp, 3 bp per mer) are required to encode them in the viral genome. Influenza H1N1 NA segments of length at least 1000 bp were downloaded from the Influenza Research Database ([7]) on 13 January 2011. The query/download parameters are shown in Figure 1.

Query parameters:

```
Select Segments: 6 (NA)
Subtype: H1N1
Date Range: 1915 to 2011
Geographic Grouping: All
Host: All
Data to Return: Segment/Nucleotide
```

Advanced Options:

```
Minimum segment length: (Segment 6) 1000
Display Fields: Sequence Accession, Date
```

```
Display: sort on (increasing) date
```

Download parameters:

```
Select: All segments
Select Download Type: Segment FASTA
Label Sequence By: Custom -- Accession Number, Date
```


Figure 1. Influenza Research Database ([7]) query/download parameters for the Influenza A/H1N1 NA segment descriptors used in this study.

The "Label" fields in the FASTA-formatted sequence descriptors obtained from the previous step were edited in *Word 2007* so that each had the form "GenBank_accession_ID-yyyy", where yyyy is the year represented in the Label. (In this paper, that year is called the "collection date". It should be noted that such a date is merely part of a free-text field; thus, that "date" could be, and mean, anything. It is relatively common practice, however, for such a date to represent the date on which the organism from which the sequence was derived was collected.) Any sequence descriptor that did not contain year information was subsequently deleted using *Word 2007*.

The file resulting from the previous step was edited in *BioEdit* v7.0.5.3 ([9]) to remove any sequences longer than 1450 bp. The *BioEdit* navigation for this filtering was

```
Sequence  -->  Filter  Out
Sequences  Containing  Certain
Characters --> Delete them  -->
are >x long (x = 1450) --> File
--> Save as (type = Fasta,
filename = ten.fasta)
```

If fewer than 10 sequences for a given year were in the resulting file, all sequence descriptors for that year were saved. Else, only the first 10 sequence descriptors in each year were saved. (This helps to reduce

time bias in the sample, some of which, due to the scarcity of specimens collected before 1930, is unavoidable). The result was a collection of FASTA-formatted sequence descriptors 1000-1450 bp long. *BioEdit* was then used to save the descriptor Labels to a separate file.

The FASTA-formatted sequences from the previous step were aligned using *MAFFT* v6.847b-win32 ([5]), invoked from a *Vista* Command Prompt window. The parameters for the alignment were

```
Order: input
Output format: clustal
Strategy:FFT-NS-i
          (Standard)
Iterative refinement
          (Maximum of 2 iterations)
All other parameters:
          defaulted
```

The resulting CLUSTAL-formatted ([16]) file was edited in *Word 2007* to remove blank lines and lines containing asterisks.

A *PAUP* ([12]) neighbor-joining (NJ, [18]) script was built in *Notepad*, incorporating the descriptor labels and aligned sequences obtained in previous steps. Hyphens in the descriptor labels were replaced by underscores. The template for the *PAUP* script is shown in Figure 2.

```
#NEXUS
begin taxa;
  dimensions ntax=385;
  taxlabels
  [descriptor labels go here (not shown)]
;
end;
```

```

begin characters;
  dimensions nchar=1477;
  format missing=? gap=- matchchar=. interleave datatype=dna;
  matrix
    [aligned data goes here (not shown)]
;
end;

begin paup;
  [1] log start file=H1N1_NA_nj_patdist.log replace;
  [2] nj;
  [3] savedist file=tenpatdist.txt format=oneColumn;
end;

```

Figure 2. Template of PAUP script used to obtain the patristic distances used in this study.

Patristic distances from a 1918 "reference" segment (AF250356 in [7]), and corresponding label-times expressed as years-since-1918, were extracted using the *get_pats* software ([11]) running under *Cygwin* (in turn running under *Vista*) from the patristic distance file produced by

PAUP. The output of *get_pats* is a comma-separated file. This file was converted to a space-separated file using *Notepad*. A linear regression of patristic distance on time was performed by the *Mathematica* ([8]) script shown in Figure 3 ([14]).

```

patdistimedata = ReadList[ToFileName[{"C:",
  "BIOCOMP2011", "Influenza", "Branch_and_age"},
  "tenpatdistime.txt"], {Number, Number}];

model=LinearModelFit[patdistimedata,x,x]

model["BestFit"]

Show[ListPlot[patdistimedata, AxesOrigin -> {0,0},
  AxesLabel -> {"Years After 1918", "Patristic Distance from
  AF250356"}], Plot[model["BestFit"], {x, 0, 100}]]

model["ParameterTable"]

model["RSquared"]

model["AdjustedRSquared"]

```

Figure 3. Mathematica script used for linear regression in this study.

3.0 Results

6816 sequences were produced by the Influenza Research Database query/download described in Section 2.0. 23 sequence descriptors in this file had no identifiable date information and were deleted, netting a file containing 6793 sequence descriptors.

The time-debiasing step in *BioEdit* yielded 385 FASTA-formatted sequences.

The *MAFFT* alignment step described in Section 2.0 yielded 385 CLUSTAL-formatted sequence descriptors with 1477 characters per sequence. 384 patristic-

distance/time pairs were produced by the patristic-distance/time extraction (via *get-pats*) from the patristic distance file produced by *PAUP*.

The linear regression computed by *Mathematica* was

$$\begin{aligned} \text{patristic_distance_from_AF250356} \\ = 0.00113267 * \text{years_since_1918} \\ + 0.0497421 \end{aligned}$$

A scatterplot and the best linear fit to that data is shown in Figure 4.

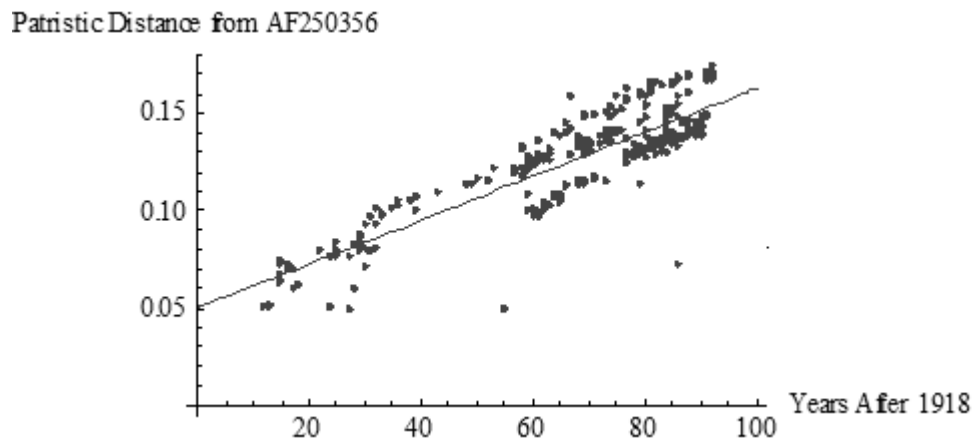


Figure 4. Scatterplot and best linear fit of patristic-distance/time data used in this study.

Some parameter statistics for this regression are:

Parm	Estimate	Standard Error	t Statistic	P-Value
b	0.0497421	0.00211585	23.5092	3.16958×10^{-76}
m	0.00113267	0.0000301765	37.535	3.17033×10^{-130}

where b is the intercept on the patristic-distance axis, and m is the slope of the regression. The regression coefficient, r^2 , is 0.786696; the adjusted r^2 , 0.786138.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The slope of the regression line suggests that the typical Influenza A/H1N1 NA segment experiences, on average, ~ 0.001 change per year. Since an NA segment has length ~ 1000 bp, we would, based on the regression formula in Section 3.0, expect $(\sim 1000 \text{ bp} \times \sim 0.001) \sim 1$ bp change per year. Such a change would be sufficient to alter at least one amide in the active site of the neuraminidase encoded by the segment about every 7 years, if we assume the active site is determined by ~ 50 bp and that mutations are uniformly distributed across the molecule. This rate is consistent with the nominal mutation rate suggested by other considerations ([15]).

In general, we could not expect "collection date" to provide any information about mutation rate. However, if specimens are collected at a rate that is comparable to the mutation rate (as is the case with flu genomic segments), collection dates will tend to exhibit a strong correlation with mutation rates.

2. The regression reported in Section 3.0 has robust significance statistics, strongly suggesting that current flu genomic segment sampling and sequencing practices are sufficient to characterize the average mutation rate of the H1N1 NA segments.

3. The sequence-descriptor sampling protocol described in Section 2.0 is intended to help mitigate time-biasing in the sample by restricting the number of sequence descriptors sampled per year to no more than 10. The results aren't perfect: for some years, [7] contains fewer than 10 (for some years, no) sequence descriptors. Other protocols are of course possible, but the one used in this study is a practical compromise between under-, or over-, sampling any given year, given the data available in [7].

5.0 Acknowledgements

This work benefited from discussions with Town Peterson of the University of Kansas Biodiversity Institute, George Hrabovsky of the Madison Area Science and Technology Institute for Scientific Computing, Tony Pawlicki, and Richard Barrett. For any problems that remain, I am solely responsible.

6.0 References.

- [1] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49.
- [2] Barry JM. *The Great Influenza*. Viking. 2004.
- [3] World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization* 58 (1980), 585-591.
- [4] Ward P et al. Oseltamivir (Tamiflu) and its potential for use in the event of an influenza pandemic. *Journal of Antimicrobial Chemotherapy* 55, supplement 1 (2005), i5-i21.
- [5] Katoh K and Toh H.. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* 9 (1 July 2008), 286-298.
- [6] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.
- [7] Squires B, Macken C, A. García-Sastre A, Godbole S, Noronha J, Hunt V, Chang R, Larsen CN, Klem E, Biersack K, and Scheuerrmann RH. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. *Nucleic Acids Research* 36

- (Database issue), D497-503 (2008). <http://www.fludb.org/brc/home.do?decorator=influenza>.
- [8] Wolfram Research. *Mathematica Home Edition* v7.0 (2010).
- [9] Hall TA. *BioEdit*: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41 (1999), 95-98.
- [10] US Centers for Disease Control. *Summary: Interim Recommendations for the Use of Influenza Antiviral Medications in the Setting of Oseltamivir Resistance among Circulating Influenza A (H1N1) Viruses, 2008-09 Influenza Season*. 19 December 2008. URL <http://www.cdc.gov/flu/professionals/antivirals/summary.htm>.
- [11] Horner JK. *get_pats*, a perl program for extracting patristic distances from a PAUP "one-column" patristic distance file. 2011.
- [12] Swofford D. *Phylogenetic Analysis Using Parsimony (PAUP)* v4.0b10. URL <http://paup.csit.fsu.edu/>. Sinauer Associates. 2004.
- [13] Luo M. Structural biology: antiviral drugs fit for a purpose. *Nature* 443 (7 September 2006), 37-38. doi:10.1038/443037a, published online 6 September 2006.
- [14] Horner JK. *statpats.nb*, a *Mathematica* notebook for performing linear regression of patristic-distance on time. Available from the author on request.
- [15] Horner JK. An estimate of the mutation rates of the active sites of Influenza A/H5N1 neuraminidases. *Proceedings of the 2010 International Conference on Bioinformatics and Computational Biology*. CSREA Press. 2010. pp. 344-349.
- [16] Higgs DG, Thompson JD, and Gibson TJ. Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* 266 (1996), 383-402.
- [17] Stoner TD, Krauss S, DuBois RM, Negovetich NJ, Stallknecht DE, Senne DA, Gramer MR, Swafford W, DeLiberto T, Govorkova EA, and Webster RG. Antiviral susceptibility of avian and swine influenza virus of the N1 neuraminidase subtype. *Journal of Virology* 84 (October 2010), 9800-9809.
- [18] Felsenstein J. *Inferring Phylogenies*. Sinauer Associates. 2004.

GAPDH architecture at low guanidine concentrations: first derivative analysis of the descending slope of the UV absorbance peak

Jessica Y. Kim¹, Christopher S. Theisen¹ and Norbert W. Seidler¹

¹Department of Biochemistry, Kansas City University of Medicine and Biosciences, Kansas City, Missouri, USA

Abstract – Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), the glycolytic housekeeping enzyme, exists as an asymmetric tetramer. We previously observed that GAPDH can appear as a dimer and in higher order structures which we proposed may be a decamer. The monomeric subunit contains two rather distinct domains, which are referred to as the nucleotide-binding domain and the catalytic domain. These two domains occupy the N- and C-termini, respectively. We examined the denaturation of GAPDH in the presence of low concentrations of the chemical denaturant, guanidine-HCl (0.5-1.5M GdnHCl). Full denaturation of proteins typically require approximately 6M GdnHCl. At various concentrations of denaturant, UV absorbance spectra were obtained. The GdnHCl-dependent changes in the descending slope of the UV absorbance spectra were further examined by computing first derivatives of this region and monitoring changes in first derivative spectral peak and trough as a function of GdnHCl concentration. The observations of multiphasic changes are consistent with a model that suggests subunit separation is followed by domain-domain separation prior to domain unfolding.

Keywords: glyceraldehyde 3-phosphate dehydrogenase, UV absorbance spectra, guanidine-HCl, denaturation.

1 Introduction

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) is an indispensable protein that plays a pivotal role in glycolysis, a vital energy-generating pathway in all human cells. The glycolytic pathway consists of two stages. The first stage converts 1mol glucose to 2mols D-glyceraldehyde 3-phosphate (Glyc3P) in five enzymatic steps. Net energy is not generated until stage two, where 1mol Glyc3P is converted to 1mol pyruvate. The first enzymatic step of the second stage of glycolysis is catalyzed by GAPDH. The substrates are Glyc3P, inorganic phosphate (Pi) and NAD⁺, and the products are 2,3BPG and NADH. The reaction is an oxidative phosphorylation and involves a covalent intermediate between the substrate Glyc3P and an active site cysteine residue. The next reaction is catalyzed by phosphoglycerate kinase (PGK) that converts ADP to ATP, in which a phosphoryl transfer occurs from 2,3BPG to ADP. Therefore in the first two

reactions of the second stage of glycolysis, both NADH and ATP are made and become available for cellular processes. Increasingly, this general event has become understood as having specific purposes. For example, synaptic vesicles are equipped to load neurotransmitter, but in order to do this the vesicles are first acidified by the activity of a proton pump. This proton gradient drives the uptake of neurotransmitter into the vesicle. The proton pump is a vesicular ATPase which derives its ATP efficiently from a GAPDH-PGK complex [1]. The nature by which a cytosolic soluble protein like GAPDH becomes membrane bound is unknown. We recently proposed that GAPDH, which is typically described as an asymmetric tetramer, can appear as a dimer and in higher order structures which we propose may be a decamer [2]. Interestingly, the presence of inhaled anesthetics, such as isoflurane, shifts the equilibrium of the oligomer states towards the decamer [2], presumably through modulation of NAD⁺/NADH binding [3]. GAPDH is rather unstable, forming turbid solutions over time and is easily denatured upon heat exposure, showing a T_m (temperature at which 50% of the proteins are denatured) of 54.7°C [4]. This intrinsic tendency towards disorder may enable GAPDH to achieve diverse interactions with binding partners and thereby participate in alternate functions, such as interactions with vesicles, receptors and cytoskeletal components [5,6]. In order to develop a better understanding of the dynamic properties of GAPDH, we examined the changes in folded states of GAPDH at low concentrations of a chaotropic agent, guanidine (GdnHCl). By focusing on the descending slope of the UV absorbance spectra, we were able to use a computational approach to understanding the early steps of protein unfolding in GAPDH.

2 Materials and methods

Materials. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) from rabbit muscle [EC 1.2.1.12] was obtained from Sigma-Aldrich (G2267) and dissolved in a 50mM sodium phosphate, pH = 7.4, buffer containing 0.3mM EDTA that was prepared with deionized (Milli Q; 18.2 MΩ) water and passed through a 0.2μm nylon filter (Millipore Millex-GN) prior to analysis. Concentration was determined by absorbance at 280nm using the molar absorbance coefficient of

$149\text{mM}^{-1}\text{cm}^{-1}$ [7]. Guandine hydrochloride (GdnHCl) was also from Sigma-Aldrich (G4505) and freshly prepared prior to each experiment.

Treatment of GAPDH with GdnHCl. Incubations were done at room temperature as follows: samples (0.6mL) of GAPDH ($10\mu\text{M}$) were mixed carefully with 0, 0.5, 0.75, 1.0, 1.25 and 1.5M GdnHCl directly in quartz cuvettes. Final buffer concentrations were 30mM sodium phosphate, $\text{pH} = 7.4$, and 0.2mM EDTA. After thorough mixing, samples were assayed immediately.

UV Absorption Spectroscopy. Absorbance spectra (240-340nm; bandwidth 2.0nm; interval 0.5nm; lamp change at 325nm; scan speed 145nm/min) were obtained for the samples using a GE Healthcare Ultrospec 4000 spectrophotometer. The deuterium lamp had less than 100hrs of use. A reference sample that contained identical buffer without GAPDH was used before each scan. Spreadsheet data of the spectra were transferred to SigmaPlot 11.0 for further analysis

Computational Analysis. Difference spectra were first determined and compared. Absorbances from the experimental samples (GAPDH treated with GdnHCl) were subtracted from the control spectra to generate difference spectra. The downslope of the 280nm peak of the original spectra was closely examined. The descending component of the spectra from 283nm to 310nm was used to get first derivatives. Since the data was acquired at intervals of 0.5nm, regression lines were computed using 3 contiguous data points successively. With one data point shifts, the regression lines were overlapping and progressed from 283nm to 310nm, using SigmaPlot 11.0 to obtain slopes, which represented the tangent first derivatives. Since this component of the spectra is a downslope, all regressions were negative. To simplify analysis, absolute values were used and multiplied by 100, thus keeping all values positive. The resulting first derivatives were plotted against wavelength. The plots presented spectra that exhibit maxima and minima that were reliable observations. In the range of 286-300nm, a single trough was seen followed by a single peak. In order to determine the exact points that represent the maximum and minimum, we used the equation for computing the center of spectral mass (CSM), by integrating over the range that pertained to the peak and trough, respectively.

$$\text{CSM} = \frac{\sum (\lambda A_i) di}{\sum A_i di} \quad (1)$$

$(i = 0.5\text{nm})$

where λ is the wavelength (nm) and A is the absorbance at that wavelength. The CSM for the spectral nadirs (or troughs) and zeniths (or peaks) changed as a function of GdnHCl concentration. These values were then plotted against concentration of chaotropic agent to assess phasic response and transitions over this range of agent.

3 Results

The UV absorbance spectrum for control GAPDH showed a pattern typically seen for globular proteins, exhibiting a peak at approximately 280nm (Figure 1A) representing the aromatic residues. The absorbance below 240nm was ignored in this study. The absorbance in the 280nm range is in part due to the contribution of the microenvironment surrounding the aromatic residues, particularly tyrosines and tryptophans. Upon addition of GdnHCl, the change in protein architecture affects the microenvironment around these residues, altering the spectra as seen in Figure 1A. The spectral changes were slight but very reproducible.

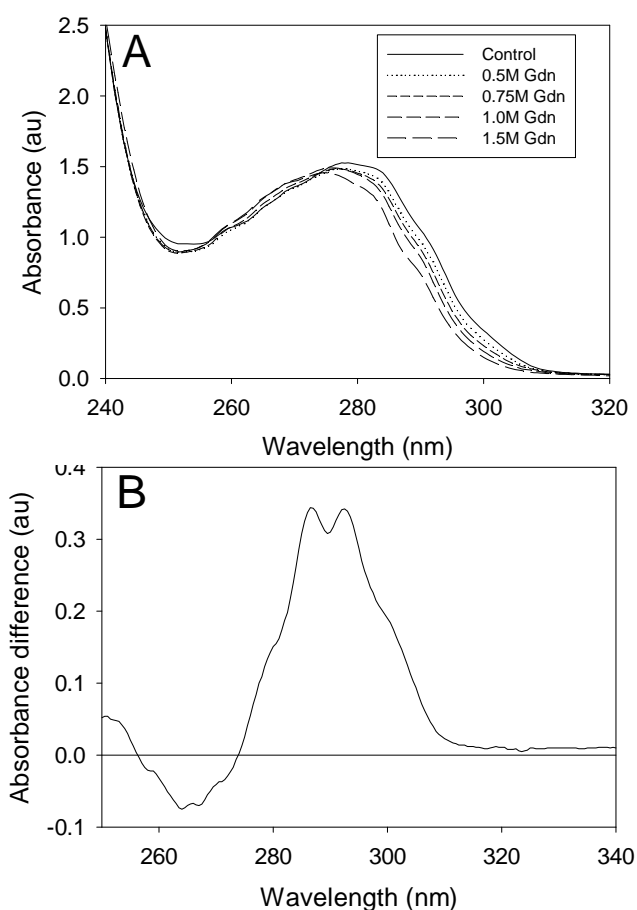


Figure 1: Spectral changes due to the chaotropic agent, GdnHCl. GAPDH ($10\mu\text{M}$) was incubated with various concentrations of GdnHCl. **A.** UV absorbance spectra in the 240 to 320nm range showing the effects of GdnHCl on the descending slope of the 280nm peak. **B.** The difference spectrum of control and GdnHCl(1.5M)-treated GAPDH.

Difference spectra were computed. Figure 1B presents the difference spectrum of control GAPDH and GAPDH-

treated with 1.5M GdnHCl. In this difference spectrum, we see the expected pattern of a trough with values below zero in the 260-270nm range and a peak (here, seen as a doublet) in the 285-295nm range.

The difference spectrum as shown in Figure 1B certainly provides information that is more readily visible than that found in the original spectra (Figure 1A). In fact, most studies that follow denaturation by absorption spectroscopy present difference spectra [8]. We extended this analysis to include an alternate approach, which involved examining first derivative conversions of the data.

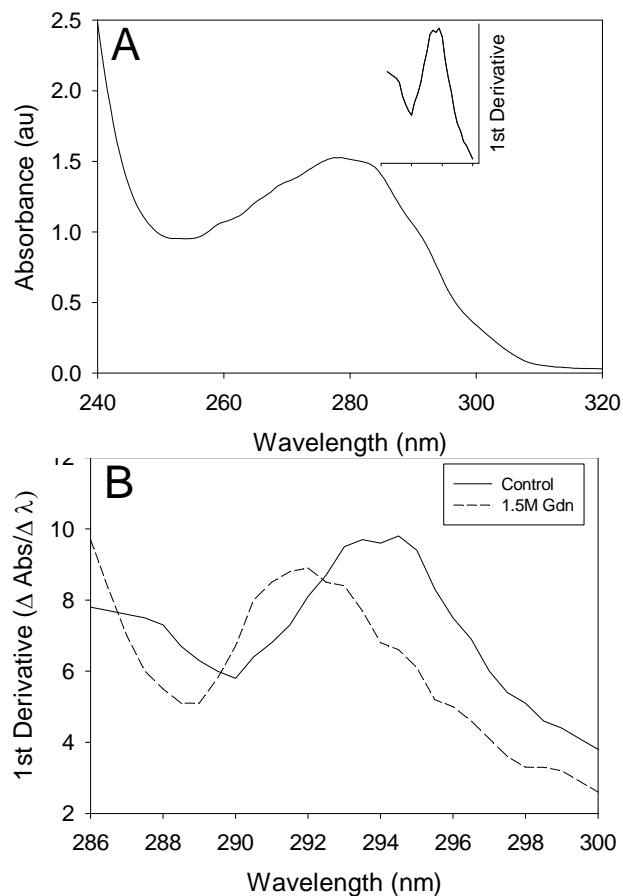


Figure 2: First derivative analysis of the descending slope of the 280nm peak. The data from Figure 1A were used to generate first derivatives. Control spectrum of untreated GAPDH was re-drawn in (A), showing the first derivatives over the 286-300nm region (inset, A), computed as described in *Materials and methods*. B. First derivatives were plotted as a function of wavelength for control and GdnHCl(1.5M)-treated GAPDH.

To pursue this aim of using an alternate approach to studying the effects of GdnHCl, we focused on the downward slope of the prominent 280nm peak from the original UV absorbance spectra. The first derivative spectrum

in this range is juxtaposed to the original control spectrum in Figure 2A. The first derivative spectrum reveals a trough (or nadir) followed by a peak (or zenith). Control first derivative spectra were then compared with those of GdnHCl-treated samples. Figure 2B presents the data from control GAPDH against the spectra of GdnHCl(1.5M)-treated GAPDH.

Visibly one notices a shift in the spectra. Since a single spectrum includes a nadir and zenith, both can be quantified as a value (λ min and λ max, respectively). These values can be tracked as a function of GdnHCl treatment. The wavelength at which the peak or trough actually reaches its exact center point was determined mathematically using center of mass (CSM) calculation, with correction of the data around the trough to invert the tracing, making calculations feasible.

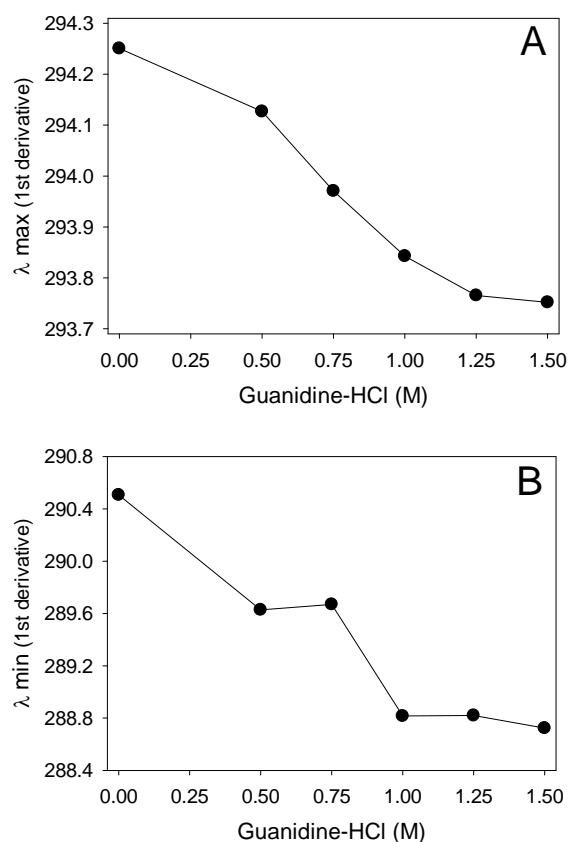


Figure 3: First derivative maxima and minima as a function of chaotropic agent, GdnHCl. First derivatives were initially computed for the downslope of the UV spectra; then the wavelengths representing zeniths (A) and nadirs (B) (λ max and λ min, respectively) were plotted as a function of GdnHCl concentration.

The wavelengths that represent the zeniths from the first derivative spectra were then examined as a function of concentration of chaotropic agent. The resulting plot (Figure 3A) exhibited two transition points: one at 0.5M and the other

at approximately 1.0M GdnHCl. We also observed that the nadir wavelengths plotted as a function of GdnHCl concentration showed three phases, an observation that was consistent with the computational data on wavelengths associated with the zenith.

Changes in secondary structure, which are indicative of domain unfolding, typically occurs at higher concentration of denaturant. For example, transferrin begins unfolding at approximately 1.6M GdnHCl [9]. The changes observed in this study, which looked at denaturant concentrations below 1.5M, were likely due to quaternary and tertiary changes that would occur prior to domain unfolding. It was previously shown that GdnHCl-induced dissociation and unfolding of a tetrameric enzyme exhibits multiple phases over a 0.5 to 5M GdnHCl concentration range [10].

4 Discussion

In looking at the conformational effects of chemical chaotropic agents, a difference spectrum, such as Figure 1B, provides discrete information that is readily visible and potentially quantifiable as compared with the slight differences found in the original spectra (Figure 1A). In fact, most of the studies in the literature on chemical denaturation and absorption spectroscopy presents the results as difference spectra [8]. The intention of this study, on the contrary, was to use an alternate approach to improve data analysis of the original UV absorbance spectra. Converting the downslope of the spectra to first derivatives does increase the information found within the original data and reveals definable transition points that are seen as nadir and zenith values that reliably shift in response to low levels of chaotropic agent.

Previous studies show changes that occur prior to domain unfolding, which may be attributed to subunit and domain interfacial disruption. Using circular dichroism (CD) measurements of porcine serum transferrin at various concentrations of GdnHCl, two conformational transitions are evident [9], one at 1.6M and the other at 3.4M. Additionally, Fe²⁺-binding is completely lost by 1.7M GdnHCl. There was a transition at approximately 1.2M, showing a slow trajectory of iron-binding loss from 0 to 1.2M GdnHCl followed by a faster trajectory from 1.2 to 1.7M GdnHCl. In this case, unfolding of secondary structure occurs above the CD-observable transition at 1.6M GdnHCl. Transferrin is a homodimer linked via a disulfide bond [11]. Each subunit consists of two domains, the N-lobe and the C-lobe with each lobe capable of binding a Fe²⁺ atom [12]. Each lobe is comprised of two subdomains. Upon more careful analysis of the authors' presentation of the iron binding data, we recognized that the process of iron-binding loss appeared to be multiphasic. The authors identified two phases that they attributed to sequential release of the two iron atoms. We think that the data shows three phases and that these phases represent changes in protein architecture that occurs at low GdnHCl concentrations. We propose that these architectural

changes would correspond to iron binding loss. Their data (fig. 4b, in [9]) suggests that there are no changes to iron bound to transferrin from 0 to 0.2M GdnHCl, a rapid change from 0.2 to 0.5M, a then slower change from 0.5 to 1.2M, and then again a fast change from 1.2 to 1.7M, as visually inspected by us in the present study with acknowledgment that these values are estimates. First derivatives were calculated and were approximated to be 0.008 $\Delta\text{Abs}_{460}/\Delta\text{GdnHCl(M)}$, 0.004 and 0.008, respectively. These data may indicate that GdnHCl may be acting on disengaging the dimeric subunits first followed by separation of domains and then disruption of subdomain to subdomain interaction. While this is rather speculative and a re-analysis of existing data, the current study examined this proposed mechanism of the initial stages of denaturation using GAPDH. The term denaturation in this regard may be inappropriate, as these events likely do not alter the secondary structure given that the concentration of the denaturant is low. We think that the subunit-subunit interface, as well as the domain-domain interface, may represent an entry point for regulatory molecules that can alter the biological properties of GAPDH and providing its diversity of function. We recently proposed that GAPDH, an asymmetric tetramer, may dissociate to a dimer or reconfigure to higher order structures that may include a decamer [2]. Interestingly, inhaled anesthetics appear to shift the equilibrium away from the tetramer towards a dimer that may proceed to a decamer [2]. We think that these interfacial events may involve modulation of association with its cofactor [3]. The present study revealed that GAPDH's oligomeric structure may be labile.

The folding of proteins are considered to be based on the principle of a domain being the functional unit of folding [13], and that contiguous folding regions may be separated by linker regions [14]. Subdomains without extensive linker regions may likewise involve a hierarchical process of folding [15]. The mechanisms of unfolding likely proceed with these same principles. GAPDH is a multimeric protein, where each subunit is composed of two distinct domains. We proposed that using GdnHCl at low concentrations would differentiate the hierarchical levels of unfolding. Similar to these effects, we think that low levels of endogenous chaotropic compounds may modulate quaternary and perhaps even tertiary structure affecting biological function. Just as inhaled anesthetics appear to act at interfacial areas, GdnHCl's primary interaction may involve disruption of interfacial contacts.

The first derivative spectra reveal a trough (nadir) and peak (zenith) that may provide insight to changes in the microenvironments of the aromatic residues, differentiating the contribution by tyrosine and tryptophan residues. Our lab will continue to explore this possibility. We intend to also explore the advantage of using second through fourth derivatives as was previously applied [16].

We observed that the wavelengths (that represent the

zeniths from the first derivative spectra of the downslope) decreased as a function of concentration of chaotropic agent. The resulting plot (Figure 3A) exhibited two transition points: one at 0.5M and the other at 1.0M GdnHCl, dividing the process into several phases. The first phase may involve separation of subunits from one another. This event would be initiated as GdnHCl was raised to 0.5M. The second phase from 0.5-1.0M GdnHCl may represent the separation of the domains from one another within the subunit structure. Above approximately 1.0M GdnHCl may involve secondary structure changes that are not as recognizable by these computational parameters. The shift of the zenith value appears to lessen and reach a limit value asymptotically. The nadir wavelengths plotted as a function of GdnHCl concentration also showed three phases, which was consistent with the zenith data.

5 Conclusion

The approach described in this study involved close inspection of the descending slope of the UV absorbance spectra, particularly examining the 283 to 310nm region, during exposure to low concentrations of GdnHCl. Molar concentrations of the chaotropic agent, such as 6M GdnHCl, completely unfolds most proteins. The effects of GdnHCl at concentrations below 1.0M remain poorly understood. Given the dynamic nature of GAPDH particularly that it exhibits multiple cellular functions with diverse binding partners, we proposed that subunit-subunit interfacial dynamics play a crucial role in GAPDH structure and function. Therefore, analysis of spectroscopic changes at low GdnHCl may reveal useful information regarding this dynamic feature. Conversion of UV spectra to first derivatives allows one to study the reliably appearing trough and peaks that exhibit quantal shifts. Our findings suggest that the GAPDH structure is easily perturbed and this intrinsic disorder, which likely resides at the interfacial regions, may contribute to functional diversity.

6 References

[1] Ikemoto A, Bole DG, Ueda T. Glycolysis and glutamate accumulation into synaptic vesicles. Role of glyceraldehyde phosphate dehydrogenase and 3-phosphoglycerate kinase. *J Biol Chem.* 2003;278(8):5929-40.
 [2] Pattin AE, Ochs S, Theisen CS, Fibuch EE, Seidler NW. Isoflurane's effect on interfacial dynamics in GAPDH influences methylglyoxal reactivity. *Arch Biochem Biophys* 2010;498(1):7-12.
 [3] Swearengin TA, Fibuch EE, Seidler NW. Sevoflurane modulates the activity of glyceraldehyde 3-phosphate dehydrogenase. *J Enzyme Inhib Med Chem* 2006;21(5):575-9.
 [4] Yeargans GS, Seidler NW. Carnosine promotes the heat denaturation of glycosylated protein. *Biochem Biophys Res Commun* 2003;300(1):75-80.

[5] Rogalski-Wilk AA, Cohen RS. Glyceraldehyde-3-phosphate dehydrogenase activity and F-actin associations in synaptosomes and postsynaptic densities of porcine cerebral cortex. *Cell Mol Neurobiol* 1997;17(1):51-70.
 [6] Wu K, Aoki C, Elste A, Rogalski-Wilk AA, Siekevitz P. The synthesis of ATP by glycolytic enzymes in the postsynaptic density and the effect of endogenously generated nitric oxide. *Proc Natl Acad Sci USA.* 1997;94(24):13273-8.
 [7] Chen YH, He RQ, Liu Y, Liu Y, Xue ZG. Effect of human neuronal tau on denaturation and reactivation of rabbit muscle D-glyceraldehyde-3-phosphate dehydrogenase. *Biochem J.* 2000;351(Pt 1):233-40.
 [8] Schmid F. Spectroscopic techniques to study protein folding and stability. In (J. Buchner and T. Kiefhaber, Eds.) *Protein Folding Handbook, Volume 1.* Wiley-VCH Verlag GmbH & Co. , Weiheim, Germany, 2005, pp. 22-44.
 [9] Shen ZM, Yang JT, Feng YM, Wu CS. Conformational stability of porcine serum transferrin. *Protein Sci* 1992; 1(11):1477-84.
 [10] Ruvinov SB, Thompson J, Sackett DL, Ginsburg A. Tetrameric N(5)-(L-1-carboxyethyl)-L-ornithine synthase: guanidine. HCl-induced unfolding and a low temperature requirement for refolding. *Arch Biochem Biophys* 1999;371(1):115-23.
 [11] Macedo MF, de Sousa M. Transferrin and the transferrin receptor: of magic bullets and other concerns. *Inflammation & Allergy Drug Targets* 2008;7(1):41-52.
 [12] Wally J, Buchanan SK. A structural comparison of human serum transferrin and human lactoferrin. *Biomaterials* 2007;20(3-4):249-62.
 [13] Baldwin RL. Early days of studying the mechanism of protein folding. In (J. Buchner and T. Kiefhaber, Eds.) *Protein Folding Handbook, Volume 1.* Wiley-VCH Verlag GmbH & Co. , Weiheim, Germany, 2005, pp. 2-21.
 [14] Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA* 1973;70(3):697-701.
 [15] Lesk AM, Rose GD. Folding units in globular proteins. *Proc Natl Acad Sci USA* 1981;78(7):4304-8.
 [16] Butler WL. Fourth derivative spectra. *Methods Enzymol* 1979;56:501-15.

Instantiation and adaptation of CRISP-DM to Bioinformatics computational processes

Santiago González, Víctor Robles, José M. Peña and Ernestina Menasalvas

Abstract—Among the many contributions made by information technologies to Bioinformatics, the techniques of intelligent data analysis combined with optimization techniques are the main application field nowadays. Many researches focused on DNA microarray field have proposed different approaches trying to obtain new undiscovered knowledge of diseases such cancer. All these researches can be represented as a standard unique process. Thus, this paper presents an overview of a common biological and computational process of DNA microarray data analysis that include these types of researches, based on the known CRISP-DM model.

I. INTRODUCTION

Nowadays, around 60 people die of diseases such as cancer every minute. The value is even more concerning if instead of thinking in minutes, we do it in hours or days. It is, therefore, a problem of high social impact that must be solved as quickly as possible. Finding a cure for diseases such as cancer would translate into a much higher life expectancy. In the scientific field, expert biologists are devoted to the study of possible solutions to these kinds of diseases. Among the many approaches, the DNA microarray technology will be the focus of this paper.

A DNA microarray is a large set of hybridized DNA molecules arranged on a solid (silicon or plastic) surface, called biochip. These types of experiments allow relative levels of mRNA abundance to be determined in a set of tissues or cell populations for thousand of genes simultaneously. A complete review of the methods used in the processing and analysis of gene expression for data generated by DNA microarrays experiments [11].

Many computer resources are needed in the work routine of an expert molecular biologist while studying DNA microarray data. That is why bioinformatics has been so important to meet the scientists' needs. The evolution of this new specialization was originally promoted by the biologists themselves and the needs they had at work. Nowadays, researchers from information technologies are beginning to work on this field, contributing on the data management and processing with their background of new tools and technology. We must bear in mind that we are talking of

rather complex information for non-biologists; therefore an intrinsic collaboration with the experts is absolutely essential.

Among the many contributions made by information technologies to bioinformatics, the techniques of intelligent data analysis combined with optimization techniques are the main application field nowadays. Many researchers contribute to improve, using these techniques, the results obtained with simple statistical studies. All researches that use Data Mining and Knowledge Discovery techniques to apply them on DNA microarray analysis are more or less supported on the same scheme or methodology. However, there is no any methodological process that describes all the possible Data Mining steps to analyze this kind of data.

Thus, this paper proposes an overview of a common biological and computational methodological process that includes practically all these types of researches. It is important to mention that the Computational Process is instanced and adapted of a known KDD model used by many Data Mining experts, which is called CRISP-DM.

The structure of the paper is as follows: The next section describes briefly the DNA Microarray technology. Section 3 presents the Biological process of DNA Microarray analysis, while section 4 describes the Computational process of these types of analysis and also analyzes each one of the steps of this process. Finally, all conclusions are presented in the last section (5).

II. DNA MICROARRAY TECHNOLOGY

DNA microarrays [17,27,31,11] are a relatively new and complex technology used in molecular biology and medicine. Microarrays present unique opportunities in analyzing gene expression and regulation in an overall cellular context. This technology has been applied in diverse areas ranging from genetic and drug discovery to disciplines such as virology, microbiology, immunology, endocrinology and neurobiology. Microarray technology is the most widely used technology for the large-scale analysis of gene expression because it provides a simultaneous study of thousands of genes by single experiment.

A DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides (shorts molecules consisting of several linked nucleotides, between 10 and 60, chained together and attached by

Santiago, Victor and Jose M. belong to the Department of Computer Architecture, Universidad Politécnica de Madrid in Spain. (emails: {sgonzalez,vrobles,jmpena}@fi.upm.es). Ernestina is from the DLSIS Department, also at the Universidad Politécnica de Madrid, in Spain. (mail: emenasalvas@fi.upm.es)

covalent bonds), called Expressed Sequence Tags (ESTs), each containing several molecules of a specific DNA sequence. This can be a short section of a gene or other DNA element.

III. BIOLOGICAL PROCESS OF DNA MICROARRAY ANALYSIS

There are several steps [28,21] in the design and implementation of a DNA microarray experiment (figure 1). Many strategies have been researched in each of these steps.

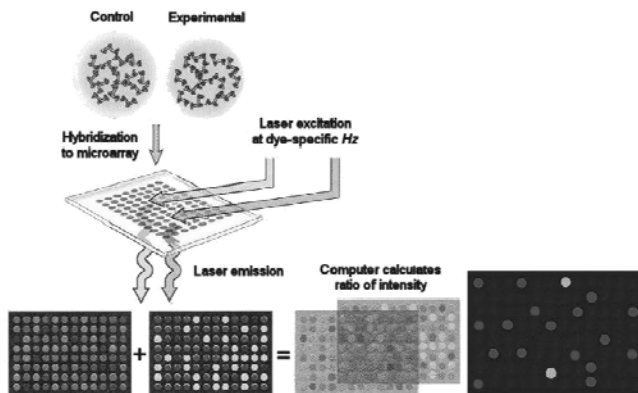


Fig. 1. Biological process of DNA Microarray analysis. Image from Gibson & Muse 2002

- **Probe:** First of all, the sample is obtained. The DNA type (cDNA/oligo with known identity) and the organism must be chosen in this step.
- **Chip manufacture:** The probes are placed on a surface. In standard microarrays, the information is attached to a solid surface by a covalent bond. The solid surface can be glass or silicon, in which case they are commonly known as gene chip or biochip. Here, several techniques have been used: Photolithography, pipette, drop-touch, piezoelectric (ink-jet), etc.
- **Sample preparation:** In this step the samples have been prepared. cDNA transcripts are prepared and labelled with a red fluorescent dye. A control library is constructed from an untreated source and labelled with a different fluorescent green dye.
- **Assay:** All information is hybridized (figure 2). Hybridization [28] is the process of combining single-stranded nucleic acids into a single molecule to the microarray.
- **Redaout:** Dual-channel laser excitation excites the corresponding dye, whose fluorescence is proportional to the degree of hybridization that has occurred. Relative gene expression is measured as the ratio of the two fluorescences: up-regulation of the experimental transcriptome relative to the control will

be visualized as a red pseudo-color, down-regulation show as green, and constitutive expression as a neutral black. The intensity of color is proportional to the expression differential.

- **Informatics:** In this final step, where new information and values are obtained from the fluorescence intensities using different computer techniques such as Robotics control, image processing [1], DBMS, etc. This step does not include data mining techniques, which have been studied as a computational process in next section.

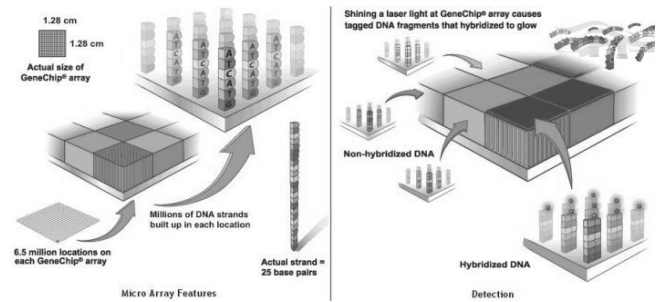


Fig. 2. Hybridization process. Image from <http://universe-review.ca/>

Nowadays, there are companies that create tools for analyzing complex genetic information such as DNA microarrays. Companies such as Affymetrix [7], Celera, Gene Logic, Xenometrix ... have built commercial platforms to carry out microarray experiments. Each platform obtains results using different methods (as Fluorescence, Mass spectrometry, Radioisotope, etc.) at each step of the microarray experiment. The use of platform determines the type of experimental design possible, the type of normalization, etc.

IV. COMPUTATIONAL PROCESS OF DNA MICROARRAY ANALYSIS

Once Biological process is finished, the Computational process starts. Trying to obtain any standard methodological process that englobes all published researches, we propose to instance the CRISP-DM model [41]. This model distinguishes six main phases of a KDD process:

- **Business Understanding:** This initial phase focuses on understanding the objectives and requirements from a business perspective, then converting this knowledge into a problem definition and a preliminary plan designed to achieve the objectives.
- **Data Understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to become familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets

to form hypotheses for hidden information.

- **Data Preparation:** The data preparation phase covers all activities for constructing the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed many times and not in any prescribed order. Tasks include record and feature selection as well as transformation and cleaning of data for modelling tools.
- **Modelling:** In this phase, various modelling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same problem type. Some techniques have specific requirements for the form of data. Therefore, stepping back to the data preparation phase is often necessary.
- **Evaluation:** Before proceeding to final deployment of the model, it is important thoroughly to evaluate the model and review the steps executed to construct the model in order to be certain it properly achieves the objectives. A key objective is to determine if there is some important issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the results should be reached.
- **Deployment:** Creation of the model is generally not the end of the project. Even if the purpose of the model is simply to increase knowledge of the data, the knowledge gained will need to be presented in a way that can be used.

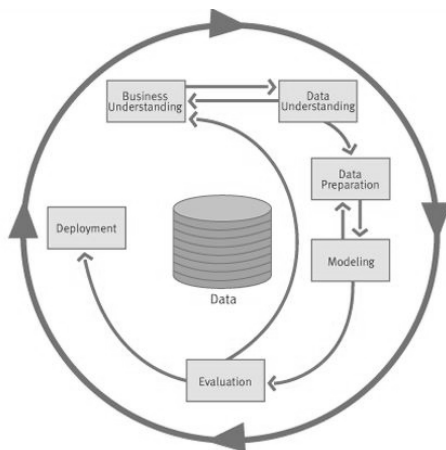


Fig.3. The phases of the CRISP-DM process model

Adapting this model to the microarray analysis process, the computational process of Microarray analysis is obtained. The Figure 4 shows a standard computational process with each phase. The life cycle of a computational process of DNA microarray analysis study consists of five phases. The sequence of the phases is not strict, moving back and forward between different phases is almost required, passing always on the Interpretation phase. It is because all decisions, results and objectives of each phase

have to be assessed and approved by expert biologists (biological interpretation). It is possible that an objective obtained in any phase has not a possible interpretation or is not a correct objective for the biologists. This can be produced, for instance, due to it is needed another Understanding iteration to understand the real objective. The lessons learned during the any phase can trigger new, often more focused questions to be answer by biologists.

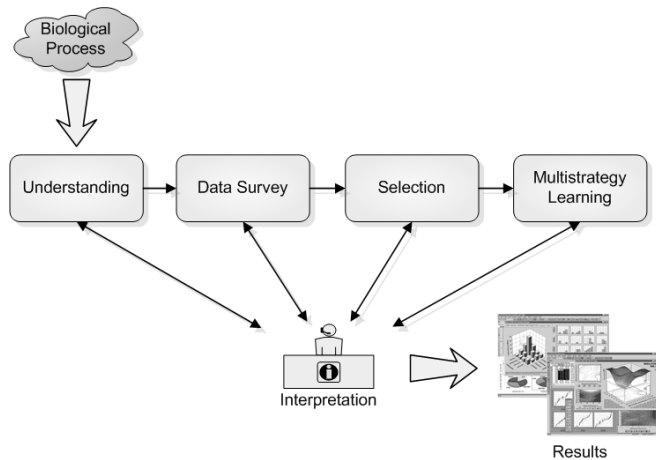


Fig. 4: Computational process of DNA Microarray analysis

A large number of data mining experiments with DNA microarray data can be represented with this methodological process, using all or not all phases, depending on the specific problem to be solved. In the next subsections each phase of the computational process is described and analyzed briefly.

A. Understanding

This initial phase focuses on understanding the research objectives and requirements from the expert biologists, and then converting this knowledge into a data mining problem definition. The biologists define and comment one specific problem. They provide a microarray expression dataset with descriptions, headers, gene identifications, patient information and possible classification of diseases, labels or outcomes.

B. Data Survey

In this phase all data is studied and prepared. Figure 3 defines the Data Survey tasks, Normalization and Pre-processing algorithms, both necessary to be able to access and compare correctly the data.

- **Normalization:** After the hybridizing and microarray image processing to obtain Cy5 and Cy3 fluorescence intensities (explained in section 3), it is needed to normalize [26,37,40] the data from each of the two scanned channels. There can be differences in labelling and detection efficiencies for the fluorescent labels and

differences in the quantity of the initial values from the two samples examined in the assay. These problems can cause a shift in the average ratio of the fluorescence intensities, so they must be re-scaled before an experiment can be properly analyzed. The normalization factor is used to adjust the data to compensate for experimental variability and to balance the fluorescence signals from the two samples.

There are many approaches for normalizing the gene expression. Some, such as total intensity normalization, are based on the assumption that the quantity of the initial RNA is the same for both labelled samples, so that consequently the total integrated intensity computed for all the elements in the array should be the same in both channels. Under this assumption, a normalization factor can be calculated and used to re-scale the intensity for each gene in the array. In addition to total intensity normalization, there are a number of alternative approaches for normalizing expressions, including linear regression analysis, log centering, rank invariant methods and Chen's ratio statistics (normalization using ratiostatistics), among others [26]. However, none of these approaches takes into account systematic biases that may appear in the data: dependence between intensity and ratio expression. Locally weighted linear regression (LOWESS) analysis [27], the most commonly used normalization method in DNA microarray experiments, can remove this dependency.

b) **Preprocess:** Obviously real data have a lot of redundancy, as well as incorrect or missing values, depending on some factors. Thus, usually it is needed some preprocessing algorithms in order to clean up and prepare the data. The most commonly used algorithms [17,8,38] are:

- Replicate handling or genes (features) that are replicated can be discarded.
- Missing value handling or patients (rows) that had more than 80% of missing gene values can be discarded.
- Imputing missing values can be estimated using different algorithms. The most known is the k-weighted nearest neighbor impute algorithm.

C. Selection

In this phase a selection of the principal features is made in order to improve the understanding of the problem and its possible solution. In figure 3, Selection phase is divided in two possible tasks to obtain these features.

a) **Dimension Reduction:** Here a feature reduction task can be applied to the data. This task is used to discard features that are not relevant for the study or can produce noise. Among all the dimension reduction algorithms [4], the most broadly used ones in

microarray data [6] are based on Principal Components Analysis (PCA), Partial Least Squares (PLS) or even discarding variables with low internal variance or with low Pearson correlation with outcome. Other approach presents an algorithm based on Penalized Logistic Regression to make a dimension reduction [32].

b) **Feature Selection:** Trying to compare Feature Subset Selection (FSS) and Dimension Reduction, the first one selects only the best features from the data and the second one discards those features that are not relevant for the study.

FSS can be used as a simple task to obtain the best features to later obtain the best new knowledge using these features. For that, simple FSS algorithms based on statistics are proposed and compared in DNA microarray field [14], such as Fold Change, ANOVA, Rank Products, etc. However, FSS is a so important task in DNA microarray data that sometimes is the final objective of many researches, that is to obtain the Biomarkers. FSS in Bioinformatics are reviewed by Saeys [30]. These techniques use wrapper and filter mechanisms with supervised and non-supervised algorithms. Thus, although the final objective is Feature Selection, it is needed to execute a Multistrategy Learning phase which includes supervised and non-supervised learning.

D. Multistrategy Learning

This phase is divided in two possible tasks (figure 3), unsupervised and supervised learning. Both tasks are used to obtain new knowledge (using different data mining algorithms) or the final objective of the process.

a) **Unsupervised Learning:** In DNA microarray technology, genes (features) classification is one of the typical final objective, although patients (rows) can be classified too. This classification can be obtained using different methods [18,25,34], such as Hierarchical cluster, EM, K-Means, QT, etc. Furthermore, it is interesting to obtain a patient classification (using the same methods) to later use this new information on the next task (supervised learning), to enrich knowledge and improve possible learners.

In Unsupervised Learning task, several studies obtain a Feature Subset Selection using wrapper mechanisms and unsupervised classification algorithm, such as EM or K-Means, to identify relationship between gene expressions [9,16]. Other researches [33] have proposed the use of biclustering technique (genes and patients simultaneously) to obtain better knowledge.

b) **Supervised Learning:** Usually in this task it is used a simple supervised learning using any supervised

classification method. Larrañaga [18] mentions the most used supervised classification methods in Bioinformatics, such as SVM, KNN, NaiveBayes, etc. However, this task can be something more complicated as a simple supervised classification algorithm. The same as Unsupervised Learning task, in Supervised Learning it is possible to make a FSS using wrapper [5] and filter mechanisms and different supervised classification algorithm, such as logistic regression [39,36], KNN, C4.5, NaiveBayes [12]. Furthermore, several researches use evolutionary algorithms, such as genetic algorithms [24], EDAs [29], or hybrid evolutionary algorithms [19] with supervised classification methods.

E. Validation

Both Unsupervised and Supervised Learning tasks have to be externally validated. This validation is based on three aspects:

- **data mining** external validation, using a validation technique (depending on the classifier) and a external dataset. For supervised classification, cross-validation and bootstrap [3] have been the most commonly used validation methods, but [13] comments that these methods are unreliable in small sample classification. For unsupervised classification,
- **using literature** and comparing obtained results with other results.
- **using biological experiments** validations for validate our work or using gene databases, such as GO or KEGG.

F. Interpretation

All interpretations, decisions, processes, feature selections and relations between genes or patients must be assessed and approved by expert biologists in order to obtain a valid result and/or objective in the microarray analysis.

V. CONCLUSIONS

A complete standard biological and computational process of DNA microarray analysis is proposed. Mention that image processing techniques have been studied out of the computational process. Approaches, such as [35,15,20,23], etc., use each one of these phases, creating an overall computational process as the proposed. Other approaches fit with the proposed process in several phases. Applications, such as Bioconductor [10], GAMS [22], Knime [2], Weka, etc., allow us to create and execute each one of the steps proposed in this paper.

ACKNOWLEDGEMENTS

The authors are grateful to the Blue Brain Project Team and the Canada IT-NRC Team, especially Fazel Famili, for their technical assistance. They also thankfully acknowledge the computer resources, technical expertise and assistance provided by the Centro de Supercomputación y Visualización de Madrid (CeSViMa) and the Spanish Supercomputing Network.

REFERENCES

- [1] P. Bajcsy. An overview of dna microarray image requirements for automated processing. In CVPR '05, page 147, Washington, DC, USA, 2005. IEEE Computer Society.
- [2] M. R. Berthold, N. Cebon, F. Dill, T. R. Gabriel, T. Kötter, T. Meinel, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. Knime: The konstanz information miner. In GfKL 2007. Springer, 2007.
- [3] U. M. Braga-Neto and E. R. Dougherty. Is cross-validation valid for small-sample microarray classification?
- [4] Miguel Carreira. A review of dimension reduction techniques. Technical Report CS-96-09, Dept. of Computer Science, University of Sheffield, January 1997.
- [5] K. Chrysostomou, S. Y. Chen, and X. Liu. Combining multiple classifiers for wrapper feature selection. IJDM, 1(1):91-102, 2008.
- [6] J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. Statistical applications in genetics and molecular biology, 5, 2006.
- [7] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada. The affymetrix genechip platform: an overview. Methods in enzymology, 410:3-28, 2006.
- [8] S. Durinck. Pre-processing of microarray data and analysis of differential expression. Methods in molecular biology, 452:89-110, 2008.
- [9] Jennifer G. Dy and Carla E. Brodley. Feature selection for unsupervised learning. J. Mach. Learn. Res., 5:845-889, 2004.
- [10] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol, 5(10), 2004.
- [11] Wolfgang Huber, Anja Von Heydebreck, and Martin Vingron. Analysis of microarray gene expression data. In Handbook of Statistical Genetics, 2nd edn. Wiley, 2003.
- [12] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. Artif Intell Med, 31(2):91-103, June 2004.
- [13] A. Isaksson, M. Wallman, H. Göransson, and M. G. Gustafsson. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters, 29(14):1960-1965, October 2008.
- [14] Ian B. Jeffery, Desmond G. Higgins, and Aedin C. Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics, 7:359+, July 2006.
- [15] K. Kaufman and R. Michalski. Discovery planning: Multistrategy learning in data mining, 1998.
- [16] Yongseog Kim and W. Nick Street. Evolutionary model selection in unsupervised learning. Intelligent Data Analysis, 6, 2002.
- [17] S. Knudsen. A biologist's guide to Analysis of DNA microarray data. JohnWiley and Sons, 2002.
- [18] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles. Machine learning in bioinformatics. Brief Bioinform, 7(1):86-112, March 2006.

- [19] A. LaTorre, J.M. Peña, S. González, O. Cubo, and F. Famili. Breast cancer biomarker selection using multiple offspring sampling. *Current Trends and Future Directions in ECML/PKDD 07*, 2007.
- [20] Seok Won Lee, Scott Fischthal, and Janusz Wnek. A multistrategy learning approach to flexible knowledge organization and discovery. In *Proceedings of AAAI-97*, pages 15–24. AAAI Press, 1997.
- [21] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405(6788):827–836, June 2000.
- [22] Dugas M., Weninger F., Merk S., Kohlmann A., and Haferlach T. A generic concept for large-scale microarray analysis dedicated to medical diagnostics. *Methods of information in medicine*, 45:146–152, 2006.
- [23] A. Naderi, A. E. Teschendorff, N. L. Barbosa-Morais, S. E. Pinder, A. R. Green, D. G. Powe, J. F. R. Robertson, S. Aparicio, I. O. Ellis, J. D. Brenton, and C. Caldas. A gene expression signature to predict survival in breast cancer across independent data sets. *Oncogene*,
- [24] C. H. Ooi and Patrick Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1):37–44, 2003.
- [25] Tuan D. Pham, Christine Wells, and Denis I. Crane. Analysis of microarray gene expression data. *Current Bioinformatics*, 1:37–53, 2006.
- [26] J. Quackenbush. Microarray data normalization and transformation - nature genetics.
- [27] J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 6(2):418–427, June 2001.
- [28] J. Quackenbush. Computational approaches to analysis of dna microarray data. *Methods Inf Med*, 45 Suppl 1:91–103, 2006.
- [29] V. Robles, C. Bielza, P. Larrañaga, S. González, and L. Ohno-Machado. Optimizing logistic regression coefficients for discrimination and calibration using estimation of distribution algorithms. *TOP*, 16(2):345–366, December 2008.
- [30] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, August 2007.
- [31] M. Schena, R. A. Heller, T.P. Theriault, K. Konrad, E. Lachenmeier, and R.W. Davis. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*, 7(16):301–306, July 1998.
- [32] Li Shen and Eng C. Tan. Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(2):166–175, April 2005.
- [33] Q. Sheng, Y. Moreau, and B. De Moor. Biclustering microarray data by gibbs sampling. *Bioinformatics*, 19:196–205, 2003.
- [34] Q. Sheng, Y. Moreau, F. De Smet, K. Marchal, and B. De Moor. Advances in cluster analysis of microarray data.
- [35] G. Sherlock. Analysis of large-scale gene expression data. *Brief Bioinform*, 2(4):350–362, December 2001.
- [36] S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.
- [37] G. K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–273, December 2003.
- [38] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, June 2001.
- [39] G. Weber, S. A. Vinterbo, and L. Ohno-Machado. Multivariate selection of genetic markers in diagnostic classification. *Artificial Intelligence in Medicine*, 31(2):155–167, 2004.
- [40] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4), February 2002.
- [41] O. Marban, J. Segovia, E. Menasalvas, C. Fernandezbaizan, *Toward Information Systems*, Vol. 34, No. 1. (March 2009), pp. 87–107. doi:10.1016/j.is.2008.04.003.

Analysis of Metabolic Networks: On the Similarity of the three Domains of Life

Karl G Kugler and Laurin AJ Mueller and Matthias Dehmer*
karl.kugler@umit.at and laurin.mueller@umit.at and matthias.dehmer@umit.at

Institute for Bioinformatics and Translational Research
University for Health Sciences, Medical Informatics and Technology (UMIT)
Eduard Wallnöfer-Zentrum 1
6060 Hall in Tyrol, Austria

Abstract—Metabolic networks summarize and represent anabolic and catabolic processes that are driven by the enzymes in every organism. It has been shown that the metabolic networks of the three domains of life (Archae, Bacteria, and Eukaryota) have certain properties in common. However, we could previously demonstrate that it is still possible to find domain-specific attributes in the corresponding networks, that allow for a good inter-domain classification performance. In this paper we aim at finding domain dependent differences based on distances between vertices in the networks. We apply three different distance-based topological network descriptors using Shannon's Entropy. Our results show that a clear distinction between the three domains of life fails when using the employed network descriptors. This indicates that certain distance-related properties are common to all organisms in this study. We expect this to be a sign of the evolutionary optimization of the information spread within these networks.

Keywords: Network biology, metabolic pathways, topological network descriptors, machine learning

I. BACKGROUND

Catabolic and anabolic processes can be represented by metabolic networks, as they represent the interlinkage of metabolic processes that make up the human metabolism [Alberts et al., 2007]. By studying how these processes are organized in pathways it is possible to derive knowledge about the underlying functions. Jeong et al. systematically investigated the organization and structure of metabolic networks from 43 organisms that were representing the three domains of life [Jeong et al., 2000]. One of their main results was, that despite the evolutionary distance, properties related to the network diameter were found to be highly conserved [Jeong et al., 2000]. However, in recent work we demonstrated that it is possible to still discriminate between the three domains of life [Mueller et al., 2011]. The main goal of this paper is to analyze path distance-based properties in the networks by Jeong et al., in order to detect domain-specific effects. We aim at detecting distance-based effects, that may hint at evolutionary differences between the domains of life. To tackle this problem we utilize entropy-based topological network descriptors [Dehmer and Mowshowitz, 2011]. The structure of a network also reflects its function [Strogatz, 2001]. Thus, applying

topological network descriptors might be useful for the analysis of complex networks, as they allow transforming structural information about a graph into a numeric value [Emmert-Streib and Dehmer, 2011]. Topological network descriptors have been employed in chemoinformatics, e.g. for predicting toxicity [Feng et al., 2003] or mutagenicity [Votano et al., 2004]. Recently, they have also been proven useful for analyzing biological networks [Mueller et al., 2010], [Emmert-Streib and Dehmer, 2011].

Jeong et al. first described the relatedness of metabolic networks, when they explored degree distributions and average path lengths [Jeong et al., 2000]. Later, Wagner and Fell found the metabolic network of *E. coli* to exhibit the small-world property for a slightly different set-up [Wagner and Fell, 2001]. Ma and Zeng analyzed the core networks and clustering properties of different organisms in their work [Ma and Zeng, 2003]. They also investigated the average path lengths of the largest subnetwork and the whole network for 65 organisms [Ma and Zeng, 2003]. A set of topological network descriptors was employed by Zhu and Qin to find differences in various single cell organisms [Zhu and Qin, 2005]. They found the average clustering coefficient and the average betweenness to differ between six Bacteria and four Archaea [Zhu and Qin, 2005]. In the current paper we focus on topological network descriptors that are based on inferring distances between vertices. We hypothesize that the analysis of these distances might reveal knowledge about the spread of information within these networks.

This paper is structured as follows: After providing background information in this Section, we describe the employed data set and the methods in Section II. Thereafter, we illustrate the results in Section III, which are discussed in IV. This paper finishes with a final summary and conclusion in section V.

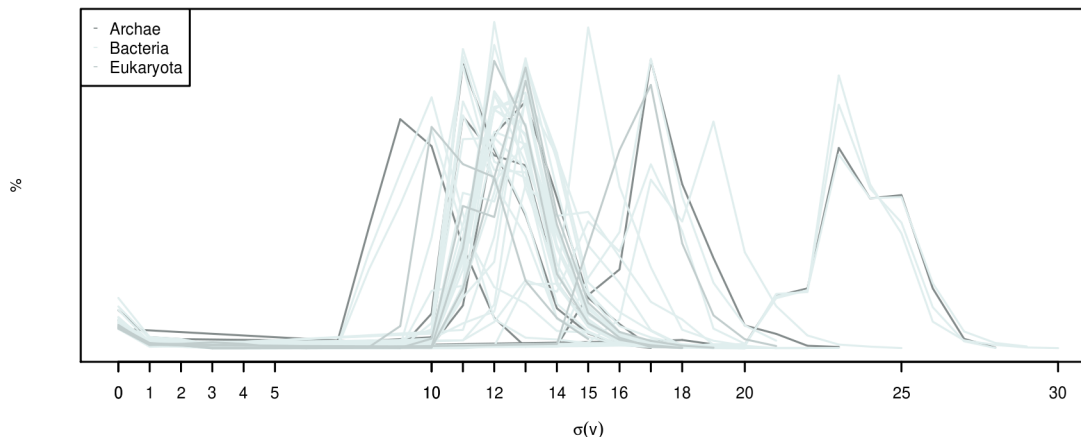


Fig. 1. The distribution of $\sigma(v)$ for the 43 organisms. Each of the three domains is depicted in a different color.

II. MATERIAL AND METHODS

Metabolic Networks

For the analysis of domain-specific effects we re-analyze the metabolic networks that have originally been studied by Jeong et al. [Jeong et al., 2000]. In their study they analyzed 43 organisms from the three domains of life ($n_{Archaea} = 6$, $n_{Bacteria} = 32$, and $n_{Eukaryota} = 5$). After we construct the networks, we extract the largest connected component, which represents the largest connected subgraph, for each organism. This results in a network G for every organism, where V is the set of labeled vertices and E is the set of directed edges. Overall, we then have 43 labeled and directed networks for the further analysis.

The eccentricity $\sigma(v)$ of a vertex v is an important feature within a network [Hage and Harary, 1995]. It gives the maximum of the distances from one vertex to all other connected vertices. In biological networks, small distances may indicate short communication processes, which allow for an organism to rapidly react to disturbances. To illustrate the distribution of $\sigma(v)$ for each of the 43 networks we plot it in Fig. 1.

Network Descriptors using Distances

Topological network descriptors represent the complexity of a network by a numeric value [Emmert-Streib and Dehmer, 2011]. Early applications of network descriptors date back to the work of Wiener [Wiener, 1947]. He utilized the sum of the distance matrix for predicting paraffin boiling points. Other well-known indices are the Balaban J index [Balaban, 1982], the Zagreb group indices [Diudea et al., 2001] or the Randić connectivity index [Li and Gutman, 2006]. Later, methods for quantifying the information content of a network were established [Bonchev and Rouvray, 2005],

[Mowshowitz, 1968], [Rashewsky, 1955], [Trucco, 1956]. Note, that many of these descriptors are correlated. Bonchev and Trinajstić introduced an information index that captures molecular branching [Bonchev and Trinajstić, 1977]. Many other real-world applications are also based on problems of relational structures, e.g. transportation or communication networks [Kolaczyk, 2009]. Networks and topological descriptors have been extensively used in the social sciences [Wasserman and Faust, 1994], e.g. for identifying opinion-leaders or the spread of information in societies.

In the present work we put an emphasis on i) descriptors, that can be used to evaluate the information spread in a network. ii) Descriptors that calculate the information-content of a network. We select entropy-based network descriptors since they were shown in previous work to possess good classification performance when capturing domain-specific effects [Mueller et al., 2011]. For a comprehensive overview on entropy-based network descriptors see e.g. [Dehmer and Mowshowitz, 2011]. We focus on studying the information-spread as we are interested in finding structural differences that present themselves in the way information is spread within the metabolic networks. Our hypothesis is, that we might find structural differences that can be clearly linked to a domain-specific origin.

It has been shown that it is possible to quantify the information-content of a network by applying special functionals to the vertices of the network and using Shannon's Entropy [Dehmer, 2008]. Dehmer presented a vertex functional that is based on the j -spheres [Dehmer, 2008]:

$$f^V(v_i) := c_1 |S_1(v_i, G)| + c_2 |S_2(v_i, G)| + \dots + c_{\rho(G)} |S_{\rho(G)}(v_i, G)|, \quad (1)$$

$$c_k > 0, 1 \leq k \leq \rho(G).$$

$S_j(v_i, G)$ is the set of vertices with distance j from vertex $v_i \in V$. Note, that c_k represents a weighting factor. In our case, we modeled it to follow an exponential function. So $c_k = \rho(G)e^k$ for $k = 0, 1, \dots, \rho(G) - 1$. This allows emphasizing on vertices that are close to v_i . The structural information content of a graph G with respect to $f^V(v_i)$ is then defined by [Dehmer, 2008]:

$$I_{f^V}(G) = \sum_{i=1}^{|V|} \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)} \log_2 \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}. \quad (2)$$

$f^V(v_i)$ can be seen as a function that represents the spread of information from v_i , so $I_f(G)$ is a model for the information spread in G [Dehmer, 2008].

Bonchev et al. introduced a descriptor that is based on the eccentricity $\sigma(v_i)$ and the mean information content [Bonchev et al., 1980]. The radial centric information index is defined by [Bonchev et al., 1980]:

$$\bar{I}_C^V(G) = \sum_{j=1}^{|V|} \frac{n_j}{|V|} \log_2 \frac{n_j}{|V|}. \quad (3)$$

Here, n_j gives the number of vertices with eccentricity $\sigma(v_i) = j$. It is common to assume that small $\sigma(v_i), v_i \in V$ indicate the possibility to spread information rapidly within G . So, $\bar{I}_C^V(G)$ should give an insight into how information is spread in G . If our hypothesis holds, organisms from different domains may exhibit systematic differences with respect to $\bar{I}_C^V(G)$.

It is possible to define information measures using local features of graphs, e.g. by quantifying the entropy of single vertices [Dehmer and Mowshowitz, 2011]. Konstantinova and Paleev introduced a measure that represents the vertex complexity by [Konstantinova and Paleev, 1990]:

$$I_D(v_i) = - \sum_{j=1}^{|V|} \frac{d(v_i, v_j)}{d(v_i)} \log_2 \left(\frac{d(v_i, v_j)}{d(v_i)} \right), \quad (4)$$

where $d(v_i)$ gives the sum of distances from vertex v_i to all other vertices in G . The entropy of G is then given as [Konstantinova and Paleev, 1990]:

$$I_D(G) = \sum_{i=1}^{|V|} I_D(v_i). \quad (5)$$

Here, we use $I_D(G)$ to model the heterogeneity of the vertices of a graph G with respect to the distances between the vertices. Based on our hypothesis we should see a domain-specific effect in this heterogeneity.

Univariate Analysis

After we calculate the three topological network descriptors for each of the 43 metabolic networks we proceed with the succeeding data analysis. First, we test for the presence of a domain-specific effect in at least one group by performing a

one way ANOVA [Chambers and Hastie, 1991].

Unsupervised Machine Learning

We use hierarchical clustering in order to explore the groups that are formed by the employed distance measures. Our clustering is based on the Euclidean distance between features [Murtagh, 1985].

Supervised Machine Learning

For supervised machine learning we make use of support vector machines [Vapnik and Lerner, 1963], with a radial basis kernel. To optimize the outcome we set the cost parameter to 100. We then calculate the accuracy and the f-score for the classification of the domains of life based on our set of topological network descriptors.

III. RESULTS

Distance-Based Network Descriptors

For each of the 43 species we calculate the three presented descriptors with the programming language R (<http://www.r-project.org>). The results are listed in Table I and illustrated as boxplots in Fig. 2.

Univariate Analysis

The results for the ANOVA are listed in Table II. To adjust for multiple testing we correct with the method by Bonferroni. However, even before the multiple testing correction no descriptor detects a significant effect in a single domain.

Unsupervised Machine Learning

The heatmap in Fig. 3 illustrates the results of the hierarchical clustering. The rows contain the 43 organisms and the columns represent the three employed topological network descriptors. We mark each domain in a specific color. We observe no meaningful clustering with respect to the three domains of life.

Supervised Machine Learning

The classification accuracy is 63%. While we reach a precision of 0.86 we only score a recall of 0.33. This leads to an overall f-score of 0.48.

IV. DISCUSSION

In the present study our goal was to detect differences and characteristics for the three domains of life by making use of their metabolic networks. Therefore, we reused a set of 43 organisms that have originally been investigated by Jeong et al. [Jeong et al., 2000]. Here, we focused on analyzing potential differences in the distances between vertices in the metabolic networks. We employed a broad range of different approaches to this problem, which all failed to detect any domain-specific effects.

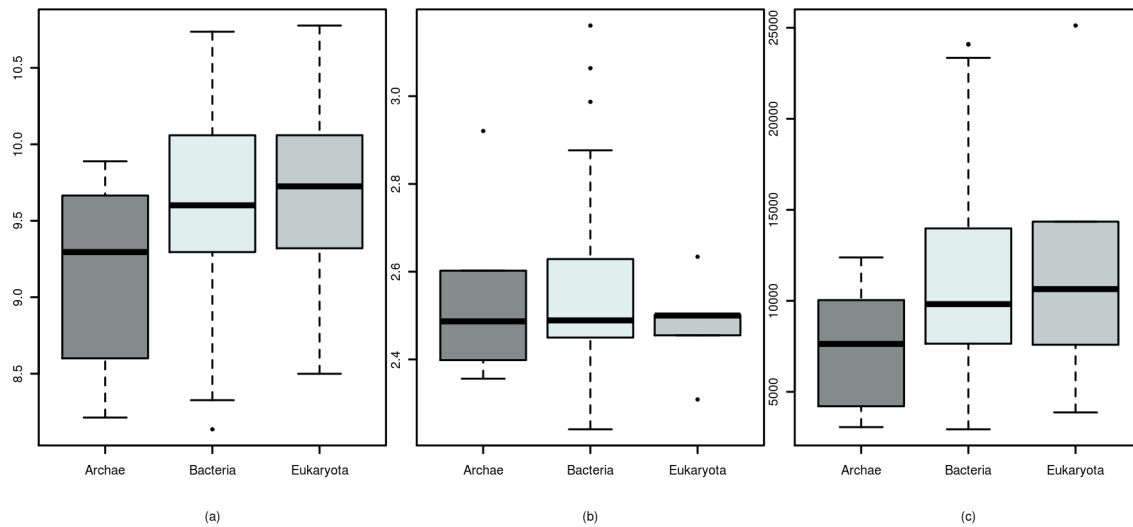


Fig. 2. For each of the 43 species we calculate three entropy-based network descriptors: (a) I_{fV} , (b) I_C^V , (c) I_D .

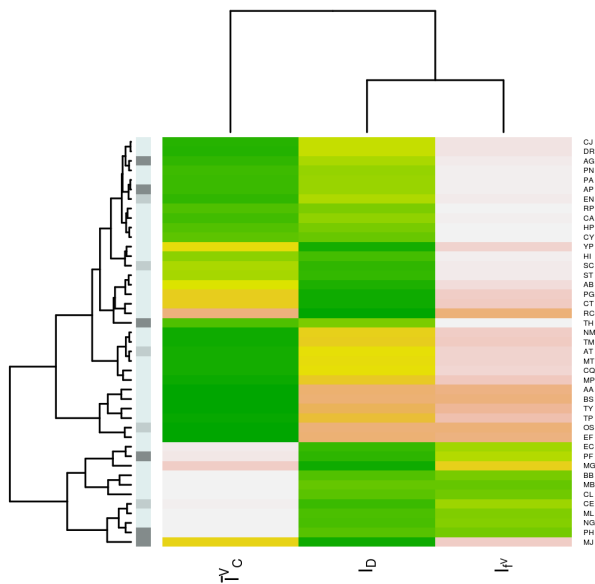


Fig. 3. We perform a hierarchical clustering for the 43 metabolic networks (rows) and the three employed network descriptors (columns). The three domains of life are depicted in three different colors.

In the original work by Jeong et al. they discovered several interesting aspects that were common to all networks. All the degree distributions of the networks were found to be scale-free and follow a power-law distribution [Jeong et al., 2000]. Moreover, the network diameters $\rho(G)$ were found to be relatively constant across all three domains of life [Jeong et al., 2000]. The similarity in the large-scale organization of the metabolic networks is also discussed in [Podani et al., 2001]. These observations indicate that

core properties of the metabolic processes are common to all species and are to a certain degree not influenced by evolutionary processes. However, in recent work we could demonstrate that it is still possible to distinguish between Archaea, Bacteria, and Eukaryota based on topological properties of their metabolic networks [Mueller et al., 2011]. In that previous study we applied a set of supervised machine learning algorithms to 33 network descriptors that were calculated for the same data, and came up with a reasonable classification performance (Accuracy: 88.4%, weighted F-score: 0.88). Such a result has not been reached in the present study. However, in contrast to this previous work we now considered directed graphs for our analysis. This hardens a direct comparison of the previous results with the current ones. Interestingly, when we ignored the directional information, two measures that are related to path length and the spheres turned out to be significantly different in at least one group [Mueller et al., 2011]. This is a striking observation that will need to be verified and interpreted in future studies.

Considering that Jeong et al. observed highly conserved distance properties in their original study and that we focused our analysis on these network invariants the observed results come to no surprise. We conclude that this highlights the fact that metabolic networks are likely to have evolved in a way that allows spreading information efficiently, and that this design is common to most organisms in the present set of networks. Our results are to a certain degree coherent with other, related observations. In a similar study, clear differences between Bacteria and Archaea were found for the average clustering coefficient and the average betweenness, but not so much for the average path length and the diameter [Zhu and Qin, 2005]. These latter two are mainly related to distance between vertices, which was also the graph invariant

TABLE I

HERE, WE LIST THE RESULTS FOR THE 43 ORGANISMS AND THE THREE EMPLOYED TOPOLOGICAL NETWORK DESCRIPTORS.

Organism	I_f	I_C^V	I_D	Domain
AP	9.665084	1.704296	10040.587666	Archaea
AG	9.564147	1.662454	9225.599594	Archae
TH	9.888839	1.803817	12382.120096	Archae
MJ	8.600314	1.63313	4206.884867	Archae
PF	9.025969	2.02442	6037.409112	Archae
PH	8.2133	1.74309	3060.56127	Archae
AA	10.658253	1.762958	23348.495827	Bacteria
CQ	9.938537	1.675467	12776.953312	Bacteria
CT	9.787773	1.887457	11276.072509	Bacteria
CY	9.505232	1.715802	8762.273987	Bacteria
PG	9.435009	1.804038	8489.037187	Bacteria
MB	8.136314	2.191041	2940.192524	Bacteria
ML	8.362968	1.837814	3543.299666	Bacteria
MT	10.070713	1.717556	14174.900512	Bacteria
BS	10.700696	1.678981	24103.108553	Bacteria
EF	10.736119	1.553148	24075.776966	Bacteria
CA	9.586382	1.701106	9517.179245	Bacteria
MG	9.416086	1.974329	8301.183302	Bacteria
MP	10.047548	1.615207	13779.564647	Bacteria
PN	9.523945	1.673842	8863.919547	Bacteria
ST	9.580282	1.792999	9745.236084	Bacteria
CL	8.350742	1.993873	3805.079281	Bacteria
RC	9.669622	1.924546	10291.659579	Bacteria
RP	9.615696	1.730073	9890.367254	Bacteria
NG	8.326707	1.783241	3343.943011	Bacteria
NM	10.142384	1.720369	15237.759713	Bacteria
CJ	9.657014	1.650279	9902.713247	Bacteria
HP	9.556632	1.711279	9098.264346	Bacteria
EC	8.964342	2.070491	5795.615593	Bacteria
TY	10.491932	1.71284	20399.468024	Bacteria
YP	9.154839	1.735033	6777.224163	Bacteria
AB	9.524078	1.806462	9364.071385	Bacteria
HI	9.174159	1.695162	6988.915709	Bacteria
PA	9.678677	1.709062	10192.080149	Bacteria
TP	10.341138	1.751982	18038.2337	Bacteria
BB	8.446315	2.123614	3699.92756	Bacteria
TM	10.171566	1.705844	15249.575415	Bacteria
DR	9.640121	1.647044	9901.92116	Bacteria
EN	9.724915	1.701616	10642.447537	Eukaryota
SC	9.320128	1.732907	7586.420683	Eukaryota
CE	8.499299	1.825795	3867.729448	Eukaryota
OS	10.776499	1.60026	25122.661425	Eukaryota
AT	10.059327	1.733995	14349.913609	Eukaryota

TABLE II

THE RESULTS FOR THE ANOVA TESTING. p_{Bonf} IS THE P-VALUES AFTER THE BONFERRONI CORRECTION.

	I_f	I_C^V	$I_D(v_i)$
p	0.384	0.638	0.342
p_{Bonf}	1.000	1.000	1.000

of interest in our study. All three utilized topological network descriptors were quantifying the information-content of the networks. In recent work we were able to demonstrate that this family of descriptors is powerful for detecting differences related to the three domains of life when using this set of data [Mueller et al., 2011]. In the present work, the low power in finding domain-specific differences is caused by the underlying graph invariant. We hypothesize that in order to find domain-specific differences in the topology of the

networks in this set it is better to focus on other graph invariants, e.g. vertex degrees or centralities.

V. SUMMARY AND CONCLUSION

Finding specific properties in groups of biological networks is a major goal in network analysis. Here, we wanted to detect topological properties responsible for the spread of information within a network and specific for the three domains of life (Archaea, Bacteria, and Eukaryota). Therefore, we employed a set of three network descriptors that capture properties related to distances within a graph. We calculated each of these three descriptors on a set of 43 metabolic networks from different organisms. To analyze the according data we utilized univariate methods as well as supervised and unsupervised machine learning procedures. However, with none of the applied approaches could we detect any meaningful discrimination or characterization of the three domains of life. Since we could demonstrate in previous work that is possible to discriminate between the three domains of life based on the present data, we conclude that the information-spread as captured by the employed measured fails to capture domain-specific properties for this set of directed networks. It will be part of future work to analyze what groups of topological network descriptors are best fitted to solve this undertaking. This could then give insights into evolutionary differences between the domains.

VI. ACKNOWLEDGEMENT

This work was funded by the Tiroler Zukunftsstiftung. This work was supported by the COMET Center ONCOTYROL and funded by the Federal Ministry for Transport Innovation and Technology (BMVIT) and the Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth (BMWA/BMWFJ), the Tiroler Zukunftsstiftung (TZS) and the State of Styria represented by the Styrian Business Promotion Agency (SFG).

We thank Armin Graber and Frank Emmert-Streib for fruitful discussions. Frank Emmert-Streib also brought our attention to the analyzed data set.

REFERENCES

- [Alberts et al., 2007] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular Biology of the Cell*. Garland Science, 5 edition.
- [Balaban, 1982] Balaban, A. (1982). Highly discriminating distance-based topological index. *Chemical Physics Letters*, 89(5):399–404.
- [Bonchev et al., 1980] Bonchev, D., Balaban, A. T., and Mekenyan, O. (1980). Generalization of the Graph Center Concept, and Derived Topological Centric Indexes. *Journal of Chemical Information and Computer Sciences*, 20(2):106–113.
- [Bonchev and Rouvray, 2005] Bonchev, D. and Rouvray, D. H. (2005). *Complexity in Chemistry, Biology, and Ecology*. Springer, New York.
- [Bonchev and Trinajstić, 1977] Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix, and molecular branching. *Journal of Chemical Physics*, 67:4517–4533.

- [Chambers and Hastie, 1991] Chambers, M. and Hastie, T. (1991). *Statistical Models in S*. Chapman and Hall/CRC.
- [Dehmer, 2008] Dehmer, M. (2008). Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation*, 201(1-2):82–94.
- [Dehmer and Mowshowitz, 2011] Dehmer, M. and Mowshowitz, A. (2011). A history of graph entropy measures. *Inf. Sci.*, 181(1):57–78.
- [Diudea et al., 2001] Diudea, M., Gutman, I., and Jäntschi, L. (2001). *Molecular topology*. Nova Science Pub Inc.
- [Emmert-Streib and Dehmer, 2011] Emmert-Streib, F. and Dehmer, M. (2011). Networks for Systems Biology: Conceptual Connection of Data and Function. *IET Syst Biol.* in press.
- [Feng et al., 2003] Feng, J., Lurati, L., Ouyang, H., Robinson, T., Wang, Y., Yuan, S., and Young, S. S. (2003). Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. *J. Chem. Inf. Comput. Sci.*, 43(5):1463–1470.
- [Hage and Harary, 1995] Hage, P. and Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, 17(1):57 – 63.
- [Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.
- [Kolaczyk, 2009] Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated.
- [Konstantinova and Paleev, 1990] Konstantinova, E. V. and Paleev, A. A. (1990). Sensitivity of topological indices of polycyclic graphs. *Vychisl. Sistemy*, 136:38–48. (in Russian).
- [Li and Gutman, 2006] Li, X. and Gutman, I. (2006). Mathematical aspects of Randić-type molecular structure descriptors. *Math Chemistry Monographs*.
- [Ma and Zeng, 2003] Ma, H.-W. and Zeng, A.-P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430.
- [Mowshowitz, 1968] Mowshowitz, A. (1968). Entropy and the complexity of the graphs. i: An index of the relative complexity of a graph. *Bull. Math. Biophys.*, 30:175–204.
- [Mueller et al., 2011] Mueller, L., Kugler, K., Netzer, M., Graber, A., and Dehmer, M. (2011). Distinguishing between the Three Domains of Life using Topological Characteristics of their underlying Metabolic Networks. submitted.
- [Mueller et al., 2010] Mueller, L. A., Kugler, K. G., Dander, A., Graber, A., and Dehmer, M. (2010). Network-based Approach to Classify Disease Stages of Prostate Cancer Using Quantitative Network Measures. *Conference on Bioinformatics & Computational Biology (BIOCAMP'10), Las Vegas/USA*, 1:55–61.
- [Murtagh, 1985] Murtagh, F. (1985). Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag, 1985*, 1.
- [Podani et al., 2001] Podani, J., Oltvai, Z. N., Jeong, H., Tombor, B., Barabási, A. L., and Szathmáry, E. (2001). Comparable system-level organization of Archaea and Eukaryotes. *Nat Genet*, 29(1):54–56.
- [Rashewsky, 1955] Rashewsky, N. (1955). Life, Information Theory, and Topology. *Bull Math Biophys*, 17:229–235.
- [Strogatz, 2001] Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–276.
- [Trucco, 1956] Trucco, E. (1956). A note on the information content of graphs. *Bull. Math. Biol.*, 18(2):129–135.
- [Vapnik and Lerner, 1963] Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.
- [Votano et al., 2004] Votano, J. R., Parham, M., Hall, L. H., and Kier, L. B. (2004). New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. *Mol Divers*, 8(4):379–391.
- [Wagner and Fell, 2001] Wagner, A. and Fell, D. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803.
- [Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge Univ Pr.
- [Wiener, 1947] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Amer. Chem. Soc.*, 69:17–20.
- [Zhu and Qin, 2005] Zhu, D. and Qin, Z. S. (2005). Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, 6:8.

Construction and Analysis of Metabolite Network of *Arabidopsis thaliana* Pathways

Kasthuribai Viswanathan*, Nita Parekh*

*Center for Computational Natural Science and Bioinformatics,
IIT- Hyderabad - 500032, India.

Abstract – The recent large scale advances in science and technology has resulted in accumulation of large amount of biological pathways data. Any metabolic pathway contains large number of enzymes, metabolites and reactions. To make sense of diverse data available on a system, one needs to correlate and analyze them as a whole. Motivated by the potential benefits of graph theory and its applications in biological data, we discuss the automated reconstruction and analysis of metabolite network of *Arabidopsis thaliana* using concepts of graph theory. *A.thaliana* metabolite network was reconstructed and the analysis of the global properties of its metabolite-centric graph shows that the network is small-world and scale-free in nature. The investigation of nodes with high centrality values like high degree and high betweenness in this network help in identifying important metabolites, reactions, etc. Newman's modularity-based approach has been used in the analysis of the metabolite network of *A. thaliana* to identify pathway clusters, isolated pathways, and orphan metabolites or products. Our analysis on network representations helps in understanding the relationship between the metabolites, enzymes and reactions of metabolic pathways in *A. thaliana*.

Keywords: metabolic pathways; graph theory; metabolite network; modularity; centrality measures

1 Introduction

In recent decades, a large number of complete and draft genomes have been sequenced very rapidly. In spite of enormous metabolic reaction data, the accurate prediction of metabolite phenotypes remains difficult. Pathway reconstruction is an approach to corroborate the experimental data and to widen its utilities. Oldest and dynamic method of pathway reconstruction is the kinetic metabolic modeling [1]. It is based on rate laws of participating reactions and corresponding kinetic parameters. Despite the utilities, kinetic approach is not handy because, the determination and interpretation of concentrations and rate reactions are much difficult. On the other hand, pathway reconstruction using graph theory becomes advantageous since only very less information is required to construct the metabolite network of the entire pathways in the organism. In depth functional analysis of metabolic pathways is succeeded by decomposition of this network.

A complete graph can be constructed using the existing knowledge of metabolites, enzymes and reactions from the

metabolic pathway databases. The undirected metabolite network was constructed by considering each substrate as a node and an edge drawn between two substrates sharing the same reaction [2]. Even though the utility of pathway reconstruction is very high in plants, only a very few plant metabolic pathways have been reconstructed. We have chosen *Arabidopsis thaliana* for the study since it is a model organism which has significant metabolic pathways with remarkable functionalities like defense against pathogens and herbivores, UV protection, resistance against oxidative stress and Auxin transport.

Here, we have used an automated and efficient metabolite pathway reconstruction of *A.thaliana* using data set extracted from Kyoto Encyclopedia of Genes and Genomes (KEGG) Release 50.0, April 1, 2009 [3]. The XML file in the KEGG FTP contains reactions grouped under pathways of a specific organism. The XML file does not contain currency metabolites like ATP, H₂O etc. List of edges and arcs that capture the biological relationship was computed. This file was visualized using open source visualization tools, such as Pajek and Centbin that help in plotting distributions, navigation within the network and calculating centralities of the biological networks.

The degree of a node in a network is the number of connections or edges the node has with other nodes. The degree distribution of the *A.thaliana* metabolite network shows that a few nodes have high degree and most of the nodes have low degree revealing the scale free nature [4]. Construction of a random network with the same number of nodes and edges as the *A.thaliana* metabolite network exhibited similar path length but smaller clustering coefficient compared to the *A.thaliana* metabolite network suggesting its small world nature. Correlation between high degree and high betweenness of the network shows that there are many nodes with high betweenness and low degree. These nodes connect pathways or two groups of reactions and are important to be analyzed.

We analyzed the robustness of the network by random and targeted removal of nodes in both the metabolite network and its random counterpart. Targeted removal was performed on nodes exhibiting high centrality values (degree and betweenness). When under attack by nodes with high degrees, the random network does not show any difference whether the nodes are selected randomly or based on the decreasing values

of degree, whereas the metabolite network shows a drastic change in diameter when the nodes are targeted for removal. The community detection analysis of *A.thaliana* metabolite network suggests its modular nature. Modularity analysis (using Newman's algorithm) of the network showed hierarchical architecture and also helped in identifying isolated and orphan metabolites.

2 Materials and Methods

In KEGG metabolic pathway database, the pathway maps are validated, manually drawn and updated frequently and the enzymes are cross-referenced to other relevant databases like GenBank, PDB, etc. [9]. Hence for reconstruction of metabolic network of *A.thaliana*, we have used the Kyoto Encyclopedia of Genes and Genomes database (<http://www.genome.jp/kegg>).

2.1 Dataset

KEGG FTP contains metabolic pathways as XML files for each listed organisms. In KEGG one hundred metabolic pathways, listed for *A.thaliana*, have been downloaded as XML files for analysis (<ftp://ftp.genome.jp/pub/kegg/xml/kgml/metabolic/organisms/ath/>).

2.2 Substrate Centric Graph

The KEGG XML file has unique reaction id for each reaction in the pathway followed by the unique ids for the reactants and products. These files are incomplete without detailed information of secondary metabolites in reactants and products. Using perl scripts, the reaction id are matched with KEGG entire reaction list which has complete reaction information and the missing information are made complete. Since the network we constructed is undirected and does not contain currency metabolites, the information on the direction of the reactions and the currency metabolites are neglected. Reactants and products of the same reaction are connected by edges. Each reactant and the product becomes each node in the network, the reaction id is assigned to the edge connecting two nodes. Edge list is computed by listing the connected edges and their corresponding reaction ids. The edge list captures the network property and this file is used for the network analysis.

3 Results and Discussion

The metabolite network constructed has metabolites as nodes and the corresponding reactions they take part as edges. The metabolite network we generated for *A.thaliana* has 2801 unique metabolites.

The network does not contain the common small molecules or currency metabolites such as ATP, NADH, water etc. There are 3639 unique reactions in the network. The diameter of the network, which is the largest distance between two nodes is 56. To know how our network differs from similar networks, we compare the properties of our metabolite network with the

random network constructed with same number of nodes and edges and with the Radrich's *Arabidopsis* metabolite network model (Table1).

Table 1: Global properties of metabolite network, random network and the metabolic network by Radrich in *A.thaliana* network construction

	Metabolite Network	Random Network	Radrich Network
Nodes	2801	2801	2288
Edges	3639	3639	6547
Diameter	56	8	10
Clustering Coeff.	0.215	0.001	0.186
Avg. Path length	3.486	4.642	3.286

The random network has low clustering coefficient compared to the metabolite network constructed by us. Radrich's semi automated genome-scale reconstruction network on *Arabidopsis* by integration of metabolic databases [6] uses current metabolites and pathway data that were common in both KEGG and AraCyc. There are more edges in Radrich model due to the currency metabolites. The diameter of the metabolite network is very high compared to that of random network with same number of nodes and edges and Radrich network. The larger diameter in our network reveals that the information flow is between metabolites of two completely unrelated pathways leading to larger path lengths between those nodes. The lower clustering coefficient of a random network compared to *A.thaliana* metabolite network explains occurrence of meaningful clustering in biological network. In metabolite network, we see the average path length depends on the system size but does not change drastically with it.

3.1 *Arabidopsis thaliana* Metabolite Network is Scale free and Small world

The degree of a node in a network is the number of connections or edges the node has to other nodes. The degree distribution $P(k)$ gives the fraction of nodes that have degree k and is obtained by counting the number of nodes $N(k)$ that have $k = 1, 2, 3, \dots$ edges and dividing it by the total number of nodes N . From Fig.1a, degree distribution graph, we see that it follows the power law which appears as a straight line on a logarithmic plot (Fig 1.b) and hence proving metabolite network follows 'scale free nature'[2]. Using this function $P(k)$ it is evident that there is a high diversity in the degree of the nodes (Fig.1).

This nature becomes more evident by comparing it with a random graph with the same number of edges and arcs. We constructed a random graph, using the Erdoes Renyi model that assumes each pair of nodes in the network is connected randomly with probability p . This graph reflects the expected properties of a network which is random with respect to the node's position and their interaction compared to a metabolite network of the same size [3]. Random network have a bell-

shaped degree distribution (Fig.1c), indicating that the majority of nodes have a degree close to the average degree $\langle k \rangle$. The average clustering coefficient of a random graph equals $\langle k \rangle / N$ and thus is very small for large N [7]. We compare the degree distribution of the metabolite network with random network containing same number of nodes (Fig 1).

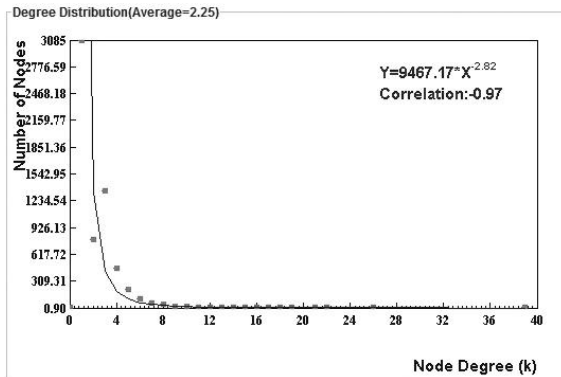


Fig. 1a: Metabolite graph in linear scale

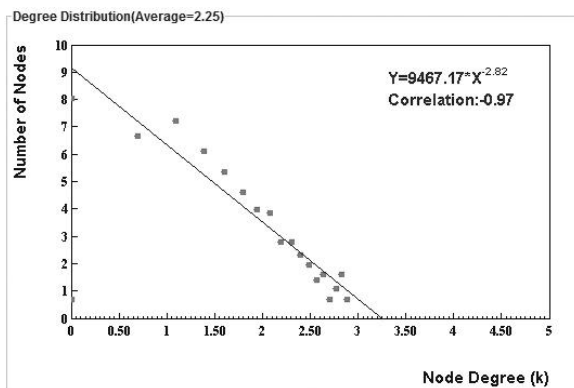


Fig. 1b: Metabolite graph in logarithmic scale

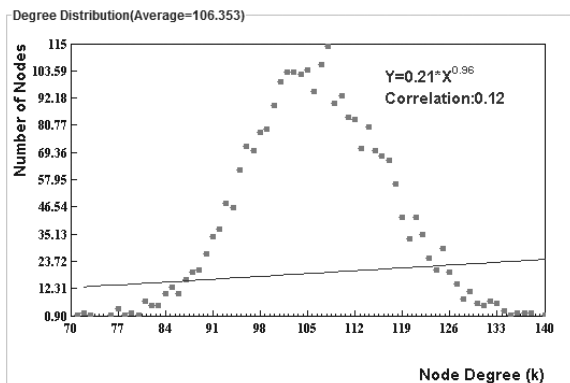


Fig. 1c: Random graph in linear scale

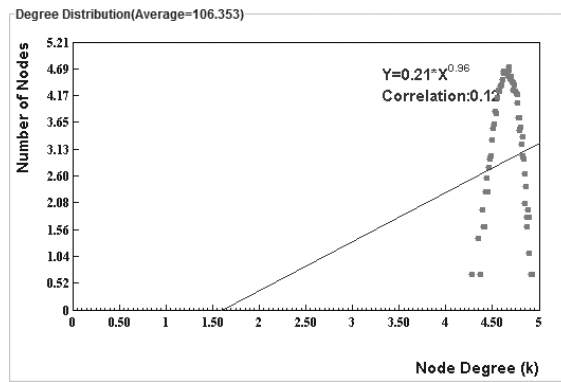


Fig. 1d: Random graph in logarithmic scale

Fig. 1: Comparison between the degree distribution of (a) Random graph having the same number of nodes and edges as the Arabidopsis metabolite network Arabidopsis thaliana metabolite network and for clarity the same two distributions are plotted both on a linear and logarithmic scale for all the networks. The bell-shaped degree distribution of random graphs peaks at the average degree and decreases fast for both smaller and larger degrees, indicating that these graphs are statistically homogeneous. By contrast, the degree distribution of the scale-free network follows the power law $P(k) = Ak^{-3}$, which appears as a straight line on a logarithmic plot.

We compare the metabolite network with random network. Another observation by comparing the metabolite network with and random network is that the average clustering coefficient of the random network is much smaller than that of *A.thaliana* metabolite network and the average path length was closer in the random graph, justifying the small world nature of the metabolite network [7].

3.2 Error and Attack Tolerance nature

The nodes in *A.thaliana* metabolite network are capable of staying interconnected and communicate even by unrealistically high failure rates. However, most networks become extremely vulnerable to attacks on selected nodes that bridge highly interconnected nodes in the network. We tested the error and attack tolerance nature of *A.thaliana* network comparing random and scale free networks.

Attack vulnerability shows a decreased performance of a network due to the selected removal of nodes or edges [8]. Here, it means the prevention of a metabolic reaction to take place due to the removal of an enzyme or primary substrate. Studying the attack vulnerability of networks is very important for identifying the weak or strong 'links' in the network [1]. Subsequently, this knowledge can be used to protect the network from outside attacks. In order to study the attack tolerance, we removed a fraction of nodes from both random and *A.thaliana* metabolite networks and studied the effect of this removal on the diameter and clustering coefficient.

We randomly removed 5, 10, 15, 20, 25 percentage of nodes from the *A.thaliana* metabolite network. In random network, due to the homogeneity all nodes contribute equally to the diameter, so the removal of each node caused the same effect

(Table 2a, 2b). But in case of metabolite network (scale free) due to the extremely inhomogeneous degree distribution, many nodes have only a few links. The nodes with small connectivity will be selected with a much higher probability and these removals changed the diameter in a small scale [4]. During attack on high degrees nodes, the random network does not show any difference irrespective of selection with random or descending degree nodes [4]. In scale-free metabolite network, targeted removal (Table 3a, 3b) show drastic change in diameter due to small number of nodes with very high connectivity. The diameter almost doubles when 5% of the nodes were removed.

Table 2a: Random Removal of Nodes from the Metabolite Network

Nodes	Clustering Coefficient	Diameter
90	0.045	56
180	0.045	51
270	0.045	39
360	0.046	37
450	0.048	34

Table 2b: Random Removal of Nodes from the Random Network

Nodes	Clustering Coefficient	Diameter
90	0.001	9
180	0.001	10
270	0.001	10
360	0.001	10
450	0.001	11

Table 3a: Targeted Removal of Nodes from the Metabolite network

Nodes	Clustering Coefficient	Diameter
90	0.01853	53
180	0.0123	30
270	0.00525	25
360	0.00225	25
450	0.002	25
540	0.002	11

Table 3b: Targeted Removal of Nodes from the Random Network

Nodes	Clustering Coefficient	Diameter
90	0.001	8
180	0.001	9
270	0.001	9
360	0.001	11
450	0.001	12

3.3 Betweenness vs. Degree Distribution

In order to understand the relation between high degree and high betweenness, the betweenness is plotted as a function of connectivity (Fig. 2). The metabolite network shows that most metabolites have low neighborhood connectivity but very high betweenness. This shows that many metabolites typically connect pathways and are potentially important metabolites. These results suggest that the network has modular organization with the high-betweenness and low-connectivity nodes as important links between these modules. The selected nodes with high degree and betweenness centrality are hubs and they are important nodes that control the overall network interaction. Hub metabolites include Pyruvate, Gibberelin, Stemmadiene, Anthracene cis-1,2-dihydrodiol which have been investigated to important in *A.thaliana*.

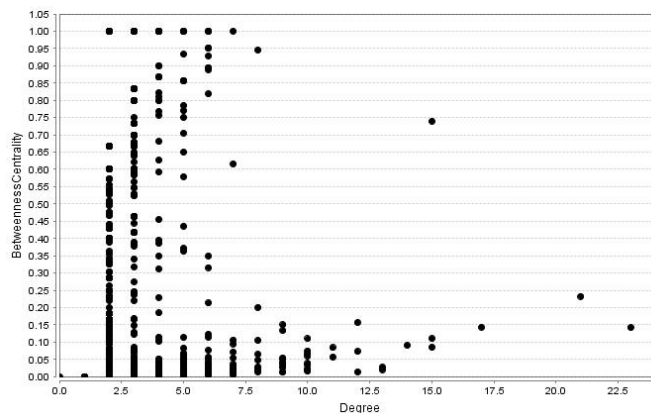


Fig. 2: Betweenness (B) is plotted as a function of connectivity (k) for metabolite network

3.4 Modularity

Modularity can be defined as a cellular functionality which can be seamlessly partitioned into a collection of modules. Each module is a discrete entity of several elementary components and performs an identifiable task, separable from the functions of other modules.

We used the Newman and Girvan edge-betweenness method to calculate the number of clusters available in the network. This algorithm identifies edges in a network that falls between communities and then removes them, leaving behind just the communities themselves [6]. We have utilized the Radatools [9] to apply the algorithm and the input files were the .NET files of the network.

Community detection using Newman's algorithm [10] detects 101 communities in metabolite network. The largest community had 506 metabolites that had the highest interaction within the group and lower interaction outside the group (Table 4a). Metabolites taking part in similar functional type of reactions will share common properties.

Table 4a: The Communities in *A.thaliana* Substrate graph and the number of Nodes in each Community

Number of Community	Number of Nodes
1	506
2	399
3	375
4	337
5	295
6	250
7	37
8	36
9	36
10	18
11	16
12	15
12	14
14	14
15	13
16	12
17	12
18	12
19	11
20	10
21	10
22-24	9
25-28	7
29-32	6
33-37	5
38-46	4
47-51	3
52-72	2

Table 4b: Pathways corresponding to the single node communities in the metabolite centric graph

	Isolated Pathways
1	Steroid hormone biosynthesis
2	Tyrosine metabolism
3	Monoterpenoid biosynthesis
4	Arachidonic acid metabolism
5	Indole alkaloid biosynthesis
6	Glycine, serine and threonine metabolism
7	Porphyrin and chlorophyll metabolism
8	Fructose and mannose metabolism

Table 4c: The Pathways corresponding to Nodes in largest Community(1) with 506 Nodes

	High Cluster Pathways
1	Biosynthesis of alkaloids derived from shikimate pathway
2	Porphyrin and chlorophyll metabolism
3	Naphthalene and anthracene degradation
4	Glycerolipid metabolism
5	Cyanoamino acid metabolism
6	ABC transporters
7	Pentose and glucuronate interconversions
8	Ascorbate and aldarate metabolism
9	Selenoamino acid metabolism
1	Biosynthesis of Terpenoids and steroids etc.,

These metabolites were further traced back to the pathways (Table 4c) that contained these metabolites and the list of pathways for the largest community was collected. They mainly constituted the amino acid metabolism pathways. There were 27 communities with only one metabolite called isolated metabolites (Table 4b). They produce similar intermediate compound and hence have interacted more closely. The pathways in the highest cluster were the amino acid metabolism pathways and the chlorophyll metabolism pathways.

4 Conclusion

Using earlier proposed methods we designed an automated method of metabolite network construction with KEGG metabolic pathway data. Analysis of this network gives us a complete idea of interaction between enzymes, reactions, and metabolites. The substrate centric graph helps in finding the conserved metabolites and reactions. This construction and analysis procedures can be further applied to an enzyme network and the enzyme evolution studies in *A.thaliana*.

5 References

- [1] Z. N. Oltvai and A. L. Barabasi, "Systems biology. Life's complexity pyramid," *Science*, vol. 298, pp. 763-4, Oct 25 2002.
- [2] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, pp. 101-13, Feb 2004.
- [3] L. H. Hartwell, *et al.*, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47-52, Dec 2 1999.
- [4] A. L. Barabasi, "Scale-free networks: a decade and beyond," *Science*, vol. 325, pp. 412-3, Jul 24 2009.
- [5] K. Radrich, *et al.*, "Integration of metabolic databases for the reconstruction of genome-scale metabolic networks," *BMC Syst Biol*, vol. 4, p. 114, 2010.
- [6] A. L. Barabasi and E. Bonabeau, "Scale-free networks," *Sci Am*, vol. 288, pp. 60-9, May 2003.

- [7] J. G. Joung, *et al.*, "Plant MetGenMAP: an integrative analysis system for plant systems biology," *Plant Physiol*, vol. 151, pp. 1758-68, Dec 2009.
- [8] H. W. Ma and A. P. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, vol. 19, pp. 1423-30, Jul 22 2003.
- [9] E. Ravasz, *et al.*, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551-5, Aug 30 2002.

A Professional Science Master Degree in Health Informatics

Kevin Daimi
 Mathematics & Computer Science Department
 University of Detroit Mercy
 4001 W. McNichols Rd., Detroit, Michigan 48221
 daimikj@udmercy.edu

Greg Grabowski
 Department of Biology
 University of Detroit Mercy
 4001 W. McNichols Rd., Detroit, Michigan 48221
 grabowgm@udmercy.edu

ABSTRACT

The demand for health informatics professionals is a response to the growing medical data bases resulting from governmental initiatives, as well as globalization of geographical information systems. This is compounded by the diversity of professions at all levels of health care. Health informatics goal is to develop integrated information systems for the exchange of data within the health care system, while ensuring security, confidentiality, and privacy. This paper proposes a Professional Master Degree in Health Informatics, which considers the multidisciplinary aspects of health care, recognizes the need for standardized assessment to ensure quality, and moves toward fulfillment of the demands placed on informatics in the health care forum.

Key Words

Health Informatics, Professional Science Master, Curriculum Design, Program Assessment, Program Outcomes

I. INTRODUCTION

Health informatics is the science of systematically processing data, information, and knowledge in medicine and clinical research [6]. The term "health" encompasses the multiple professions involved in the delivery of health care. These professions not only involve individuals considered to be at the forefront of delivery, such as medical doctors, dentists, physician assistants, nurse practitioners, and nurses; but also include laboratory and imagery technicians, pharmacists, social scientists, librarians, financial and budget managers, and public health and clinical researchers. Informatics has become a critical component in the management and effective use of ever-changing and continuously-growing data generated in clinical care and research. [3], [6], [8], [10]. As the medical field technologically advances and grows, demands for the organization and management of

information increase in complexity. Challenges facing health informatics in response to these demands include the development of coherent and integrated information systems, consideration of the high degree of information exchange resulting from the multidiscipline nature of health care and research, and security associated with confidentiality, privacy, and propriety issues [6], [16].

In the forum of public health, health informatics includes applications of knowledge from computer/information science, management, organizational theory, psychology, political science, communications, epidemiology, toxicology, microbiology, and law [4], [7], [16]. Its varied applications focus on health trends within populations, rather than clinical settings. Demand for proficiency in health informatics at the patient and population level is expected to grow, and to match medical advancement and increased health care accessibility. Professionals educated in health informatics are needed to meet these challenges, resulting in the improvement in the quality and efficiency of health care [4], [6], [8], [10].

The call for an electronic medical record for all Americans by 2014, which was made by President Bush in his 2004 State of the Union Address, recognized the demands and challenges of health informatics. President Obama continued this pledge, as well as called for digitized medical records within five years and passed the American Recovery and Reinvestment Act [ARRA] with Congress [12]. This Act provided significant momentum to health care reform through informatics technology [12]. In order to fulfill this pledge it is estimated that 50,000 health informatics professionals will be required [7]. The Canadian government is also committed to universal electronic medical records for its citizens by 2016 [14]. The overall goal of governmental incentives is the promotion of health within populations, prevention of disease with regard to effectiveness, expediency, cost, and acceptability, and direct governmental responsiveness to health issues [12] [16]. In order to meet the 2014 goal, the ARRA has provided funding for

the development of health informatics curriculum, degree programs, and competency certification to train and increase the workforce in response to health care reform [12].

Projected workforce needs are also being addressed by the American Medical Informatics Association's 10 x10 initiative program. The goal of this initiative was to train 10,000 health care professions in the field of health informatics by 2010 [7]. Likewise, medical schools and health care systems in the United States, France, Germany, and the Netherlands have incorporated health informatics into their medical curricula and training. However, many nations are lagging behind despite the impending flood of universal electronic medical records their governments are calling for and the dire need of health informatics professionals [5], [6], [14]. Such projection-based goals are feasible in developed countries; however, the health informatics work force needs of developing countries are unknown, yet anticipated to exceed that of developed countries [7].

On national and global levels, geographical information system technology and its entry into industry resulted in the generation of numerous data sets. [11]. These data sets were created independently and therefore lack a common structure, as well as a reluctance of their creators to share this information. Many more data sets are not documented. These issues call for the continued improvement and implementation of support mechanisms, and development of guidelines for geoportals technologies. Geoportals and repositories for data can provide the infrastructure to support the management of data and open its accessibility to the public for research, industry, academia, and public health policy [11]. Global consideration of health informatics still needs to encompass work force profiles, cultural and language variations, and recruitment and training across various and non-compatible systems [7].

Library science and informatics share the same objective, that being the management, sorting, and delivery of information: however the professional boundary between the two disciplines is becoming less distinctive [13]. Library science is typically seen as a sub-discipline of informatics that focuses on the user's need and the purpose of information. This focus has become less the domain of a "traditional library", but has transcended its physical confinement through digital media and network infrastructures. As the focus of library science broadens it overlaps with the broader focus of informatics. Both engage in information management and the development of its delivery, with librarians contributing to the former and informaticians contributing to the latter. Librarians, as expert searchers, have the training in data base organization

and understand the "architecture" of search systems, which allows them to provide relevant and high impact information despite lacking expertise in specific subject matter [15]. This complementary relationship benefits the developing field of health informatics as the repositories of information with their reservoir of data digitally coexisting with research and medical institutions. This symbiotic relationship is integral in the development and implementation of medical knowledge bases.

Because all health care professionals will be confronted with health informatics as global data sets consolidate and medical knowledge bases grow due to nations' conversion to electronic medical records, professionals will be exposed to informatics education within their professional studies and during career development. The variety of professions and their specific needs within health care require different modes of education methodologies, and require qualified educators that demonstrate competence in health informatics. The overarching outcome for informatics users in health care is to enable health care professionals to efficiently and responsibly use information processing methodology and technology; whereas, graduates specializing in informatics must be prepared for careers in health informatics in academic, health care, or industrial settings [3], [8], [10].

Master and higher graduate degrees provide the skills and knowledge for researchers, system developers, and educators. Most health informatics programs are at the graduate level and either share curricula with related programs or have a separate tract. Of 177 surveyed programs, 91 graduate level degrees fall within the bioinformatics area [1]. Those outside of bioinformatics are in the focused areas of nursing, dentistry, and Cheminformatics. Likewise, the housing of health informatics programs is also variable, with graduate programs affiliated with health science (31%), medical (25%), public health (16%), and computer science (16%) schools. Within these schools 37% of the health informatics programs are interdepartmental, and the remainder housed in biology (21%) and computer/information science [13%] departments, and medical schools (10%), [9]. Survey of admissions into the Health Informatics program at the University of Victoria indicated most students entered the program directly upon completion of their bachelor degree and have at least five years of work experience. Approximately half of their graduates intend to enter academia [2]. This reflects the general trend of Health Informatics programs, which are typically oriented towards primary health providers from various disciplines with the goal of producing educators and

researchers. Development of contemporary health informatics programs must recognize the need for health informatics professionals at all levels of healthcare, and adopt a broader focus if the ARRA goal of universal electronic health record is to be attained, developed, and managed.

The aim of this paper is to propose a curriculum design for a Professional Science Master in Health Informatics to contribute to meeting the enormous demand for health informatics professionals. The Professional Science Master will prepare leaders, researchers, entrepreneurs, and educators in the Health Informatics field.

II. Health Informatics Tasks

As health informatics continues to develop, priorities and goals need to be identified. The following is a tentative composition for the design of a Professional Master Degree Program [1], [9], [12], [16].

- Interdisciplinary professionals required for business practice, healthcare delivery and medical research require re-engineering of information systems, so that information can be shared.
- Facilitate the electronic gathering, storing, and interchange of patient data for analysis in a collaborative manner.
- Synergize health informatics with evidence-based medicine and its implementation through the development of clinical guide lines.
- Educate and train healthcare providers to use health informatics and its accompanying technology effectively.
- Improve security for patient confidentiality in regards to diagnostic reporting, use of patient clinical records for research, or finances.
- Create an informed patient environment through linking educational systems that involve prevention, diagnosis, and treatment to electronic patient records and distance care.
- Identify new technologic opportunities for their use in healthcare, such as bioinformatics.
- Enhance the fields of epidemiology and public health.
- Support transducers and mobile devices that report a patient's physiologic status.
- Contribute to good practice decision-making policies in the management and financial aspects of healthcare and public health.

- Reduce health care disparities within populations.

III. HEALTH INFORMATICS APPLICATIONS

Health informatics has many critical real life applications. Examples of its applications include: [1], [4], [9], [11], [12], [16].

- Development of patient-oriented interactive computer-based programs that provide information, support, patient status, and decision-making formats for underserved populations and high risk patients.
- Conversion of medical records into electronic formats to be shared amongst the various professional levels within the health care system.
- Re-engineer processes to maintain quality of health at permissible costs.
- Surveillance of disease incidence and vaccination patterns to identify trends, make predictions, and improve efficiency.
- Development of e-health (electronic accessible) and m-health (mobile phone accessible) applications in underserved populations.
- Create health spatial data infrastructures supported by geoportals, such as the OneGeology Portal, which delivers environmental data for medical research

II. GRANTS SUPPORTING BODIES

A number of organizations provide support for health informatics research and educational programs. Among these are the following:

- National Institute of Health
- National Library of Medicine
- Strategic Health IT Advanced Research Projects [SHARP]
- Informatics Training for Global Health Program
- International Medical Informatics Association
- Beacon Community Program
- Robert Wood Johnson Foundation

III. PROFESSIONAL SCIENCE MASTER

A. Program Objectives

- 1) Develop within graduates the level of Health Informatics proficiency needed for the professional practice.
- 2) Instill within graduates the ability to effectively communicate ideas and outcomes, both orally and in writing, in a logical manner.
- 3) Develop within graduates the appreciation for and an understanding of the need to maintain high ethical standards.
- 4) Instill within graduates the ability to demonstrate effective leadership and entrepreneurial thinking.
- 5) Prepare graduates for pursuing a doctoral degree in Health Informatics.

B. Program Outcomes

Students should be able to:

1. Students will be able to demonstrate proficiency in storing, retrieving, and interpreting health-related data sets in computer systems, and an awareness of their limitations.
2. Students will be aware of the need to communicate effectively to recognize the specific informational needs of different professionals in health care such as researchers, physicians, nurses, health economists, laboratory technicians, and librarians.
3. Students will be able to understand the role of information systems in the development and implementation of interactive programs that monitor patient physiology or provide supportive services.
4. Students will be able to recognize the various types of data, and effectively filter information and adopt new methods of searching information.
5. Students will be able to assess the quality of data as it pertains to specific health care areas and ensure its accuracy.
6. Students will be able to define and implement the principles of data protection, confidentiality, and privacy rights as they pertain to health care.
7. Students will be able to comprehend the supportive role of informatics in research, diagnostics, health management, public health, and decision making processes.
8. Students will be able to value ethical principles as they apply to patient rights and the data management.
9. Students will be able to demonstrate knowledge of leadership effectiveness in various health informatics fields, and innovational thinking, as well as functioning in teams.

C. Admission Requirements

In order to be admitted to the program, the applicant:

1. Must hold a bachelor's degree [or equivalent] with a minimum GPA of 3.0.
2. Must have a bachelor degree in health informatics, computer science/informatics, health science, health administration, biology, or public health.
3. Must have taken courses in a programming language (such as C++, Java, or Perl), Data Structures, Machine Organization, Calculus and Discrete Mathematics.
4. Must make up for deficiencies in undergraduate preparation by taking some prerequisite courses.
5. May have courses waived after passing a department test with a grade of at least a "B", if applicants have academic or work experience equivalent to any of the courses mentioned above.

D. Degree Requirements

The Professional Science Master in Health Informatics consists of 45 credits of coursework. Students must complete a 3-month internship in a health care industry, or a healthcare research institution. The 45 credits are distributed as follows:

▪ Health Informatics Core	12 cr.
▪ Health Informatics Elective	03cr.
▪ Leadership and Entrepreneurship	12 cr.
▪ Computing Core	12 cr.
▪ Computing Elective	03 cr.
▪ Research	03-06 cr.

E. Course Requirements

Courses representing each of the above areas are provided in Tables I - VI below. If a thesis is pursued, only one course of the list of classes in Tables II and V below is needed.

TABLE I

HEALTH INFORMATICS CORE	
Course Title	Credits
Introduction to Health Informatics	3
Clinical Informatics	3
Consumer Health Informatics	3
Public Health Informatics	3

TABLE II
HEALTH INFORMATICS ELECTIVES

Course Title	Credits
Electronic Health Care Records	3
Clinical Decision Support	3
Telemedicine	3
E-Health Systems	3
Legal and Business Issues	3
Health Systems Simulation	3
Advanced Topics	3

TABLE III
LEADERSHIP AND ENTREPRENEURSHIP CORE

Course Title	Credits
Healthcare Management	3
Healthcare IT Project Management	3
Health Informatics Entrepreneurship	3
Health Informatics Internship	3

TABLE IV
COMPUTING CORE

Course Title	Credits
Software Engineering	3
Web Technology	3
Security and Privacy	3
Data Mining	3

TABLE V
COMPUTING ELECTIVES

Course Title	Credits
Geographic Information Systems	3
Database Design	3
Systems Design	3
Software Requirements Engineering	3
Knowledge Management	3
Human Computer Interface	3
Quantitative Methods	3

TABLE VI
RESEARCH

Course Title	Credits
Health Informatics Design Project	3
Health Informatics Thesis	6

F. Degree Assessment

Many Health Informatics programs are customized to the suit the students professional needs in healthcare. As a result programs can become informal and self-directed [2]. The following assessment measures are recommended to assure overall program quality:

- 1) Individual course assessment to ensure that each course is achieving its learning outcomes and supporting the program outcomes.
- 2) A comprehensive program self study will be prepared for the purposes of any program review.
- 3) A Graduate Survey will be employed to measure students' satisfaction with individual courses and the program as a whole.
- 4) A survey for the students taking the Health Informatics Design Project course or Thesis will be provided to measure the extent to which the program will achieve its learning outcomes and how well their learning experience matches the program objectives.
- 5) A Comprehensive Test will be devised to measure how well students are prepared to meet the learning objectives. This test will be offered as part of a capstone course with a weight of 30% and focus on the course requirements for the general knowledge areas recommended above
- 6) An Exit Survey will be offered to students completing the program to solicit their feedback on the program and on how to improve it.
- 7) An Alumni Survey will be used to discover how well our graduates feel they were prepared for their current position.
- 8) An Employer Survey will be prepared to obtain the feedback of employers on how well our graduates are prepared for their positions.
- 9) An Internship Survey will be used to measure students' performance in prospective organizations.

IV. CONCLUSION

The growing demand for health informatics professionals is a direct result of developed countries moving forward with universal electronic medical records for their citizens in an effort to efficiently

provide a higher quality and more cost effective health care system. As medical knowledge bases grow in size and complexity, the demand for their organization and management grows. The projected work force in health informatics required to manage this growth is in the ten of thousands for the immediate future, and is expected to increase with the development of geoportals supporting spatial data infrastructures. The former will facilitate the exchange of data amongst health and research institutions across state and national boundaries.

Professional Science Master Degree program design, proposed in this paper, is in direct response to the demand for health informatics professionals and encompasses the multidisciplinary aspect of health care. Program objectives emphasize the basic knowledge areas of informatics, while program outcomes account for the professional diversity an informatician would encounter. Admission requirements and degree assessment ensure overall program quality, a necessity in face of the "specifically professionalized" trend health informatics programs tend to follow. Because of the professional diversity inherent in health care, housing of the program can occur in multiple departments or schools; however, most health informatics programs are found in health science, medical, and public health schools, or biology and computer /information science departments.

REFERENCES

- [1] J. Brender, C. Nohr, and P. McNair, "Research Needs and Priorities in Health Informatics," *International Journal of Medical Informatics*, Vol. 58-59, pp. 257-28, 2000.
- [2] H.D. Covvey and A. B. Pidduck, "Health Informatics Education, Working Paper," in *Waterloo HIP Position Paper*, pp. 1-50, 1999.
- [3] S. Garde and E. Hovenga, "Australian Health Informatics Educational Framework," in *Australian Health Informatics Educational Framework*, pp. 1-15, 2005.
- [4] D.H. Gustafson, R.P. Hawkins, E.W. Boberg, F. McTavish, B. Owens, M. Wise, H. Berhe, and S. Pingree, "CHESS: 10 Years of Research and Development in Consumer Health Informatics for Broad Populations, Including the Underserved," *International Journal of Medical Informatics*, Vol. 65, pp. 169-177, 2002.
- [5] A. Hasman and A. Albert, "Education and Training in Health Informatics: Guidelines for European Curricula," *International Journal of Medical Informatics*, Vol. 45, pp. 91-110, 1997.
- [6] R. Haux, "Aims and Tasks of Medical Informatics," *International Journal of Medical Informatics*, Vol. 44, pp. 9-20, 1997.
- [7] W. Hersh, A. Margolis, F. Quiros, and P. Otero, "Building a Health Informatics Workforce in Developing Countries," *Health Affairs*, Vol. 29, No. 2, pp. 275-278, 2010.
- [8] International Medical Informatics Association, "Recommendations of the International Medical Informatics Association [IMIA] on Education in Health and Medical Informatics" *Methods of Information Medicine*, Vol. 39, pp. 267-277, 2000.
- [9] J. Kampov-Polevoi, and B. M. Hemminger, "Survey of Biomedical and Health Care Informatics Programs in the United States," *Journal of Medical Library Association*, Vol. 98, No. 2, pp. 178-181, 2010.
- [10] J. Mantas, E. Ammenwerth, G. Demiris, A. Hasman, R. Haux, W. Hersh, E. Hovenga, K.C. Lun, H. Marin, F. Martin-Sanchez, and G. Wright, "Recommendations of the International Medical Informatics Association [IMIA] on Education in Biomedical and Health Informatics," *Methods of Information Medicine*, Vol. 2, pp. 1-16, 2010.
- [11] T. Mathys, and M. N. K. Boulos, "Geospatial Resources for Supporting Data Standards, Guidance, and Best Practice in Health Informatics," *BMC Research Notes*, Vol. 4, No. 19, pp. 1-18, 2011.
- [12] J. Murphy, "The Journey to Meaningful Use of Electronic Health Records," *Nursing Economic*, Vol. 28, No. 4, pp. 283-286, 2010.
- [13] G.J. Perry, N.K. Roderer, and S. Assar, "A Current Perspective on Medical Informatics and Health Sciences Librarianship," in *Journal of Medical Library Association*, Vol. 93, No. 2, pp. 199-204, 2005.
- [14] S. Strauss, "Canadian Medical Schools Slow to Integrate Health Informatics into Curriculum," *Canadian Medical Association Journal*, Vol. 182, No. 12, pp. E551-E552, 2010.
- [15] E. C. Whipple, J.J. McGowan, B.E. Gixon, and A. Zafar, "The Selection of High-Impact Health Informatics Literature: A Comparison of Results between the Content Expert and Expect Searcher," *Journal of Medical Library Association*, Vol. 97, No. 3, pp. 212-218, 2009.
- [16] W. A. Yasnoff, P. W. O'Carroll, D. Koo, R.W. Linkins, and E. M. Kilbourne, "Public Health Informatics: Improving and Transforming Public Health in the Information Age," *Journal of Public Health Management Practice*, Vol. 6, No. 6, pp. 67-75, 2000.

Laterality of Motor Control before the Advent of Experimental Psychology: Revisiting David Kinnebrook's "Error of Judgement" at Greenwich in 1796

Iraj Derakhshan, MD, Neurologist

Formerly, Associate Professor of Neurology, Case Western Reserve and Cincinnati Universities,
Cleveland and Cincinnati, Ohio, USA

Currently, Private Practice of Neurology
415 Morris St, Ste. 401
Charleston, WV 25301
Tel 304 343 4098
Fax 304 343 4598

Abstract:

Background: The dismissal of David Kinnebrook as an astronomical laborer in 1796 has afforded him a special position in the history of experimental psychology: "a martyr of science." This is because he was "ushered away" from his work at the Royal Greenwich Observatory through no fault of his own. Here, using data available in the literature and insights from a new understanding in laterality of motor control (i.e. one-way callosal traffic circuitry) it is shown that Kinnebrook, though right handed, was wired as a left handed person would be; with delayed reaction times in noticing events arising from his right hemisphere (delayed saccades to the right).

Keywords: Neurology, handedness, 1-way callosal traffic theory, reaction time, brain anatomy

1 Introduction

"In the 18th and early 19th centuries, astronomers were required to make difficult judgments, based on a combination of auditory and visual cues, in order to time stellar transits. A well-known story from the history of science is the firing in 1796 of Kinnebrook, an assistant to Maskelyne, the Astronomer Royal of England.

Kinnebrook was relieved of his job for giving inaccurate readings of stellar transits. Although he had provided readings in agreement with Maskelyne's 18 months prior to his dismissal, the hapless Kinnebrook

by August 1795 had begun to give times that differed from Maskelyne's by one-half second. Subsequently, Kinnebrook's readings grew even more discrepant, so by the time of his firing they were almost a second later than Maskelyne's. This matter might not have attracted much interest had not Maskelyne recorded it in *Astronomical Observations at Greenwich*. Seventeen years later, in a history of Greenwich Observatory published in German, Kinnebrook's tribulation came to the attention of Bessel, an astronomer at Königsberg. Bessel conducted a series of studies culminating in the notion of the personal equation [reaction time], the name given the systematic difference in recording times found to characterize the stellar transits of almost any pair of astronomers. From the perspective of reliability theory, the personal equation [reaction time] itself was not a highly significant discovery, for it refers to systematic error, not the random error treated by reliability theory. What interests us, instead, is Bessel's finding that the personal equation [reaction time] itself is a variable quantity, one that differs from one pair of astronomers to another. This variation suggests random or accidental errors in observations, errors that, if neither controllable nor amenable to elimination, at the least demand an explanation grounded in a theory or a scientific law."¹

2 Methods and Results

In many of the accounts of the subject, David Kinnebrook is considered a "martyr of science"

because of the role of his dismissal in inaugurating experimental psychology as a scientific discipline.^{2,3} As an assistant astronomer, Kinnebrook was constantly and regularly late (by 500-800 milliseconds) in marking down the transits of stars as they crossed the meridian; but only in his second year of employment and beyond. In the citation above, Traub asks for an explanation (scientific law) of the delay as shown by Kinnebrook as well as an explanation for the time line of its occurrence. The latter undertaking, however, has never been done before.

Based on time resolved anatomical data supporting 1-way callosal traffic theory^{4,5}, the present article provides an explanation for the widening of the gap between the performance of Kinnebrook and that of his superior Nevil Maskelyne (the royal astronomer) in the years of his employment as an assistant to Maskelyne.

From the neurological perspective, the key to explaining Kinnebrook's performance is the direction of motion of the stars monitored by the contestants and the number of such observations in each direction over the period under scrutiny. However, to my knowledge there are no published reports concerning the issue. We know that the procedure followed by observers in calibrating the clock called for single or multiple saccades in the direction of the appearance of the star (s) (right or left of the observer). According to the information available, Kinnebrook's performance variability over the two years of employment was contrary to the performance of other assistants working in the same observatory, as documented by a later Astronomer Royal (Sir Spencer Jones).⁶ Spencer-Jones also recorded that two of his six assistants (R.C. & W.D) consistently lagged behind a "standard observer" in reacting to transiting stars as they watched a list of the so-called "clock stars." Clearly, therefore, we are not dealing with a very rare phenomenon though the physiological nature of the phenomenon has remained obscure thus far. Remarkably, this occurred despite the fact that astronomers had already discovered that "direction of star's motion could introduce a change in the personal equation."^{7,8} Thus, while comparing his own reaction times with those of a colleague while using a chronograph in an observatory in Madras, a certain officer, named W.M. Campbell, became aware of his own tardiness in catching a glimpse of the objects appearing to his right compared to that of his colleague who clocked them that same way. In the words of Campbell, "Captain Heaviside observing in advance of me [by 64 milliseconds]."⁸

This latter experiment performed in 1877 is equivalent to that of visual half-field paradigm conducted in today's laboratory, employing the so-called Poffenberger paradigm.⁴

To summarize, according to the 1-way callosal traffic circuitry (see below for details), by recording the fact that he was delayed in observing objects moving from right to left, Captain Campbell was documenting his own status as a neural left hander compared to his comrade in arms, Captain Heaviside, who was faster responding to the events appearing on his right side.

3 Discussion

The generally accepted view that each hemisphere controls the movement of the contralateral side has been questioned recently. There is overwhelming evidence that our handedness is a reflection of the fact that only one hemisphere houses the command center with the nondominant hemisphere engaged in carrying out the commands issued by the dominant for movements planned for the nondominant side of the body.

According to 1-way callosal traffic circuitry,^{4,5} it is the directionality of callosal traffic (i.e. whether signals move from left to right hemisphere or the reverse) that determines the status of one hemisphere as that of action hemisphere (the command center, dominant hemisphere), where all commands are issued for movements occurring on either side of the body. According to this understanding, a person's behavioral (avowed) handedness is only a guide to his or her directionality of callosal traffic (i.e. neural handedness); the neural and behavioral handedness in an individual subject are in agreement in only ~ 80 percent of the population. In the remaining 20 percent of individuals display an avowed (behavioral) handedness opposite for which they are truly wired (see below for further explanation).

The above estimates as to the laterality of command center are derived from a variety of clinical sources. Thus, since the action hemisphere is the same as the speech hemisphere, the incidence of crossed aphasia and crossed nonaphasia in penetrating brain injuries does provide an estimate of the incongruities under consideration;⁹ as do anomalous occurrences of neglect in lesions affecting the left hemisphere in ostensibly right handed subjects,¹⁰ occurrences of aphasia after removal of supratentorial tumors of the right hemisphere in right handed subjects,¹¹ as well as occurrences of alien hand syndrome on the ostensibly dominant side of the subject following lesions affecting the minor hemisphere or its afferent callosal connection.^{12,13} Experimentally, persons incongruous in neural and behavioral handedness display a faster manual reaction time to stimuli on their (ostensibly) nondominant side, or a negative crossed uncrossed differential (negative CUD) in applications of Poffenberger paradigm.^{4,14}

According to 1-way callosal traffic circuitry, all actions originate in the major hemisphere, including those of moving the eyes to the side (saccades) and swallowing, with the command traversing the corpus callosum to activate the minor hemisphere which in turn moves the nondominant side of the body once it receives the command.^{5,15} Electrophysiologically, the abovementioned callosally mediated delay has been repeatedly documented in bimanual “simultaneous” movements recorded with different techniques, indicating precedence of the neurally dominant side in moving when a simultaneous movement was intended.¹⁶⁻¹⁸ For the saccades, a similar ratio of faster response to the stimuli from the left hemisphere was found in two of the twelve (presumably) right handed subjects described by Honda,¹⁹ confirming an earlier study by Hamers and Lambert in a lexical decision task on 15 right handed subjects (wherein three of the participants responded faster to stimuli from the left side).²⁰ Elsewhere, I have provided detailed explanation regarding the subjects reported by Honda.⁴ To the above may be added the reports on those ostensible right handers who drew longer lines or larger geometrical designs with their nondominant hands, while drawing simultaneously with both hands^{21,22} and the three of seventeen right handers who showed higher refractory cue-cost for their ostensibly dominant right hand (instead of the left) in a study by Buckingham et al.²³

Since movements of the eyes to the sides is governed by the same circuitry that underpins hand movements, moving the eyes to the neurally dominant side occurs at a faster speed than moving them to the opposite direction; by an amount equal to the interhemispheric transfer time (IHTT, i.e. the time needed for transfer of the command signals issued in the action hemisphere for movements occurring on the nondominant side of the body). It has been shown that such commands are implemented by the minor hemisphere upon receiving the same via corpus callosum and anterior commissure.^{4,5,22,24}

According to the above sketched circuitry, David Kinnebrook must have been a member of the above described neural-left but behavioral-right handers who saw the objects arising from his right hemifield at a significant delay compared to a real (neuro-behavioral congruent) right hander (such as Maskelyne). For objects arising from his left hemisphere, however, Kinnebrook would have reacted faster than his superior Maskelyne. This provides a plausible explanation for his acceptable performance in his first year of employment at Greenwich. Accordingly, vast majority of transit trials performed by Kinnebrook in the years 1795 and 1796 must have been instances in which the transiting stars were moving from right to left, resulting in his ever

worsening performance compared to Maskelyne as the time went on (leading to his eventual dismissal).

The validity of 1-way callosal traffic circuitry has been confirmed in several recent studies.²⁴⁻²⁸ The criticism raised by Goble,²⁹ is based on a failure to fully understand the import of the circuitry, i.e. that the “critical” issue as to dominance of one limb over its counterpart is the comparative speed with which the two arms move, *regardless of the subject's claim as to his/her own handedness*. Thus, in addressing the problems of classification of handedness by employing a dexterity evaluation method (i.e. the speed of performance), Satz et al³⁰ found that “roughly 69 per cent of the left-handers showed superior performance on the left hand, and 75 per cent of the dextrals showed superior right hand performance. In this study, “self-classified right-handers displayed less variable and better performance with their preferred right hand.” In the same vein, Wyke in an experiment involving speed of performance concluded that *“handedness influences the speed of arm movements, and the results are in line with previous observations showing that tests of rapid repetitive movements of the arms might provide a more critical index of handedness than is obtained from observations of non-repetitive arm movements.”*³¹ The above described motor asymmetry is reflected as an asymmetry in perceptual span in experiments involving the oculomotor system; and as a wider excursion of the neurally dominant side of the body in bimanual simultaneous drawing test (a simple paper and pencil test for determining the laterality of motor control in those able to hold a pen in each hand and draw a line simultaneously with both).^{32,33}

Finally, the clinical import (validity) of the above mentioned time-resolved observations in the motor realm is corroborated by the hitherto ignored observation that only one-half of the 35 supratentorial cases of cerebral herniation described by Kernohan and Woltman in their 1929 article displayed (false localizing) pyramidal signs ipsilateral to the tumor; corroborating the fact that callosal interhemispheric transfers are one-way in directionality and excitatory in nature (i.e. from the major to the minor hemisphere).³⁴

4 Conclusion

Approximately one in five people in society displays a handedness for which he or she is wired in the opposite direction. The dismissal of Kinnebrook by Astronomer Royal of England was based on an assumption that all right handers are created equal. Kinnebrook was in fact wired as a left hander. Bimanual simultaneous drawing task, an inexpensive and very accurate method, based on the existence of

laterality in motor control, has shown quantitatively that this assumption has been flawed. Similarly flawed was Maskelyne's methodology, i.e. failure to control for the direction of motion of objects that the two observers were tracking at the time; thus the sad outcome for the "hapless" David Kinnebrook.

5 References

- [1] Traub RE. Classical test theory in historical perspective. *Educ Meas Issues Pract* 1997; 16:8–14.
- [2] Brooks GP, Brooks RC. The improbable progenitor. *J Roy Astron Soc Can* 1979; 73: 10-23.
- [3] Schaffer S. Astronomers Mark Time: Discipline and the Personal Equation. *Sci Context* 1998; 2: 115-145.
- [4] Derakhshan I. Handedness and macular vision: laterality of motor control underpins both. *Neurol Res* 2004; 26: 331-337.
- [5] Derakhshan I. How do the eyes move together? New understanding help explain eye deviations in patients with stroke. *CMAJ* 2005; 172: 171-173.
- [6] Spencer-Jones H. The measurement of time. *Rep Prog Phys* 1937; 4:1-26 (table 2)
- [7] Sanford EC. Personal equation; ii variations in the mount of personal equation. *Am J Psychol* 1889; 2: 271-298.
- [8] Campbell WM. On a peculiarity of personal equation. *Monthly Notices of the Royal Astronomical Society* 1877; 37: 283-284. (suppl)
- [9] Mohr JP, Weiss GH, Caveness WF, et al. Language and motor disorders after penetrating head injury in Viet Nam. *Neurology* 1980; 30: 1273-1279.
- [10] Beis JM, Keller C, Morin N, et al. Right spatial neglect after left hemisphere stroke. Qualitative and quantitative study. *Neurology* 2004; 63: 1600-1605.
- [11] Thomson AM, Taylor R, Whittle IR. Assessment of communication impairment and the effects of resective surgery in solitary, right-sided supratentorial intracranial tumours: a prospective study. *Br J Neurosurg* 1998; 12: 423-429.
- [12] Derakhshan I. Laterality of motor control and the alien hand always coincide: further observations on directionality in callosal traffic underpinning handedness. *Neurol Res* 2009; 31: 258-264.
- [13] Jeannerod M. The origin of voluntary action. History of a physiological concept. *C. R. Biologies* 2006; 329: 354-362.
- [14] Derakhshan I. Crossed-uncrossed difference (CUD) in a new light: anatomy of the negative CUD in Poffenberger's paradigm. *Acta Neurol Scand* 2006; 113: 203-208.
- [15] Teismann IK, Rainer D, Steinstraeter O, Pantev C. Time-Dependent Hemispheric Shift of the Cortical Control of Volitional Swallowing. *Hum Brain Mapp* 2009; 30:92–100.
- [16] Kristeva R, Keller E, Deecke L, Kornhuber HH. Cerebral potentials preceding unilateral and simultaneous bilateral finger movements. *Electroencephalogr Clin Neurophysiol* 1979; 47: 229-238.
- [17] Shen YC, Franz EA. Hemispheric competition in Left-Handers on bimanual reaction time tasks. *J Mot Behav* 2005; 37: 3-9. (tables 1-3)
- [18] Walsh RR, Small SL, Chen EE, Solodkin A. Network activation during bimanual movements in humans. *Neuroimage*. 2008; 43: 540–553.
- [19] Honda H. Idiosyncratic left-right asymmetries of saccadic latencies: examination in a gap paradigm. *Vision Res* 2002; 42: 1437-1445.
- [20] Hamers JF, Lambert WE. Visual field and cerebral hemisphere preferences in bilinguals. In *Language Development and Neurological Theory*, Sj Segalowitz and FA Gruber (Eds.). Academic Press, New York, 1977 (PP 57-62)
- [21] Derakhshan I. Right sided weakness with right subdural hematoma: motor deafferentation of left hemisphere resulted in paralysis of the right side. *Brain Inj* 2009; 23: 770-774.
- [22] Semjen A, Summers JJ, Cattaert D. Hand Coordination in Bimanual Circle Drawing. *J Exp Psychol: Human Perception and Performance* 1995; 21: 1139-1157. (subject DA)
- [23] Buckingham G, Main JC, Carey D. Asymmetries in motor attention during a cued bimanual reaching task: Left and right handers compared. *Cortex* 2010, 45: 1-9. (Fig. 4a)
- [24] Ioannides AA, Fenwick PB, Pitri LI, Liu E. A step towards non-invasive characterization of the human frontal eye fields of individual subjects. *Nonlinear Biomed Phys* 2010; 4 suppl 1: S 11
- [25] Azémar G, Stein SF, Ripoll H. Effects of ocular dominance on eye-hand coordination in sporting duels. *Sci & Sports* 2008; 23:263-277.
- [26] Thummler P, Karnath HO. Eye deviation after left hemispheric stroke. *Tubingen University School of Medicine*, 2009. (suppl)

- [27] Badoud S, Rouiller EM. Relation between hand preference and hand dominance in human: A behavioral study. University of Fribourg School of Medicine, 2009. (suppl)
- [28] Becker E, Karnath HO. Neuroimaging of eye position reveals spatial neglect. *Brain* 2010; 133: 909-914.
- [29] Goble D. Validity of using reaction time as a basis for determining motor laterality. *J Neurophysiol* 2007; 97: 868.
- [30] Satz P, Achenbach K, Fennell E. Correlations between assessed manual laterality and predicted speech laterality in a normal population. *Neuropsychologia* 1967; 5: 295-310.
- [31] Wyke M. Influence of direction on the rapidity of bilateral arm movements. *Neuropsychologia* 1969; 7: 189-194.
- [32] Lindell AK, Nicholls MER, Kwantes PJ, Castles A. Sequential processing in hemispheric word recognition: The impact of initial letter discriminability on the OUP naming effect. *Brain Lang* 2005; 93: 160-172.
- [33] Derakhshan I. Attentional asymmetry or laterality of motor control? Commentary on Buckingham et al. *Cortex* 2011; 47: 509-510.
- [34] Derakhshan I. The Kernohan-Woltman phenomenon and laterality of motor control: Fresh analysis of data in the article "Incisura of the crus due to contralateral brain tumor". *J Neurol Sci* 2009; 287: 296

SESSION

COMPUTER-ASSISTED MEDICAL CARE + SERVICE SYSTEMS, DIAGNOSTIC TOOLS IN BIOMEDICAL, AND COMPUTER-BASED MEDICAL SYSTEMS

Chair(s)

TBA

Diagnosis of Breast Cancer using Averaged Proximity Measure between Samples

R.I. Andrushkiw¹, E.N. Golubeva², D.A. Klyushin², Yu.I. Petunin², and N.V. Boroday³

¹New Jersey Institute of Technology, Newark, NJ 07102, USA

²Kyiv National Taras Shevchenko University, Kyiv, Ukraine

³R.E.Kavetsky Institute of Experimental Pathology, Oncology and Radiobiology, Kyiv, Ukraine

Abstract - In this paper we determine the risk degree of malignancy in tumors that have been diagnosed as benign. To do this we compute the proximity measure between corresponding morphological and densitometrical indexes of digital images of interphase nuclei of buccal epithelium in patients with benign tumors, malignant tumors and individuals that are practically healthy (without tumors).

Keywords: Breast Cancer, Diagnosis, Proximity Measure

1 Introduction

The development of a neo-plastic process in an organism is usually accompanied by changes in the functional interrelations between its organs [1]. In a series of investigations [2-4], it was proved that changes in oral mucosa are an early indicator of some pathological processes in an organism. Hence, it is possible to use buccal epithelium for the investigation of changes which are going on in epitheliocytes of oral mucosa in patients with oncological pathology. Such changes are called MAC – malignancy associated changes. Since violation in the function of organs and systems in an organism are related to changes in the functional state of a cell genome, the morphometric and densitometric parameters of epitheliocytes in buccal epithelium may be used as a criterion of MAC [5]. The use of quantitative automatic image analysis opened the possibility of estimating the content of DNA in the nuclei and compactness of the chromatin, which characterizes the functional state of cells in various pathological processes, including tumors [6].

2 Materials

We consider three groups of patients: G_1 – patients suffering from breast cancer (38 cases), G_2 – patients suffering from fibroadenomatosis (44 cases) and G_3 – group of practically healthy women (33 cases). Smears

from various depths of the spinous layer were obtained (conventionally they were denoted as median and deep), after gargling and removing the superficial cell layer of the buccal mucous. The DNA content stained by Feulgen was estimated using the Olympus computer analyzer, consisting of the Olympus BX microscope, Camedia C-5050 digital zoom camera and a computer. We investigated from 40 to 60 nuclei in each preparation. The DNA-fuchsin content in the nuclei of the epitheliocytes was defined as a green component of a RGB-value.

3 Methods

3.1 Proximity measure

Let $x = (x_1, \dots, x_n) \in G$ and $x' = (x'_1, \dots, x'_m) \in G'$ be samples from general populations G and G' , and $x_{(1)} \leq \dots \leq x_{(n)}$ and $x'_{(1)} \leq \dots \leq x'_{(m)}$ be their order statistics. We test the hypothesis on the identity of absolutely continuous distribution functions $F_G(u)$ and $F_{G'}(u)$ of the general populations G and G' . Suppose that $F_G(u) = F_{G'}(u)$. Denote by $A_{ij}^{(k)}$, $k = 1, 2, \dots, m$, a random event that x'_k lies in the interval $(x_{(i)}, x_{(j)})$:

$$A_{ij}^{(k)} = \{x'_k \in (x_{(i)}, x_{(j)})\}, \quad (i < j).$$

The probability of this event is determined by the formula [7, p. 126]:

$$P(A_{ij}^{(k)}) = P(x'_k \in (x_{(i)}, x_{(j)})) = p_{ij}^{(n)} = \frac{j-i}{n+1}.$$

Let

$$p_{ij}^{(1)} = \frac{h_{ij}^{(n)} m + g^2/2 - g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m + g^2/4}}{m + g^2},$$

$$p_{ij}^{(2)} = \frac{h_{ij}^{(n)} m + g^2/2 + g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m + g^2/4}}{m + g^2},$$

where $h_{ij}^{(n)}$ is the relative frequency of the event $A_{ij}^{(k)}$ in m trials and $g = 3$.

Denote by N the number of all confidence intervals $I_{ij}^{(n,m)} = (p_{ij}^{(1)}, p_{ij}^{(2)})$, ($N = n(n-1)/2$) and by L the number of intervals $I_{ij}^{(n,m)}$ containing probabilities $p_{ij}^{(n)}$. Then we get the p-statistics:

$$h^{(n,m)} = \rho(x, x') = \frac{L}{N}.$$

Letting $h_{ij}^{(n)} = h^{(n,m)}$, $m = N$, $g = 3$, we get the confidence interval for the p-statistics $h^{(n,m)}$:

$$p^{(1)} = \frac{h^{(n,m)}N + g^2/2 - g\sqrt{h^{(n,m)}(1-h^{(n,m)})N + g^2/4}}{N + g^2},$$

$$p^{(2)} = \frac{h^{(n,m)}N + g^2/2 + g\sqrt{h^{(n,m)}(1-h^{(n,m)})N + g^2/4}}{N + g^2}.$$

3.2 Averaging of proximity measure

Let $x = (x_1, x_2, \dots, x_n)$ be a sample from the general population G , which is obtained by simple random sampling. Let $y_1 = (y_1^{(1)}, \dots, y_{n_1}^{(1)})$, ..., $y_K = (y_1^{(K)}, \dots, y_{n_K}^{(K)})$ be similar samples, which are obtained from general populations G'_1, \dots, G'_K accordingly. Let G^* be a group which consists of the populations G'_1, \dots, G'_K :

$$G^* = \{G'_1, \dots, G'_K\}.$$

Based on the obtained statistics, let us calculate the p-statistics $\rho(x, y_i)$ [7] and define the quantity

$$\bar{\rho}(x, G^*) = \frac{1}{K} \sum_{i=1}^K \rho(x, y_i), \tag{1}$$

The value $\bar{\rho}(x, G^*)$ is called the averaged p-statistics between the sample x and the group of general population $G^* = \{G'_1, \dots, G'_K\}$.

3.3 Method of diagnostics

Let P be a patient with unknown diagnosis: breast cancer or fibroadenomatosis. Let x_1, \dots, x_n be the sample which consists of the areas of interphase nucleus of buccal epithelium of some patient P . We'll denote x_1, \dots, x_n as a sample x . The group G^* is a teaching sample, which consists of similar indexes of patients $P_1^{(1)}, \dots, P_K^{(1)}$ with

breast cancer, or patients $P_1^{(2)}, \dots, P_m^{(2)}$ with fibroadenomatosis.

Consequently, the teaching sample with indexes of patients with cancer is

$$G_1^* = \{G_1^{(1)}, \dots, G_K^{(1)}\},$$

and the teaching sample with indexes of patients with fibroadenomatosis is

$$G_2^* = \{G_1^{(2)}, \dots, G_m^{(2)}\}.$$

The criterion for the diagnostics of breast cancer consists of two parts. First, patients with cancer and their averaged p-statistics in group G_1^* and their averaged p-statistics in group G_2^* are considered.

Thus, the first patient $P_1^{(1)} \in G_1^*$ is considered. After excluding $P_1^{(1)}$ from group G_1^* we compute the averaged p-statistics between $P_1^{(1)}$ and the group

$$\tilde{G}_1^{(1)} = G_1^* \setminus P_1^{(1)} = \{P_2^{(1)}, \dots, P_K^{(1)}\}: \bar{\rho}(P_1^{(1)}, \tilde{G}_1^{(1)}).$$

Then, patient $P_2^{(1)}$ is excluded from group G_1^* and in this way the group $\tilde{G}_2^{(1)}$ is obtained:

$$\tilde{G}_2^{(1)} = G_1^* \setminus P_2^{(1)}.$$

After that, the averaged p-statistics $\bar{\rho}(P_2^{(1)}, \tilde{G}_2^{(1)})$ is computed. Then the next patient is excluded and the averaged p-statistics is computed, and so on. This method is called «one-out». The results of the computations are in table 1.

In the same way, computations of the averaged p-statistics for patients $P_i^{(1)}$, ($i = 1, \dots, K$) and group G_2^* are done. The obtained values $\bar{\rho}(P_i^{(1)}, G_2^*)$ are given in table 1.

Table 1. Averaged p-statistics between patients with breast cancer and patients with fibroadenomatosis

№	Averaged p-statistics between patients with breast cancer and	
	patients from group with breast cancer	patients from group with fibroadenomatosis
101	0,85	0,47
130	0,86	0,48
132	0,79	0,45
135	0,88	0,49
139	0,7	0,43
154	0,78	0,46

155	0,8	0,46
156	0,84	0,48
157	0,87	0,48
159	0,7	0,42
160	0,67	0,42
161	0,83	0,48
165	0,86	0,48
170	0,89	0,47
180	0,79	0,45
183	0,83	0,48
185	0,84	0,47
191	0,86	0,48
194	0,89	0,49
196	0,8	0,46
197	0,71	0,44
198	0,78	0,45
200	0,77	0,45
201	0,79	0,47
204	0,87	0,48
208	0,77	0,46
209	0,87	0,49
210	0,86	0,48
212	0,65	0,43
34	0,79	0,45
36	0,82	0,46
37	0,69	0,43
39	0,75	0,45
41	0,86	0,48
43	0,8	0,45
46	0,79	0,46
54	0,77	0,45
87	0,74	0,44

Analysis of table 1 shows that all values of the averaged p-statistics between patients with cancer and the group of patients with fibroadenomatosis are situated between $x_{(1)} = 0,649$ and $x_{(n)} = 0,887$, and all values of the averaged p-statistics between patients with cancer and group of patients with fibroadenomatosis are situated between $x_{(1)} = 0,415$ and $x_{(n)} = 0,493$.

From the above it follows that the diagnosis of patients with breast cancer was made without error. Let H denote the hypothesis that a patient has cancer and let \bar{H} be the alternative hypothesis that a patient has fibroadenomatosis. Then the probability of type I error is equal to zero: $P(\bar{H}/H) = 0$. So for all patients with breast cancer the diagnosis is correct.

For the diagnostics of patients with fibroadenomatosis we used averaged p-statistics between patients with fibroadenomatosis and group G_1^* , as well as the averaged p-statistics between patients with fibroadenomatosis and group G_2^* . The results of the computation are given in table 2.

Table 2. Averaged p-statistics between patients with fibroadenomatosis and patients from the group with breast cancer and the group with fibroadenomatosis

	Averaged p-statistics between patients with fibroadenomatosis and	
	patients from group with breast cancer	patients from group with fibroadenomatosis
158	0,81	0,46
162	0,82	0,47
17	0,83	0,48
1	0,84	0,47
203	0,82	0,47
33	0,76	0,45
401	0,32	0,34
402	0,33	0,34
403	0,32	0,34
406	0,32	0,34
407	0,33	0,34
418	0,33	0,34
419	0,32	0,33
422	0,33	0,34
423	0,33	0,34
424	0,33	0,34
434	0,32	0,34
435	0,33	0,34
440	0,32	0,34
443	0,33	0,34
459	0,33	0,34
460	0,33	0,34
464	0,33	0,34
472	0,33	0,34
473	0,33	0,34
478	0,32	0,34
47	0,82	0,47
486	0,32	0,34
490	0,33	0,34
491	0,32	0,34
494	0,32	0,33
496	0,32	0,34

498	0,32	0,33
499	0,32	0,34
500	0,32	0,34
501	0,33	0,34
506	0,33	0,35
507	0,33	0,34
509	0,33	0,34
510	0,33	0,35
57	0,8	0,47
59	0,69	0,43
61	0,87	0,49
63	0,84	0,48

Analysis of the data in the first column of table 2, using confidence interval $I = (0,649;0,887)$ constructed by order statistics $x_{(1)} = 0,649$ and $x_{(n)} = 0,887$, shows that 11 patients with fibroadenomatosis were diagnosed as having cancer. Indexes of patients numbered 58, 162, 17, 1, 203, 33, 47, 57, 59, 61, 63 belong to the confidence interval $I = (0,649;0,887)$. The rest of the patients (33 persons) were diagnosed correctly.

Hence, the error is equal to 25%. The same results were obtained using the second confidence interval $I = (0,415;0,493)$ and the second column from table 2 with averaged p-statistics $\bar{p}(P_i^{(2)}, \tilde{G}_i^{(2)})$. In this case the same patients were diagnosed incorrectly.

In order to decrease this error we apply the second part of the diagnostics. Thus, to increase the accuracy of the criterion we consider the data of the group of practically healthy women.

We compute the averaged p-statistics α_2 between patients with fibroadenomatosis and group G_2^* , as well as averaged p-statistics β_2 between patients with fibroadenomatosis and group G_3^* . Then we calculate the ratio of the obtained averaged p-statistics α_2 and β_2 . This ratio is denoted as γ_2 :

$$\gamma_2 = \frac{\alpha_2}{\beta_2} .$$

The results of the computations are given in table 3.

Table 3. Ratio of averaged p-statistics $\gamma_2 = \alpha_2 / \beta_2$ for patients with fibroadenomatosis

	Ratio of averaged p-statistics γ_2
158	0,755

162	1,18
17	1,074
1	1,015
203	1,263
33	0,706
401	1,17
402	1,173
403	1,171
406	1,203
407	1,173
418	1,171
419	1,169
422	1,179
423	1,175
424	1,16
434	1,16
435	1,176
440	1,178
443	1,203
459	1,149
460	1,179
464	1,158
472	1,175
473	1,171
478	1,17
47	0,882
486	1,206
490	1,169
491	1,181
494	1,167
496	1,209
498	1,158
499	1,181
500	1,172
501	1,204
506	1,175
507	1,162
509	1,167
510	1,166
57	1,366
59	0,567
61	1,066
63	1,2

Similarly, we obtain the ratio between α_1 and β_1 , where α_1 is the averaged p-statistics between all breast cancer patients and the group of women patients with

breast cancer, and β_1^* is averaged p-statistics between all breast cancer patients and the group of practically healthy women:

$$\gamma_1 = \frac{\alpha_1}{\beta_1}$$

The results of the computations is given in table 4.

Table 4. Ratio of averaged p-statistics $\gamma_1 = \alpha_1 / \beta_1$ for the patients with breast cancer

	Ratio of averaged p -statistics γ_1
101	1,719
130	1,469
132	1,473
135	1,833
139	0,945
154	1,644
155	1,473
156	1,628
157	1,651
159	0,96
160	0,794
161	1,721
165	2,182
170	1,686
180	1,2
183	2,26
185	1,565
191	1,677
194	2,002
196	2,291
197	2,167
198	1,115
200	2,251
201	1,906
204	1,97
208	2,104
209	1,927
210	2,09
212	2,204
34	1,273
36	2,089
37	2,098
39	1,99
41	1,097
43	2,025

46	1,1
54	1,1
87	0,963

Analysis of table 4 shows that the ratio γ_1 is situated between minimal $x_{(1)} = 0,794$ and maximal 2,291 order statistics. So, the confidence interval $I = (0,794; 2,291)$ covers the main distributed mass of the general population for γ_1 .

On the other hand, the data from table 3 shows that the ratio of the averaged p-statistics γ_2 , of patients with indexes 158, 33 and 59, does not belong to the confidence interval I . So, these patients are diagnosed as patients with fibroadenomatosis. Hence, only 8 patient are diagnosed incorrectly. After applying the second part of the criterion, the type II error is equal to 18,18%.

Let $H_0 = H$ denote the hypothesis that a patient has breast cancer and let $H_1 = \bar{H}$ be the hypothesis that a patient has fibroadenomatosis. Then $P(\bar{H}/H) = 0$, $P(H/\bar{H}) = 0,1818$.

4 Conclusions

Let us formulate the main conclusions based on the obtained results:

1) If after applying the criterion we diagnose fibroadenomatosis, then the probability of such event is close to one. The probability of the event that these patients have breast cancer is practically zero. The results are unexpected, since according to medical statistics the error in the diagnosis of fibroadenomatosis is approximately 20%.

2) If after applying the criterion we diagnose breast cancer, then the probability of such event is equal to 81,8%. The probability of not detecting a patient suffering from breast cancer is equal to zero.

In order to the increase the accuracy of detecting cancer, one can use another method [8, 9] in conjunction with the method discussed above. In that case all patients with fibroadenomatosis are diagnosed correctly, and patients with breast cancer are diagnosed with an error of 7,9%.

The application of the two methods together gives an accuracy of 92% and sensitivity of 100% .

5 References

[1] Shabalkin I.P. et al. (2000) Violation of systemic relations in an organism under cancerogenesis. Proc. Russian Academy of Science. 375 (3):404-409 (in Russian).

- [2] Nieburgs H.F. et al. (1962) Buccal all changes in patients with malignant tumors. *Lab. Invest.* 11(1):80-88.
- [3] Ogden G.R. et al. (1990) The effect of distant malignancy upon quantitative cytologic assessment of normal oral mucosa. *Cancer.* 65(3):477-480.
- [4] Mayansky A.N. et al. (2004) Reactivity of buccal epithelocytes: indication of local and general homeostasis (survey). *Clinical laboratory diagnostics*, 8:31-34 (in Russian).
- [5] Avtandilov T.T. et al. (2004). Ploidometrical diagnosis of precancerous processes and cancer of cervix on cytological preparations. *Clinical laboratory diagnostics* . 11:45-47 (in Russian).
- [6] Linder J. (1994) Imaging cytometry: applications in diagnostic pathology. / *Anal. And Quant. Cytology and Histology.* 16(1):53-57.
- [7] Klyushin D.A., Petunin Yu.I. (2008) Evidence medicine. Application of statistical methods – Williams , Moscow, (in Russian).
- [8] R. Andrushkiw, D. Klyushin, M. Boroday, Yu. Petunin, K. Golubeva (2009) Determination of risk degree and making diagnosis of cancer by averaging of p-statistics – BIOCOMP'09, Las Vegas, Nevada, USA, 2009, pp 892-895.
- [9] Klyushin D.A., Petunin Yu.I., Golubeva K.M. (2010) Computer citogenetic diagnostics of breast cancer and fibroadenomatosis based on areas of nucleus of buccal epithelium – *Bulletin of University of Kiev, Series: Physics and Mathematics*, №1, Kiev, 2010, P. 138-143 (in Russian).

An Extensible Software Architecture to Facilitate Disaster Response Planning

Martin O'Neill II, Armin R. Mikler, and Tamara Schneider

Center for Computational Epidemiology and Response Analysis, University of North Texas, Denton, TX USA

Abstract—*Disaster mitigation planning must rely on an analysis of available data. However, the vast amounts and different types of data make this data analysis intractable without the use of computational tools. The RE-PLAN Response Plan Analysis framework was designed to create the computational tools needed for these analyses. Although the methodology it employs was originally designed to facilitate validation of mitigation plans for biological emergencies arising from a release of hazardous biological substances, the RE-PLAN framework has been generalized to serve as a launching point for the development of a wide variety of disaster mitigation and evacuation planning scenarios. A tool using the RE-PLAN framework for feasibility analysis of ad hoc clinics for treating the population following a biological emergency event has been created. This paper focuses on the design and implementation of the RE-PLAN framework and how it has been used to address the hazardous biological substance release mitigation data analysis problem.*

Keywords: biological emergencies, disaster mitigation planning, emergency response, evacuation planning, POD throughput, public health preparedness

1. Introduction

The RE-PLAN Response Plan Analysis framework was designed to facilitate the construction of computational tools for the analysis and development of disaster mitigation and evacuation plans. Although this framework was originally designed around a specific disaster mitigation problem, its modules are generalized and may be used in the context of a wide variety of disaster and evacuation situations. Additional modules may be added to the framework in order to address concerns peculiar to specific disaster or evacuation situations. However, the existing framework comprises a significant set of analysis techniques relevant to a wide variety of different situations.

The RE-PLAN framework emerged from a methodology developed for analyzing the feasibility of ad hoc facilities for treating populations following a release of hazardous biological substances [1][2]. A set of facilities is considered feasible if its operational efficiency [3] is capable of meeting service requirements (e.g. specific time frames for service completion or proportions of populations to be served) without exceeding available resources (e.g. transportation network capacities or limitations of facility infrastructure).

This paper will highlight the following main architectural components of the RE-PLAN framework and the modules designed to implement them:

- Facility selection and service area determination - Sets of facilities in existing plans may be analyzed or sets of feasible facilities may be generated with respect to the populations' geographic distributions. This component is primarily responsible for the selection of facilities and generation of service areas.
- Logistics calculator - Calculates how the population utilizes the transportation network to travel to the facilities. These calculations facilitate the analysis of conditions on the transportation network resulting from response plan implementation.
- Facility requirement and traffic analysis - Population distribution among the facilities can be examined to facilitate resource distribution, and parking lot entry and exit rates at each facility are determined. Parameters may be modified to increase or decrease the number of individuals each facility is capable of serving per day. Traffic conditions resulting from the placement of facilities may be analyzed using geographic population data, road network data, and traffic count observation data. Parameters such as people per car, time of day, and day of week may be modified to facilitate mitigation planning.

Computational models of biological emergency events show the importance of a policy of aggressive mass treatment [4][5], and delays in this treatment can lead to increased numbers of casualties [6]. Routing and scheduling for timely delivery of medications to treatment facilities have been examined in [7], and strategies regarding medication distribution among the facilities have been explored in [8]. However, the distribution of medications to the population remains a challenging problem [9]. To aid larger cities in planning for these contingencies, the United States Department of Health and Human Services instituted the Cities Readiness Initiative (CRI) in 2004 [10]. An initial evaluation of CRI indicates that the initiative has improved mass treatment preparedness [11]. Studies have been conducted regarding shortcomings and optimization strategies inside service facilities during a biological emergency [12][13]. However, less attention has been paid to how the population will be delivered to facilities for treatment during response plan implementation.

Surveillance systems, such as the BioSense system created by the Centers for Disease Control and Prevention [10], use data from disparate sources to facilitate early detection of biological emergency events. The World Health Organization endorses the use of public health surveillance systems, referring to them as being, “the cornerstone of public health security.” [14] A wide variety of different data sources such as over-the-counter drug sales [15][16], internet search query patterns [17][18], and personal web log (blog) data [19] have been explored to detect public health events. However, further exploration of available data sources to allow decision makers to develop and analyze their response plans is needed [12].

Disaster mitigation and evacuation planning must rely on an analysis of quantitative data [20]. However, the vast amounts and different types of data make this analysis intractable without the use of computational tools. A set of generalized modules was designed around data analysis problems relevant to a variety of disaster mitigation and evacuation planning and analysis. These modules, which facilitate the analyses of many different types of geographic, spatio-temporal, and demographic data, comprise the RE-PLAN Response Plan Analysis framework.

2. Plan Analysis Problems

A feasible response plan must be able to accomplish its assigned tasks without exceeding the available resources. The assigned task of facilities hosting ad-hoc clinics in a biological emergency is to serve the entire population within specific time frames. This facility throughput problem may be affected by geographic population distribution and facility location, constraints of the transportation network, limitations of the facilities, and the availability of personnel and supplies. All of these factors can be separated into two groups: problems that may occur at the facilities themselves, and problems that may occur in each facility’s service area.

2.1 Problems at the Facilities

The primary question to be answered when analyzing problems that may occur at a facility is, “Can the facility serve the number of people in its assigned population under the given time constraints?” To answer this question, the service area of each facility must be determined. Once the service areas have been determined, the population of each service area may be analyzed to estimate the requirements of each facility. These requirements lead to further questions such as:

- “Can the parking lot at each facility support the number of cars which must enter and exit under the time constraints?”
- “How long will it take for each facility to serve its assigned population?”
- “Based on the assigned population of each facility, are there any special requirements for the facilities?”

2.2 Problems in the Service Areas

The population in each service area is unlikely to be uniformly distributed. Further, transportation network infrastructure is likely to be irregularly distributed across the service areas. Therefore, the locations of the facilities in relation to population distribution and transportation network resources must be examined. A facility may be capable of serving its assigned population under the time constraints. However, if the transportation network is incapable of delivering individuals to the facility in a timely manner, resources at the facility may be under-utilized. This motivates the question, “How will the implementation of the facilities in a given plan affect the traffic situation on the transportation network?”

3. Methodology

The methodology employed by the RE-PLAN Framework was designed to analyze plan feasibility using large amounts of quantitative data from disparate sources. These data include population data, transportation network data, and traffic count observation data. Combined with assumptions from public health officials, this methodology is used to create a model which facilitates the analysis of conditions resulting from the implementation of response plans. Further, the model allows public health officials to experiment with alternate plan scenarios while exploring the data underlying the computational model.

3.1 Facility Selection and Service Area Determination

A variety of facility selection and service area determination methods have been developed. The most simple method allows users to select facility locations directly and then uses these locations as a basis for determining service areas. A variation on this method allows the user to select a set of facility locations and to set the number of desired facilities. The most feasible of the selected facility locations are then automatically determined, and service areas are created based on the locations of these facilities. Yet another method creates uniform service areas based on selected demographic variables. Once the service areas have been created, the facility locations are selected within them.

3.2 Service Facility Analysis

Once service areas have been determined for each facility, the population assigned to each is also known. Each facility’s requirements and feasibility can then be analyzed with respect to the population it is assigned to serve. Depending on the outcome of this analysis, users may choose to redistribute personnel and resources among the facilities or to modify their response plans. Methods for analyzing the set of facilities include examining the distribution of the population among them, calculating the amount of time it

will take for each to serve its assigned population, and estimating the traffic situation in each's parking lot.

Decisions regarding the distribution of personnel and supplies among the facilities requires an analysis of the distribution of the load (in this case, the population) across the set of facilities. Facilities assigned a greater load must be assigned more personnel and resources to adequately handle this load. Those assigned drastically smaller or larger loads can be identified, allowing public health officials to adjust their response plans accordingly. Once the distribution of load across the set of facilities is acceptable, and personnel and resource distributions have been determined, analysis can continue.

The estimated service time T_i required for each facility i to serve its assigned population p_i can be calculated by $T_i = \frac{p_i s}{w_i}$, where s is the amount of time it would take to serve a single individual and w_i is the number of individuals who may be served at facility i in parallel. Any facility whose estimated service time is close to or exceeds the mandated time constraints may be infeasible. The feasibility of a particular facility may be improved by increasing the number of individuals whom it may serve in parallel, by decreasing the assigned population, or by shortening the amount of time it would take to serve a single individual.

The total number of cars a_i which must visit each facility i can be calculated by $a_i = \frac{p_i}{f}$, where f is the average number of people who will travel in each car. This average is determined by public health officials and may be based upon such factors as demographic data or familiarity with the population to be served. The rate r_i at which cars must enter and exit the parking lot at each facility i can be calculated by $r_i = \frac{T_i}{a_i}$. Further, the rate ρ_i at which cars must enter and exit each parking lot in order to meet the time constraints ω can be calculated by $\rho_i = \frac{\omega}{a_i}$.

3.3 Traffic Analysis

If the transportation network of a service area is incapable of delivering the population to the facility in a timely manner, the facility will be under-utilized, causing it to be infeasible. Therefore, the traffic situation resulting from implementation of specific response plans must be analyzed. To accomplish this task, a model of how the population travels inside each service area must be used. This model must combine geographic population distribution data with transportation network data into a context which facilitates the analysis of traffic conditions at specific points inside each service area.

If each service area is divided into rings around the facility as shown in Figure 1, the population of the outer rings must travel through the inner rings to arrive at the facility as shown in Figure 2. Further, it must also travel back through these inner rings to return to its origin. Each ring of the service area may then be divided further into a series of segments around links in the transportation network which connect

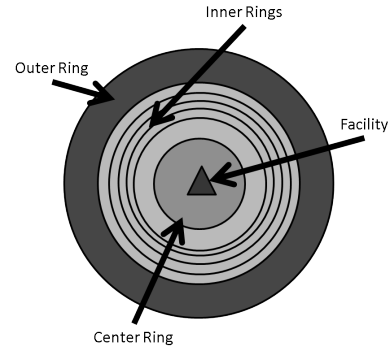


Fig. 1: Breaking a service area into rings of proximity to the facility

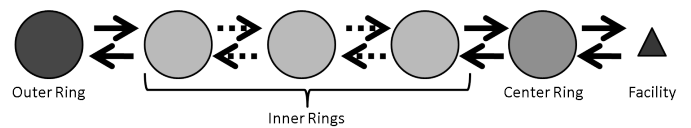


Fig. 2: How the population travels through the rings of proximity to the facility

them as shown in Figure 3. The population of the outer rings may then be modeled crossing from segment to segment on its way towards the facility. The links on the road network, shown in Figure 4, where the population crosses from ring to ring may be examined with respect to the number of individuals who must cross these links.

The population crossing each link may be divided by the average number of individuals per car to determine the load on each link caused by the implementation of the plan under the time constraints. The constraints of the transportation network are likely to vary. Therefore, the load caused by the implementation of the plan at a specific point must be analyzed with respect to the properties of the transportation network at that point (e.g. speed limit, number of lanes, functional class, or maximum physical capacity).

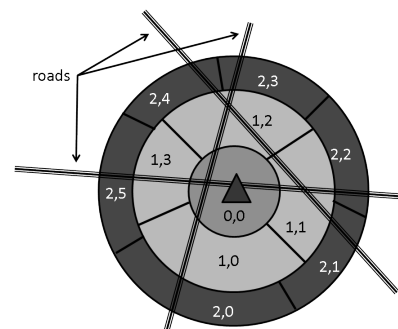


Fig. 3: Using only three rings of proximity, dividing rings into segments to model population flow across roads

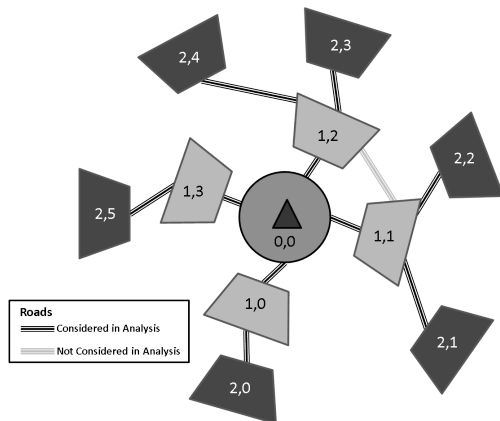


Fig. 4: Using only three rings of proximity, links in road network where population must cross from one ring to another are examined.

Disasters do not occur in a vacuum. Traffic to the facilities will not be the only load on the transportation network. This necessitates the analysis of normal “business as usual” base traffic side-by-side with traffic to the facilities. To accomplish this task, traffic count observation data must be used with other properties of the transportation network to project the load of base traffic across the network. Further, time-of-day and day-of-week become important considerations when dealing with base traffic to adequately represent peak and off-peak traffic periods.

4. RE-PLAN Framework

The RE-PLAN Architectural Framework consists of the Plan Designer, Plan Analysis Tools, and the Logistics Calculator. These three main architectural components communicate with the RE-PLAN Database to facilitate plan creation and analysis. Each component consists of a series of modules which may be redesigned, augmented, or replaced in order to change the underlying model being used. The work flow diagram in Figure 5 shows an overview of how the modules are used in each of the architectural components to design plans, calculate logistics, and perform analysis.

4.1 Plan Designer

The Plan Designer allows the user to create a response plan consisting of a set of facilities and assigned service areas. Figure 5 shows three different example paths through the modules of the Plan Designer. Each of these paths results in a set of facilities and service areas being stored on the RE-PLAN Database. The flexibility of these modules lies in the specific tasks they were chosen to perform.

The Facility Editor module allows users to create, edit, delete, import, and export facilities. All of these functions are accomplished through a point-and-click graphical interface. This module modifies the RE-PLAN Database as the user

modifies the facilities in a plan. The information regarding each facility which currently affects the plan analysis calculations are the facility’s longitude, latitude, type, status, and width. The longitude and latitude are used to specify a facility’s geographic location. A facility’s type may be used to modify how a facility affects logistical calculations or analysis of a plan. A facility’s status reflects whether it is *on* or *off*. Only facilities which are *on* are included in the logistic and analysis calculations. The width of a facility is the number of individuals which may be treated at this facility in parallel. Additional information (such as name, address, and comments) may be stored with each facility to assist officials in their planning, but this information is not used in the logistic or analysis calculations.

The Automatic Facility Selector module chooses locations for a user-specified number of facilities. If a list of facilities has already been chosen, this module selects the number of facilities from them. Once facilities have been selected, the Creator of Service Areas for Facilities module breaks down the area of interest such that every point in the area of interest is assigned to exactly one facility. Once these facilities and service areas have been determined, they are stored in the RE-PLAN Database.

The Creator of Uniform Service Areas module breaks down the area of interest into uniform service areas based on the demographic characteristics of the area’s population. After the service areas have been created, feasible facility locations are selected for each service area. Once the service areas and facilities have been determined, they are stored in the RE-PLAN Database.

The POD Analysis problem was examined and broken down into a series of modules which comprise the RE-PLAN tool. Together with the RE-PLAN database, these modules facilitate the analysis of POD-based biological emergency response plans.

4.2 Logistics Calculator

Once a plan has been created by the Plan Designer and stored in the RE-PLAN Database, the Logistics Calculator prepares this plan for analysis. Each facility’s service area is dissolved into rings of proximity. Each ring is dissolved into a series of segments around crossing points where links in the transportation network connect the rings, and the population of each segment is calculated. The population is then cascaded across the crossing points leading to the facility such that the number of individuals who must traverse each crossing point to reach the facility is known. The crossing points are stored in the RE-PLAN Database with their corresponding loads (in numbers of individuals).

4.3 Plan Analysis Tools

The Plan Analysis Tools facilitate analysis of response plans at the facility and on the transportation network. The Facility Requirement Analyzer module uses data from

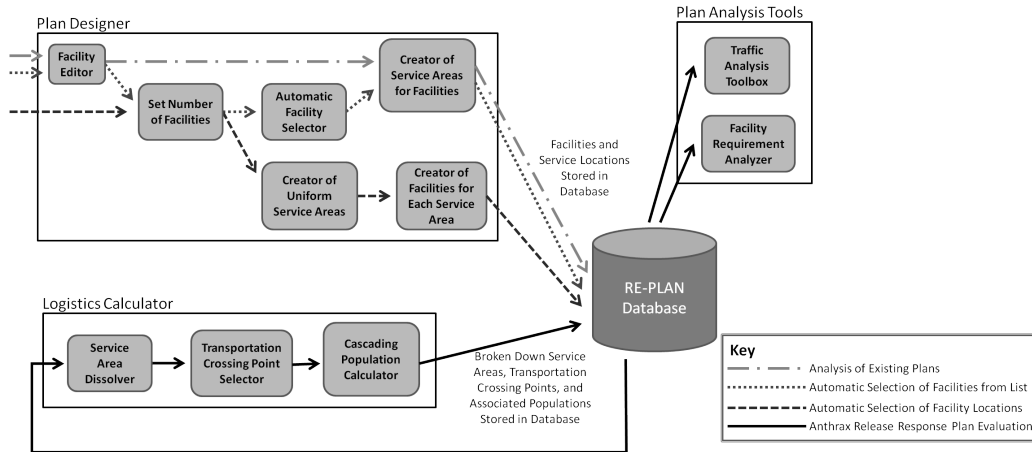


Fig. 5: RE-PLAN Framework Architectural Components and Work flow Diagram

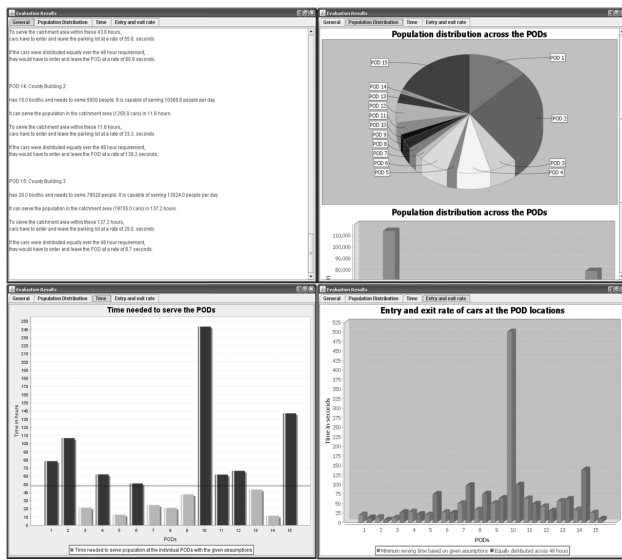


Fig. 6: Facility Requirement Analyzer graphical display

the RE-PLAN Database created by the Plan Designer to explore the expected situation at each facility. The Traffic Analysis Toolbox module uses data from the RE-PLAN Database calculated by the Logistics Calculator to facilitate exploration of traffic conditions resulting from response plan implementation. Although these modules are functionally separate, they comprise the set of analysis tools available in the RE-PLAN Framework.

4.3.1 Facility Requirement Analyzer

The Facility Requirement Analyzer module facilitates the analysis of response plan data created by the Plan Designer through a series of four tabs shown in Figure 6. Graphical representations of population load distribution among the facilities are created, allowing public health officials to easily

analyze personnel and supply distribution requirements of the facilities. Further, the population load on each facility is combined with user-specified assumptions regarding the width of each facility, the average number of people who will travel to the facility in each car, and the amount of time required to serve each person to create graphical representations for the analysis of each facility.

The first of the four tabs provides a written report for the entire plan and for each individual facility. The second tab includes two different graphical representations of the population distribution. The third tab shows the time required for each facility to serve its population under the current assumptions. Facilities which are capable of serving their populations within the mandated time constraints may be considered feasible and are shown in green while their infeasible counterparts are shown in red. The fourth tab shows estimates of the situation in each facility’s parking lot under the mandated time constraints as well as under the total amount of time required for each facility to serve its assigned population.

4.3.2 Traffic Analysis Toolbox

The Traffic Analysis Toolbox module combines geographic population data with transportation data in the context of the response plan logistics data. Traffic conditions at specific points may be examined with respect to the physical properties of the transportation network and the load on these points resulting from “business as usual” base traffic, traffic caused by implementation of the response plans, or a combination of both. Traffic conditions at each point are classified into one of six different classes which represent the ratio of load to the maximum physical capacity. Figure 7 shows how these classes are visually represented on a map to facilitate analysis of traffic conditions by personnel without the need for Geographic Information Systems (GIS) or computer programming expertise.

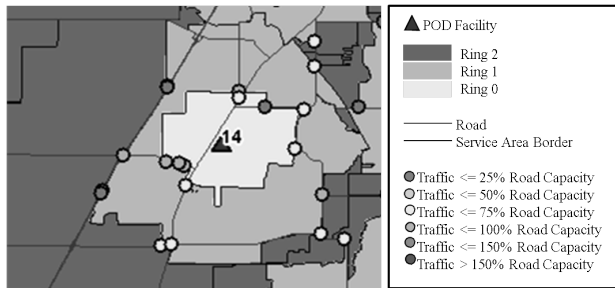


Fig. 7: Screen capture of traffic analysis using RE-PLAN hazardous biological substance release mitigation plan analyzer tool

Fig. 8: Traffic toolbox interface for analysis of Point of Dispensing (POD) facilities for hazardous biological substance release mitigation plans

The Traffic Toolbox graphical interface in Figure 8 allows users to easily change analysis parameters through a point-and-click interface. Traffic resulting from implementation of response plans and base traffic may be toggled *on* or *off* independently. Parameters which affect base traffic projections such as *weekday* versus *weekend* traffic and *time of day* may be adjusted. Assumptions which affect response plan traffic such as *people per car* and time constraints may also be adjusted.

4.4 RE-PLAN Database

The RE-PLAN Database stores all information about the response plans. It is also the primary way data is shared among the different architectural components of the RE-PLAN Framework. Three main categories of data are stored on the RE-PLAN Database: data concerning the area of interest, data which comprises response plans, and other data which facilitates the user experience and collaboration among users.

Data concerning the area of interest must be loaded into the RE-PLAN Database before plan creation or analysis begins. This includes spatial demographic and population data, spatial transportation network data, and traffic count observation data. If traffic count observation data is unavailable for the area of interest, data from another area may be loaded and used to train the Traffic Analysis Toolbox module. Data for multiple areas may be loaded into the RE-PLAN Database, and users may adjust parameters to choose a specific area of interest among them.

The RE-PLAN Database also stores all data resulting from the creation and analysis of individual response plans. The set of facilities, their service areas, and the load distributed across the transportation network are all stored in the database. For each facility, the facility's name, location, address, city, zip code, status, type, and other comments are stored in the database. If the Logistics Calculator has been used on a response plan, the dissolved service areas and population load on the transportation network are also stored in the database. Notes about each plan may also be stored with each plan in the database.

Each plan is owned by a specific user, and this association is stored in the RE-PLAN Database. This facilitates collaboration among users who may share their plans. Data concerning the currency of response plans is stored for the purpose of automatically deleting temporary database tables. Further data regarding the analysis progress of each response plan is also saved to enable RE-PLAN tools to open saved plans with the correct options and features enabled. For example, if a plan was saved before the Logistics Calculator was executed, when the plan is reloaded, the Traffic Analysis Toolbox should not be available to the user until the Logistics Calculator is executed.

5. Implementation

The RE-PLAN framework employs a client-server model. Client-side modules are written in Java for portability, and a PostgreSQL database with PostGIS is used on the RE-PLAN server for flexibility. This model facilitates collaboration among users whose client programs connect to the same server, thus allowing users to access each other's mitigation plans. Further, the client programs have minimal system requirements since most of the complex calculations are executed on the server. As a result, new hardware may not have to be deployed to execute the client programs.

Modules in the framework have been designed to be interoperable by incorporating import and export features. Sets of facilities may be imported or exported as Comma Separated Values (CSV) files. Many software packages commonly used in public health, disaster management, and city planning are capable of importing and exporting data as CSV files. Therefore, existing data may be imported into tools created with the RE-PLAN Framework and exported to other commonly used software packages.

Functionality to export entire plans as standard ESRI shapefiles allows data to be shared and further analyzed by those with GIS expertise. These plans may be published by creating maps in software packages such as ArcGIS or may be used to create online, interactive maps using OpenMap, Google Maps, or Microsoft Bing Maps. Therefore, the RE-PLAN Framework facilitates not only plan analysis, but plan distribution and implementation as well.

6. Discussion

The hazardous biological substance release mitigation tool created using the RE-PLAN Framework answers important questions regarding the implementation of specific response plans. The tool harnesses large amounts of quantitative data to estimate conditions at and requirements for each facility during implementation of specific mitigation plans. These conditions and requirements include the projected traffic situation at each facility's parking lot, the amount of time each facility requires to serve its assigned population, and the infrastructure needed by each facility to serve its assigned population. Further, the tool facilitates the analysis of the traffic situation on the transportation network resulting from the implementation of specific response plans.

The RE-PLAN Framework was developed in collaboration with public health officials. Their suggestions and comments were included in the methodology and implementation. Graphical interfaces were incorporated into the framework to allow use without the need for GIS or computer programming expertise. Although large amounts of quantitative data are used and may be accessed through the Framework, the design of graphical displays focused on specific aspects of the response plan analysis methodology. Therefore, while the data underlying the computational model may be accessed through the graphical interface, it is hidden by default to avoid clutter and confusion.

A version of the hazardous biological substance release mitigation tool has been created and deployed at a local county public health department. County public health officials have been trained to use this tool for analyzing their mitigation plans. Analysis performed at the county using this tool has led to the revision and modification of response plans. Local stakeholders have been trained regarding the modified plans, and preparations to implement these plans are underway.

The RE-PLAN Framework is comprised of a set of generalized modules. These modules may be used together in different contexts to address different response or evacuation scenarios. The methodology used may be adjusted by modifying existing modules, and additional modules may be created to address problems peculiar to specific scenarios. Further, the existing framework may be used to create tools with a web interface, thus enabling widespread distribution of RE-PLAN tools.

7. Limitations

As with all computational models, fidelity is limited by the accuracy and availability of underlying data sets. The RE-PLAN Framework uses several sets of data for which availability or currency may be problems. Examples of these data sets are population data, transportation network data, and traffic count observation data. Nonetheless, the RE-PLAN Framework has been developed to be as flexible as possible in accepting alternate data sets.

Geographic population distribution data is available in the United States from the U.S. Census Bureau. However, a full census is only conducted once every decade, leading to the use of potentially out-of-date data. If more accurate or current population distribution data is available from other sources, the RE-PLAN Framework is capable of incorporating this data into the model. The only limitation regarding which population distribution data may be used is that the framework is only designed to incorporate vector (not raster) data sets. However, raster data can easily be transformed into vector data using a wide variety of available tools.

Transportation network data may be incomplete or out-of-date for certain areas. Smaller neighborhood roads may be excluded from available data sets. However, these smaller roads do not greatly affect traffic analysis. Although newer links on the transportation network may not be included in available data, local public health officials who are using the tool will likely be familiar with the roads in their area. Attributes available for each link of the transportation network may differ from location to location. To address this problem, a variety of methods have been developed to use the framework with the available attributes.

Traffic count observation data may be unavailable for the vast majority of links in the transportation network. The RE-PLAN Framework addresses this sparse data problem by classifying the links in the transportation network and assigning links of the same class the same traffic loads. Traffic count observation data may contain inaccuracies due to the methods used in the collection of this data. If this data is too sparse or entirely unavailable for an area, the RE-PLAN Framework may be trained on data from a similar, but different, area. Nonetheless, it may not be difficult to accurately determine whether a specific traffic count represents conditions of high traffic speed and low traffic density, or of low traffic speed and high traffic density.

Acknowledgment

The project described was supported by Grant Number NIH 1R15LM010804-01 from the National Institutes of Health.

References

- [1] T. Schneider and A. R. Mikler, "RE-PLAN: A Computational Tool for Response Plan Analysis," *International Journal of Functional Informatics and Personalised Medicine*, vol. 3, no. 2, pp. 103–121, 2010.
- [2] T. Schneider, A. R. Mikler, and M. J. O'Neill II, "Computational Tools for Evaluating Bioemergency Contingency Plans," in *Proceedings of the 2009 International Conference on Disaster Management*, 2009.
- [3] Centers for Disease Control and Prevention, "Receiving, Distributing, and Dispensing Strategic National Stockpile Assets: A Guide for Preparedness, Version 10.02," 2006. [Online]. Available: http://www.kdheks.gov/cphp/download/SNS_Planning_Guide_V10.02.pdf
- [4] E. H. Kaplan, D. L. Craft, and L. M. Wein, "Emergency response to a smallpox attack: the case for mass vaccination." *Proceedings of the National Academy of Sciences of the United States of America*, no. 16, pp. 10935–40, Aug.

- [5] L. M. Wein, D. L. Craft, and E. H. Kaplan, "Emergency response to an anthrax attack." *Proceedings of the National Academy of Sciences of the United States of America*, no. 7, pp. 4346–51, Apr.
- [6] P. Baccam and M. Boechler, "Public health response to an anthrax attack: an evaluation of vaccination policy options." *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, vol. 5, no. 1, pp. 26–34, Mar. 2007. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/17437349>
- [7] J. W. Herrmann, S. Lu, and K. Schalliol, "A Routing and Scheduling Approach for Planning Medication Distribution," in *Proceedings of the 2009 Industrial Engineering Research Conference*, 2009.
- [8] Y. M. Lee, "Analyzing Dispensing Plan for Emergency Medical Supplies in the Event of Bioterrorism," in *Proceedings of the 2008 Winter Simulation Conference*, 2008, pp. 2600–2608.
- [9] L. D. Rotz and J. M. Hughes, "Advances in detecting and responding to threats from bioterrorism and emerging infectious disease." *Nature medicine*, vol. 10, no. 12 Suppl, pp. S130–6, Dec. 2004. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15577931>
- [10] U.S. Department of Health & Human Services, "FY 2005 Performance and Accountability Report," 2005. [Online]. Available: <http://www.hhs.gov/of/library/par05/pdfmenu/>
- [11] H. H. Willis, C. Nelson, S. R. Shelton, A. M. Parker, J. A. Zambrano, E. W. Chan, J. Wasserman, and B. A. Jackson, "Initial Evaluation of the Cities Readiness Initiative," Santa Monica, CA, 2009. [Online]. Available: <http://www.rand.org>
- [12] E. K. Lee, C.-H. Chen, F. Pietz, and B. Benecke, "Modeling and Optimizing the Public-Health Infrastructure for Emergency Response." *Interfaces*, vol. 39, no. 5, pp. 476–490, Oct. 2009. [Online]. Available: <http://interfaces.journal.informs.org/cgi/doi/10.1287/inte.1090.0463>
- [13] M. Giovachino, T. Calhoun, N. Carey, B. Coleman, G. Gonzalez, B. Hardeman, and B. McCue, "Optimizing a District of Columbia Strategic National Stockpile Dispensing Center," *J Public Health Manag Pract*, vol. 11, no. 4, pp. 282–90, 2005.
- [14] World Health Organization, "The World Health Report 2007: A Safer Future - Global Public Health in the 21st Century," Geneva, Switzerland, 2007. [Online]. Available: http://www.who.int/whr/2007/whr07_en.pdf
- [15] H. Chen, D. Zeng, and P. Yan, "Syndromic Surveillance Data Sources and Collection Strategies," *Infectious Disease Informatics*, vol. 21, no. 1, pp. 33–48, 2010.
- [16] E. Krenzelok, E. MacPherson, and R. Mrvos, "Disease surveillance and nonprescription medication sales can predict increases in poison exposure." *Journal of medical toxicology*, vol. 4, no. 1, pp. 7–10, Mar. 2008. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/18338303>
- [17] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data." *Nature*, vol. 457, pp. 1012–4, Feb. 2009. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19020500>
- [18] A. Hulth, G. Rydevik, and A. Linde, "Web queries as a source for syndromic surveillance." *PloS one*, no. 2, Jan.
- [19] C. D. Corley, A. R. Mikler, K. P. Singh, D. J. Cook, F. Worth, and C. Science, "Monitoring Influenza Trends through Mining Social Media," in *Proceedings of the 2009 International Conference on Bioinformatics and Bioengineering (BIOCOMP09)*, Las Vegas, NV, 2009.
- [20] M. L. Brandeau, J. H. McCoy, N. Hupert, J.-E. Holty, and D. M. Bravata, "Recommendations for modeling disaster responses in public health and medicine: a position paper of the society for medical decision making." *Medical decision making : an international journal of the Society for Medical Decision Making*, vol. 29, no. 4, pp. 438–60, 2009. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/19605887>

Computer Modeling of Diffuse Axonal Injury Mechanisms

Igor Szczyrba¹, Martin Burtcher², and Rafał Szczyrba³

¹School of Mathematical Sciences, University of Northern Colorado, Greeley, CO 80639, U.S.A.

²Department of Computer Science, Texas State University, San Marcos, TX 78666, U.S.A.

³Funiosoft, LLC, Silverthorne, CO 80498, U.S.A.

Abstract—We investigate numerically which properties of the human brain cause Diffuse Axonal Injuries (DAI) to appear in a scattered and pointwise manner near the gray/white matter boundary, mostly in the white matter. These simulations are based on our dually-nonlinear, viscoelastic, fluid Traumatic Brain Injury model, which includes a nonlinear stress/strain relation. We simulate rotational accelerations and decelerations of a human head that replicate realistic traumatic scenarios. The rotational loads are quantified by our Brain Injury Criterion, which extends the translational Head Injury Criterion to arbitrary head motions. Our simulations show that: (i) DAI occurrences near the gray/white matter boundary can be explained by the difference in the gray and the white matter's shear modulus values, (ii) the scattered/pointwise DAI character can be attributed to the nonlinear fluid aspect of the brain tissue, and (iii) the scattering of DAI deeper in the white matter appears to be caused by the complicated shape of the brain. Our results also show that the nonlinear stress/strain relation plays a secondary role in shaping basic DAI features.

Keywords: computer modeling, diffuse brain injury, nonuniform shear modulus, nonlinearity

1. Introduction

The most 'mysterious' kind of Traumatic Brain Injuries (TBI) are Diffuse Axonal Injuries (DAI). DAI predominantly appear during abrupt head rotations [1], [2]. However, despite many experimental and numerical studies, the way DAI are created in the brain matter is still not well understood. In particular, the following main characteristics of DAI require explanation [3]:

- The injuries are highly localized, i.e., some neurons are affected while their close neighbors are not.
- The injuries are randomly scattered, mostly in the white matter along its boundary with the gray matter.

In his initial studies with a nonlinear fluid TBI model, one of the co-authors investigated implications of the difference in the shear moduli between the gray matter and the white matter on the propagation of shear waves in human brain tissue. The results of a simulated *idealized instant* motion

of two-layer brain tissue indicated that the different shear moduli could explain some features of DAI [4]. More recent studies have shown that the nonlinear stress/strain relation in brain tissue should also be taken into account when modeling scenarios leading to brain trauma [5].

In this paper, we present results of a systematic study of possible mechanism of DAI. The computer simulations are based on our new viscoelastic dually-nonlinear TBI model that includes a nonlinear fluid term as well as a nonlinear stress/strain relation derived from experimental data. Our new model uses a brain facsimile that reflects the *realistic general shape* of a human brain. The gray matter and the meninges are represented as thin layers that follow the skull's shape. We focus on simulating rotational accelerations and decelerations of a human head that recreate *realistic* dynamic conditions leading to severe brain trauma, e.g., a forceful helmet-to-helmet hit during a football game.

2. Dually-nonlinear TBI model

Our computational TBI model is rooted in the biophysical approach that describes the brain dynamics based on the viscoelasticity theory—the brain is injured when the strain field, created in the brain by shear waves due to the head motion, assumes sufficiently high values. To model the dynamic evolution of this strain field, we use the following system of nonlinear Partial Differential Equations (PDEs):

$$\frac{D\mathbf{v}}{Dt} = -\nabla\tilde{p} + \Delta(s^2\mathbf{u} + \nu\mathbf{v}), \quad \frac{D\mathbf{u}}{Dt} = \mathbf{v}, \quad \nabla\cdot\mathbf{v} = 0. \quad (1)$$

Here, $D/Dt \equiv \partial/\partial t + (\mathbf{v}\cdot\nabla)$ is the nonlinear Lie (material) derivative, where $\mathbf{v}(\mathbf{x}, t) \equiv (v_1(\mathbf{x}, t), v_2(\mathbf{x}, t), v_3(\mathbf{x}, t))$ with $\mathbf{x} \equiv (x_1, x_2, x_3)$ denotes the brain matter velocity vector field evaluated at time t in an external coordinate system; $\mathbf{u}(\mathbf{x}, t)$ is the corresponding displacement vector field; $\tilde{p}(\mathbf{x}, t)$ denotes the generalized pressure term consisting of the density normalized pressure and the hydrostatic compression term; $s(\mathbf{x}, t)$ describes the brain's shear wave phase velocity; and ν is the brain's kinematic viscosity.

PDE system (1) generalizes the linear solid Kelvin-Voigt (K-V) model (successfully used to develop a DAI criterion [6]) by introducing *two* nonlinear terms $s(\mathbf{x}, t)$ and $\mathbf{v}\cdot\nabla$, and the term $\tilde{p}(\mathbf{x}, t)$ that is necessary in such a case cf. [4].

The material derivative allows us to model the nonlinear fluid (gel-like) aspect of the brain tissue, whereas $s(\mathbf{x}, t)$ describes how the brain matter stiffens under larger deformations, i.e., how the shear wave velocity increases with the strain. Experiments imply that this relation is linear only for small strains [5], [7] and that it can be approximated by an exponential function for larger strains [8].

Thus, we model the stress/strain relation by $s(\mathbf{x}, t) \equiv c(\mathbf{x}) \exp(qP(\mathbf{x}, t))$, where $c(\mathbf{x}) \equiv \sqrt{G(\mathbf{x})/\delta(\mathbf{x})}$ denotes the basic shear wave velocity in the absence of strain ($G(\mathbf{x})$ and $\delta(\mathbf{x})$ are the brain matter shear modulus and density, respectively), and $P(\mathbf{x}, t)$ describes the time evolution of the spatial distribution of the maximum strain. For strains larger than 50%, we assume that $s(\mathbf{x}, t)$ smoothly becomes proportional to the basic shear wave velocity $c(\mathbf{x})$.

Experiments, cf. [5], [8]-[10], imply that:

- the basic wave velocity in the white matter is $c_w \approx 1\text{m/s}$ and c_g in the gray matter is up to 4 times larger,
- the coefficient q determining the stress/strain relation is within the range $0.4 \leq q \leq 2.5$, and
- the brain's viscosity ν equals approximately $0.013\text{m}^2/\text{s}$.

3. Simulation setup and display method

We simulate sideways head rotations about a fixed vertical axis through the brain's center of mass and forward or backward head rotations about horizontal axes located at the brain's center of mass, the neck, and the abdomen. Keeping the axes fixed allows us to solve PDEs in separate horizontal or sagittal 2D brain cross sections, which simplifies the analysis and presentation of the results.

We show the effects of head rotations in a form of time snapshots presenting (in horizontal and sagittal brain cross sections) the distribution of:

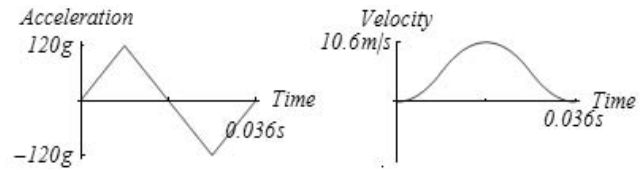
- the vector field $\mathbf{V}(\mathbf{x}, t)$ describing the brain matter velocity relative to the skull,
- this relative velocity's magnitude $|\mathbf{V}(\mathbf{x}, t)|$,
- and the values $P(\mathbf{x}, t)$ of the maximum strain in the white and the gray matter as well as in the meninges.

To better present the character of the brain matter oscillations, we depict the vector field \mathbf{V} in form of curved vectors [11]. The dark to light shading of the curved vectors indicates the motion's direction. Animated 'movies' built from the snapshots of various head rotations are available at our website: <http://www.funiosoft.com/brain/>.

The average (around the skull's perimeter) tangential acceleration loads we apply are quantified by the value of our universal Brain Injury Criterion BIC_{1000T} , where T is the load's duration [12]. It means that the average power per unit mass transmitted from the skull to the vicinity of the considered 2D brain cross section is equal to the average power transmitted to this vicinity under the translational

load corresponding to the Head Injury Criterion HIC_{1000T} successfully used by the automotive industry to determine critical loads [13], [14].

The results presented are obtained using the following triangularly shaped acceleration/deceleration load characterized by the critical value $BIC_{36}=1000$:



Under this tangential load, the sideways rotations of about 110° replicate, e.g., a blow to a boxer's head, whereas similar forward or backward rotations simulate a head motion, e.g., during a car accident.

4. The role of a nonuniform shear modulus and brain geometry

We have previously shown that the brain's geometry influences the character of traumatic brain oscillations [11], [15]. To separate the role played by the brain geometry in shaping DAI features from the role of the difference in the gray and white matter shear moduli and the role of the brain's nonlinear properties, we first simulate rotations of the brain with a uniform or nonuniform shear modulus using the linear K-V TBI model.

Fig. 1 (resp. 2) shows the velocity and the maximum strain distributions at time $t=0.025\text{s}$ in a horizontal brain cross section (separated by the falx cerebri) with a uniform (resp. nonuniform) shear modulus during a counter-clockwise sideways rotation of the head.

In a case of a uniform shear modulus with $c_g = c_w = 1\text{m/s}$, the velocity magnitude $|\mathbf{V}|$ is distributed quite smoothly with $|\mathbf{V}|_{max} \approx 0.6\text{m/s}$, Fig. 1 left panel, even where the skull's shape creates (at the top and bottom of the cross section) secondary vortices with 'opposite' oscillations than those appearing in the major two vortices, Fig. 1 middle panel. Consequently, high strain magnitudes appear only in the meninges, where the transfer of energy between the skull and the brain takes place, Fig. 1 right panel.

In a case of a nonuniform shear modulus with $c_g = 1.75\text{m/s}$ and $c_w = 1\text{m/s}$, the gray matter tends to oscillate along the skull and the falx cerebri in the opposite direction than the white matter, Fig. 2 middle panel. This leads to very steep changes in magnitudes $|\mathbf{V}|$ at the gray/white matter boundary, Fig. 2 left panel, and hence to high strain values there, Fig. 2 right panel. The largest strain values exceed 30%, which suffices to severely damage neurons [6], [16]-[18], most likely due to a chemical imbalance [19], [20].

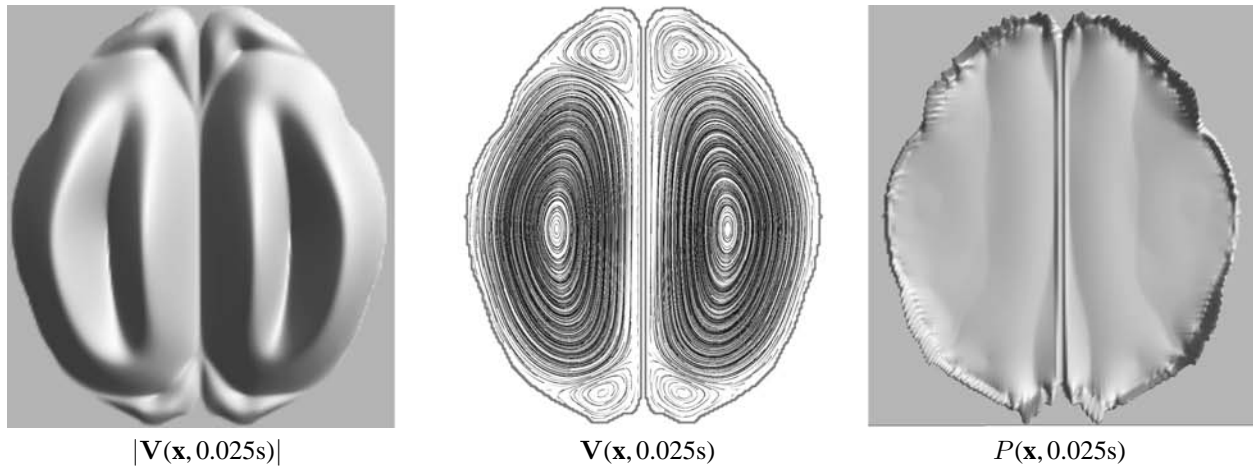


Fig. 1

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A HORIZONTAL CROSS SECTION DURING SIDEWAYS ROTATION ABOUT THE CENTER OF MASS; LINEAR KELVIN-VOIGT MODEL; UNIFORM SHEAR MODULUS: $c_g = c_w = 1 \text{ M/S}$.

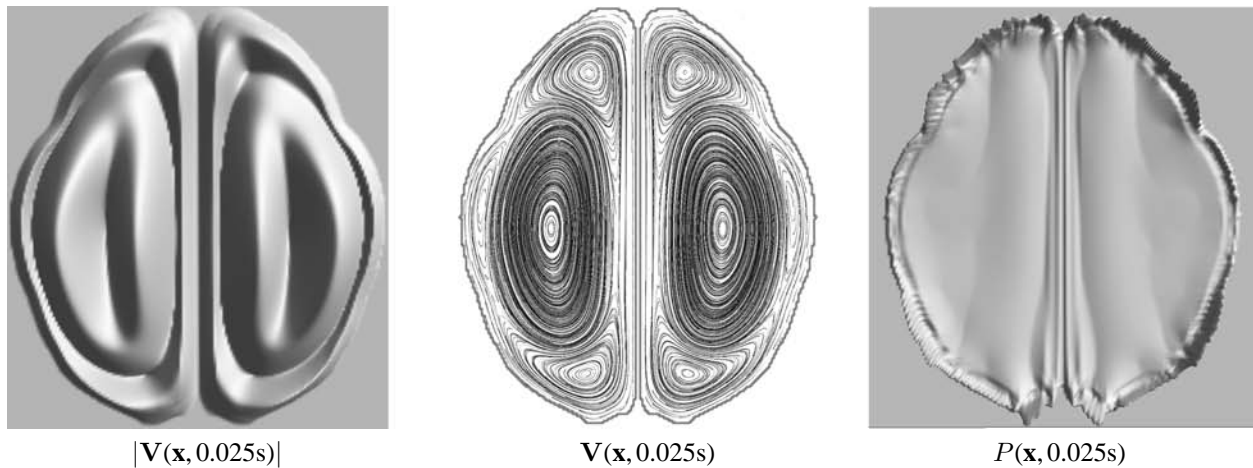


Fig. 2

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A HORIZONTAL CROSS SECTION DURING SIDEWAYS ROTATION ABOUT THE CENTER OF MASS; LINEAR KELVIN-VOIGT MODEL; NONUNIFORM SHEAR MODULUS: $c_g = 1.75 \text{ M/S}$, $c_w = 1 \text{ M/S}$. NOTE THE HIGH VALUES OF $|\mathbf{V}|$ AT THE GRAY/WHITE MATTER BOUNDARY IN THE LEFT PANEL, WHICH ARE THE RESULT OF THE 'OPPOSITE' OSCILLATIONS OF THE GRAY MATTER ALONG THE SKULL AND THE FALX CEREBRI WHEN $c_g > c_w$, MIDDLE PANEL. CONSEQUENTLY, HIGH STRAIN MAGNITUDES APPEAR ALONG THIS BOUNDARY, RIGHT PANEL, WHICH ARE NOT PRESENT IN FIG. 1.

Our simulation results of forward and backward head rotations further show that the brain's shape plays a major role in the localization of oscillatory vortices within the gray and the white matter.

Fig. 3 (resp. 4) on the next page depicts the relative velocity and the maximum strain distributions predicted by the linear K-V model in a sagittal cross section with a uniform (resp. nonuniform) shear modulus when the head is rotated forward about the neck.

In both cases, the shape and the position of the major oscillatory vortex reflects the general semi-circular shape of the upper part of the brain and the fact that the rotational axis is substantially lower than the brain's center of mass, Figs. 3 and 4 middle panels.

A head rotation about an axis located at the abdomen (not shown here) shifts the major vortex towards the top of the brain whereas a head rotation about the brain's center of mass pushes the position of the major vortex down.

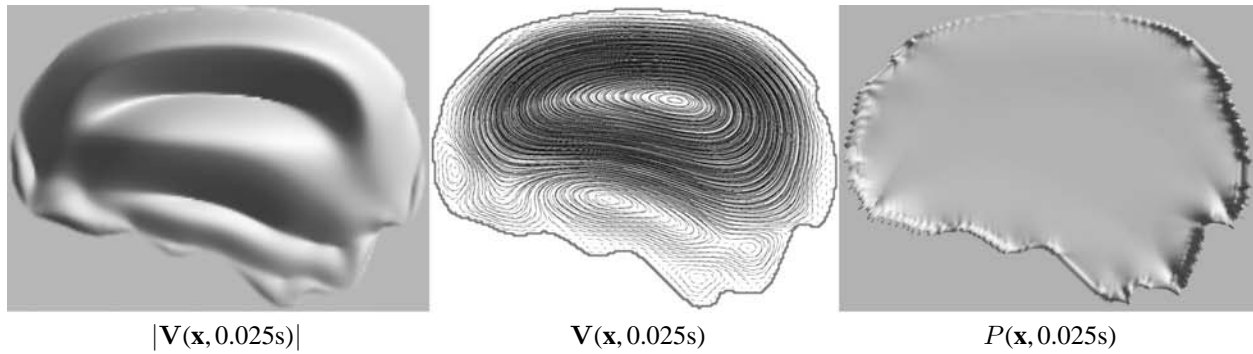


Fig. 3

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A SAGITTAL CROSS SECTION DURING FORWARD ROTATION ABOUT THE NECK; LINEAR KELVIN-VOIGT MODEL; UNIFORM SHEAR MODULUS: $c_g = c_w = 1\text{M/S}$.

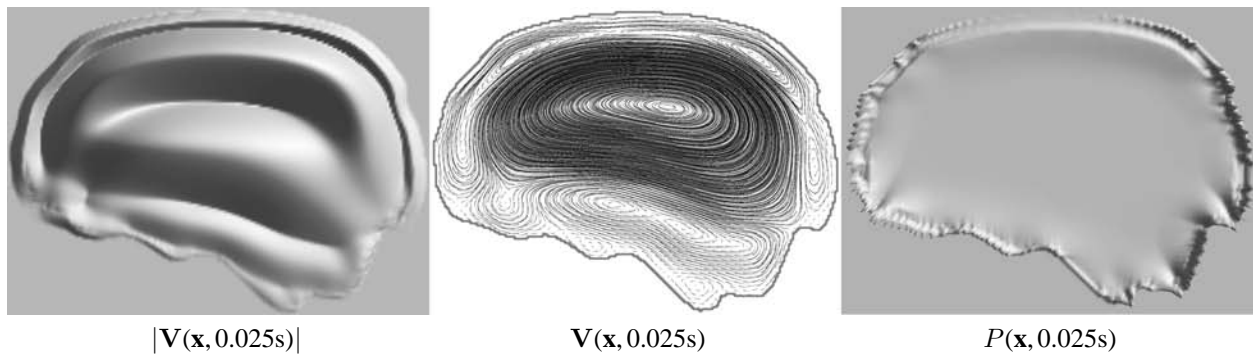


Fig. 4

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A SAGITTAL CROSS SECTION DURING FORWARD ROTATION ABOUT THE NECK; LINEAR KELVIN-VOIGT MODEL; NONUNIFORM SHEAR MODULUS: $c_g = 1.75\text{M/S}$, $c_w = 1\text{M/S}$. NOTE THE HIGH VALUES OF $|\mathbf{V}|$ AT THE GRAY/WHITE MATTER BOUNDARY IN THE LEFT PANEL, WHICH ARE THE RESULT OF THE 'OPPOSITE' OSCILLATIONS OF THE GRAY MATTER ALONG THE SKULL WHEN $c_g > c_w$, MIDDLE PANEL. CONSEQUENTLY, HIGH STRAIN MAGNITUDES APPEAR ALONG THE GRAY/WHITE MATTER BOUNDARY, RIGHT PANEL, WHICH ARE NOT PRESENT IN FIG. 3.

The secondary oscillatory vortices at the bottom of the sagittal cross section, Figs. 3 and 4 middle panels, appear regardless of whether the head is rotated about an axis located at the brain's center of mass, the neck, or the abdomen, i.e., they are created mainly due to the brain's geometry. The specific character of these oscillations changes essentially when the head is rotated backwards, which again highlights the role of the brain's geometry in the distribution of the strain values.

Similar to what we observed in sideways head rotations, in forward head rotations under the linear K-V model neither the major nor the secondary oscillatory vortices create very steep changes in the values of $|\mathbf{V}|$ in the brain interior and consequently they do not lead to high strain values there, Figs. 3 and 4 left and right panels.

When forward or backward head rotations are simulated assuming a nonuniform shear modulus, the results near the gray/white matter boundary are also similar to those obtained during sideways head rotations—the gray matter tends to oscillate in the opposite direction than the white matter, Fig. 4 middle panel. Hence, very steep changes in the velocity magnitudes are created near the gray/white matter boundary, Fig. 4 left panel, that result in high strain magnitudes there, Fig. 4 right panel.

Although, according to the K-V model, the brain geometry substantially influences the character of the brain oscillations, it does not change the maximum velocity magnitude $|\mathbf{V}|_{max}$ and the largest maximum strain values, which are very similar during sideways, forward and backward rotations under the same load.

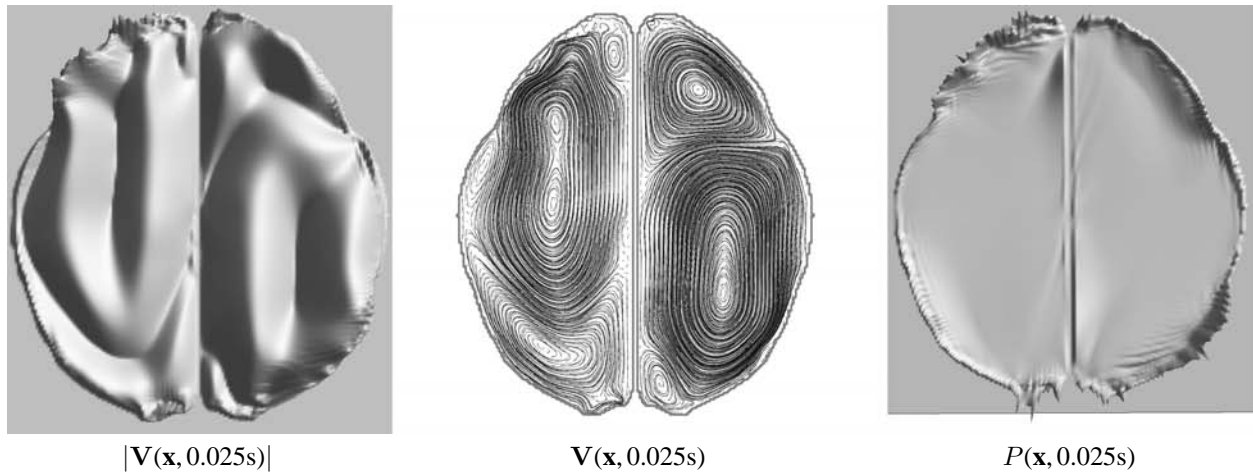


Fig. 5

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A HORIZONTAL CROSS SECTION DURING SIDWAYS ROTATION ABOUT THE CENTER OF MASS; NONLINEAR FLUID MODEL; UNIFORM SHEAR MODULUS: $c_g = c_w = 1\text{ M/S}$. NOTE THAT THE ASYMMETRIC OSCILLATIONS, MIDDLE PANEL, LEAD TO AN ASYMMETRIC SCATTERING OF THE HIGH STRAIN VALUES ALONG THE BRAIN'S PERIMETER, RIGHT PANEL.

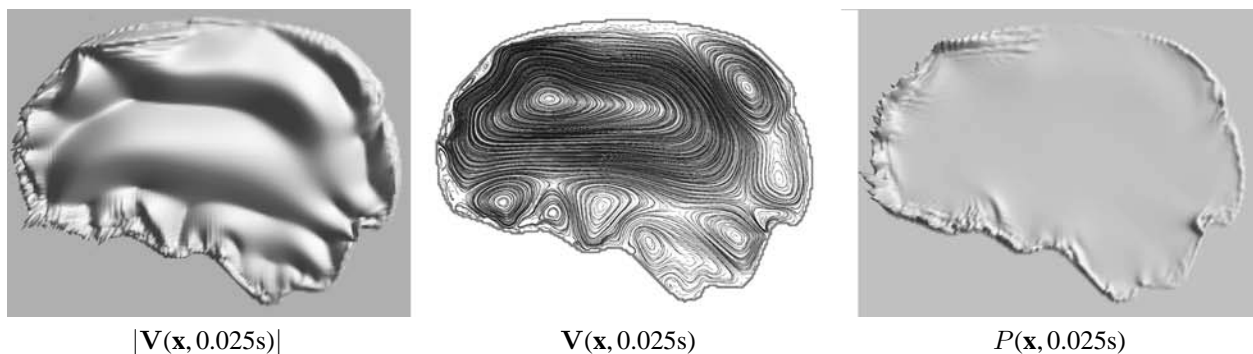


Fig. 6

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A SAGITTAL CROSS SECTION DURING FORWARD ROTATION ABOUT THE NECK; NONLINEAR FLUID MODEL; UNIFORM SHEAR MODULUS: $c_g = c_w = 1\text{ M/S}$. NOTE THE RANDOM SCATTERING OF OSCILLATORY VORTICES, MIDDLE PANEL, AND OF HIGH STRAIN VALUES, RIGHT PANEL, DUE TO THE BRAIN'S GEOMETRY.

5. The role of the brain's fluidity

Replacing the linear temporal derivative in the Kelvin-Voigt model with the nonlinear material derivative allows us to reflect the fluid (gel-like) nature of the brain. This nonlinear fluid (N-F) model predicts more complicated oscillatory patterns than the linear K-V model, even when a uniform shear modulus is assumed, cf. middle panels of Figs. 1 and 5 as well as of Figs. 3 and 6.

In particular, the sideways rotations under the N-F model create asymmetric oscillatory patterns in the brain hemispheres, Fig. 5 middle panel, which is not the case under the K-V model. Thus, the localization of injuries can strongly depend on the rotational direction.

Similarly, the forward head rotations under the N-F model create multiple localized vortices in the back and the bottom of the brain, Fig. 6 middle panel, which are not predicted by the K-V model. The number of these vortices increases when the rotational axis is moved down to the abdomen and decreases when it is moved up to the brain's center of mass.

Moreover, under the N-F model with a uniform shear modulus, the value of $|\mathbf{V}|_{max}$ is up to three times higher than in the K-V model, and steep changes in the velocity magnitudes appear also at the brain's perimeter, Figs. 5 and 6 left panels. This leads to scattered high strain magnitudes near the brain's perimeter, which are not predicted by the K-V model, Figs. 5 and 6 right panels.

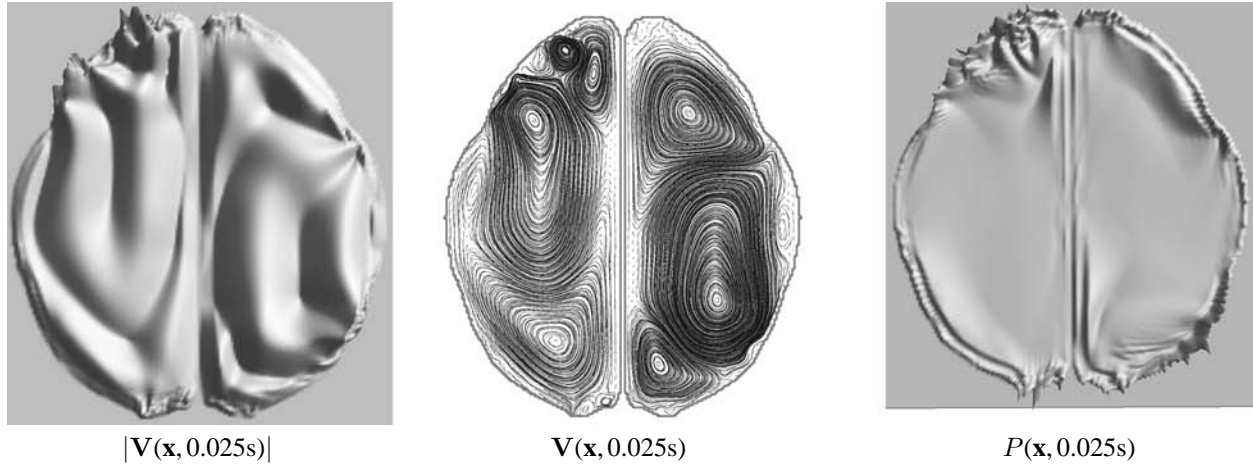


Fig. 7

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A HORIZONTAL CROSS SECTION DURING SIDWAYS ROTATION ABOUT THE CENTER OF MASS; NONLINEAR FLUID MODEL; NONUNIFORM SHEAR MODULUS: $c_g = 1.75\text{M/S}$, $c_w = 1\text{M/S}$.

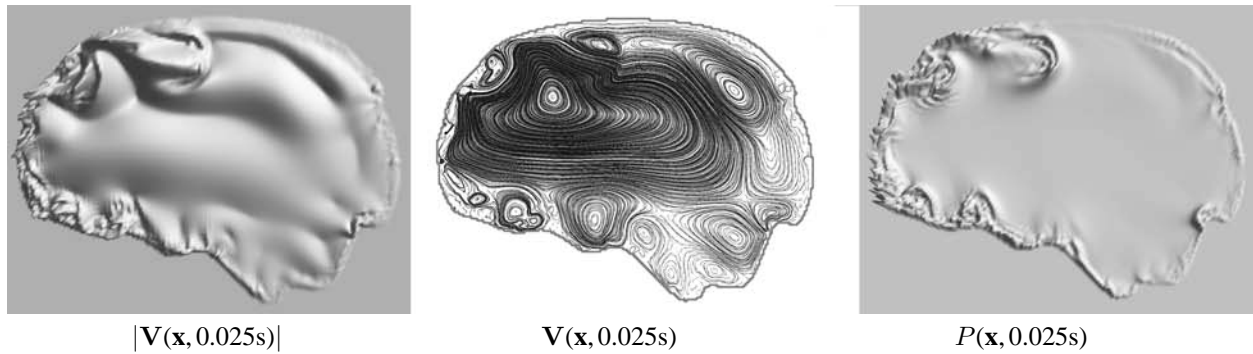


Fig. 8

RELATIVE VELOCITY AND MAXIMUM STRAIN IN A SAGITTAL CROSS SECTION DURING FORWARD ROTATION ABOUT THE NECK; NONLINEAR FLUID MODEL; NONUNIFORM SHEAR MODULUS: $c_g = 1.75\text{M/S}$, $c_w = 1\text{M/S}$.

The introduction of a nonuniform shear modulus into our N-F model allows us to satisfactorily explain why Diffuse Axonal Injuries are highly localized and randomly scattered, mostly in the white matter along the boundary with the gray matter. Indeed, introducing a nonuniform shear modulus results in multiple oscillatory vortices that:

- are characterized by 1/3 higher values of the maximum velocity magnitudes $|\mathbf{V}|_{max}$ than in the case of a uniform shear modulus,
- create steep changes in $|\mathbf{V}|$ along the gray/white matter boundary as well as deeper in some regions of the white matter near this boundary, Figs. 7 and 8 left panels,
- are quite randomly scattered along the boundary between the gray and the white matter, Figs. 7 and 8 middle panels, and

- lead to localized very high strain magnitudes P that are also quite randomly scattered near the gray/white matter boundary as well as deeper inside the white matter, Figs. 7 and 8 right panels.

According to both the K-V and N-F models, the localization of high strain values depends essentially on whether the head is rotated forward or sideways. This outcome is consistent with results obtained by means of one of the most advanced finite element brain injury simulators SIMon [21].

However, the results of our simulations also imply that a specific type of traumatic head motion strongly influences the localization of high strain values. Thus, DAI localization can be quite different when the head is rotated forward or backward, about the brain's center of mass, the neck, or the abdomen, and counter-clockwise or clockwise.

6. The role of a nonlinear stress/strain relation

We have shown in our previous studies that including a nonlinear stress/strain relation with a high value of the parameter q into the K-V model with a uniform shear modulus has the following consequences [15]:

- during head rotations, it reduces strain magnitudes, especially near the skull, and
- after the forcing stops, it creates relatively higher strain magnitudes scattered within the white matter.

Our new simulations lead to similar results under the dually-nonlinear fluid (D-N-F) model with a nonlinear stress/strain relation and both uniform and nonuniform shear moduli. However, the increased strain magnitudes within the white matter due to the nonlinear stress/strain relation are smaller than the critical strain magnitudes appearing due to the nonuniform shear modulus and the brain geometry.

In fact, under the D-N-F model, a nonlinear stress/strain relation only slightly changes the spatial distribution of critical strain magnitudes appearing during head rotations and moderately increases the scattering of high strain magnitudes after the forcing stops. Thus, the nonlinear stress/strain relation seems to play a secondary role in shaping DAI features.

7. Conclusions

Simulations based on our dually nonlinear Traumatic Brain Injury model show that:

- the difference between the values of shear moduli in the gray and in the white matter can explain why Diffuse Axonal Injuries are primarily localized at the gray/white matter boundary,
- the nonlinear gel-like nature of the brain matter together with the complicated shape of the brain can explain the scattered random distribution and pointwise character of DAI, and
- the brain matter's nonlinear relation between stress and strain and the specific position of a fixed rotational axis influence DAI localization and may enhance the random scattered nature of neuronal injuries.

Because the brain's *general* shape and its fluidity already 'scatter' high strain values, one can expect the *convoluted* folding of the brain to cause further scattering of the localized high strain magnitudes along the gray/white matter boundary.

Moreover, since the position of the fixed rotational axis and the rotational direction significantly influence the localization of potential injury points, it is likely that a complicated head rotation about a *varying* axis will further 'randomize' the distribution of axonal injuries.

References

- [1] J. Meythaler *et al.*, "Amantadine to Improve Neurorecovery in Traumatic Brain Injury-associated Diffuse Axonal Injury," *J. of Head Trauma and Rehabilitation*, vol. 17(4), pp. 303-313, 2002.
- [2] J. H. McElhaney, "John Paul Stapp Memorial Lecture: In Search of Head Injury Criteria," *Stapp Car Crash J.*, vol. 49, pp. v-xvi, 2005.
- [3] W. Maxwell, Povlishock J., and Graham D., "A Mechanistic Analysis of Nondisruptive Axonal Injury: A Review," *J. Neurotrauma*, vol. 14, pp. 419-440, 1997.
- [4] C. S. Cotter, P. K. Smolarkiewicz and I. N. Szczyrba, "A Viscoelastic Fluid Model for Brain Injuries," *Int. J. for Numerical Methods in Fluids*, vol. 40, pp. 303-311, 2002.
- [5] E. G. Takhounts, J. R. Crandall and K. Darvish, "On The Importance of Nonlinearity of Brain Tissue Under Large Deformations," *Stapp Car Crash J.*, vol. 47, pp. 79-92, 2003.
- [6] S. S. Margulies and L. Thibault, "A Proposed Tolerance Criterion for Diffuse Axonal Injury in Man," *J. Biomech.*, vol. 25, pp. 917-923, 1992.
- [7] S. Mehdizadeh *et al.*, "Comparison between Brain Tissue Gray and White Matters in Tension Including Necking Phenomenon," *Am. J. Appl. Sc.*, vol. 12, pp. 1701-1706, 2008.
- [8] B. R. Donnelly and J. Medige, "Shear Properties of Human Brain Tissue," *J. Biomech. Engineering*, vol. 119, pp. 423-432, 1997.
- [9] Y. Tada and T. Nagashima, "Modeling and Simulation of Brain Lesions by the Finite-Element Method," *IEEE Eng. Med. Biol. Mag.*, pp. 497-503, 1994.
- [10] K. Paulsen *et al.*, "A Computational Model for Tracking Subsurface Tissue Deformation," *IEEE Trans. Biomech. Eng.*, vol. 46, pp. 213-225, 1999.
- [11] M. Burtcher and I. Szczyrba, "Computational Simulation and Visualization of Traumatic Brain Injuries," in *Proc. 2006 Conf. Modeling, Simulation and Visualization Methods*, pp. 101-107, CSREA Press 2006.
- [12] I. Szczyrba, M. Burtcher and R. Szczyrba, "A Proposed New Brain Injury Tolerance Criterion, based on the Exchange of Energy Between the Skull and the Brain," in *Proc. ASME 2007 Summer Bioeng. Conf.*, paper SBC 2007-171967.
- [13] M. Kleinberger *et al.*, "Development of Improved Injury Criteria for the Assessment of Advanced Automotive Restraint Systems", NHTSA 1998, <http://www-nrd.nhtsa.dot.gov/pdf/nrd-11/airbags/criteria.pdf>.
- [14] R. Eppinger *et al.*, "Development of Improved Injury Criteria for the Assessment of Advanced Automotive Restraint Systems-II," NHTSA 2000, http://www-nrd.nhtsa.dot.gov/pdf/nrd-11/airbags/finalrule_all.pdf.
- [15] I. Szczyrba, M. Burtcher and R. Szczyrba, "On the Role of a Nonlinear Stress-Strain Relation in Brain Trauma," in *Proc. 2008 Conf. Bioinformatics and Computational Biology*, vol. 1, pp. 265-271, CSREA Press 2008.
- [16] B. Morrison III *et al.*, "A Tissue Tolerance Criterion for Living Brain Developed with *In vitro* Model of Traumatic Mechanical Loading," *Stapp Car Crash J.*, vol. 47, pp. 93-105, 2003.
- [17] B. S. Elkin and B. Morrison III, "Region-Specific Tolerance Criteria for the Living Brain," *Stapp Car Crash J.*, vol. 51, pp. 127-138, 2007.
- [18] Y. Matsui and T. Nishimoto, "Nerve Level Traumatic Brain Injury in *in Vivo/in Vitro* Experiments," *Stapp Car Crash J.*, vol. 54, pp. 197-210, 2010.
- [19] J. M. Spaethling *et al.*, "Calcium-Permeable AMPA Receptors Appear in Cortical Neurons after Traumatic Mechanical Injury and Contribute to Neuronal Fate," *J. Neurotrauma*, vol. 25, pp. 1207-1216, 2008.
- [20] J. M. Hinzman *et al.*, "Diffuse Brain Injury Elevates Tonic Glutamate Levels and Potassium-Evoked Glutamate Release in Discrete Brain Regions at Two Days Post-Injury: An Enzyme-Based Microelectrode Array Study," *J. Neurotrauma*, vol. 27, pp. 889-1899, 2010.
- [21] E. G. Takhounts *et al.*, "Investigation of Traumatic Brain Injuries Using the Next Generation of Simulated Injury Monitor, (SIMon), Finite Element Head Model," *Stapp Car Crash J.*, vol. 52, pp. 1-31, 2008.

A Computational Linguistics Approach to the Identification of Biological Factors that Contribute to the Development and Progression of Lung Cancer

C. M. Frenz^{1*}, C. Luo², E. Urgard² and A. Metspalu²

¹In Silico Biotechnologies, New York, USA

²Department of Biotechnology, University of Tartu, Estonia

*Corresponding Author: chris@insilicobiotechnologies.com

Abstract - *The copious volumes of biomedical literature being generated have created a need for the development of text mining algorithms to identify and extract and pertinent biological information. This pilot study demonstrates a computational linguistics approach to identifying genes, proteins, and other biological factors that are associated with the development and progression of lung cancer.*

Keywords: lung cancer, linguistics, bioinformatics, text mining

1 Introduction

Over the course of the last couple of decades the biomedical sciences have undergone an explosion in the amount of biomedical literature that is published, with indexes of biomedical literature, such as Pubmed, housing over 20 million articles (as of March 2011). While the massive amount of information available in such a large corpus of literature is of clear benefit to researchers, the sheer numbers of documents that can match any given query often makes the task of finding needed pieces of information a difficult one. For this reason, researchers in the biomedical sciences are continually turning toward the development of computational tools to perform data mining tasks, ranging from the identification of genes that play a role in certain biological outcome [1-3] and the identification of mutations [4] to the identification of high quality Web resources [5], and many areas in between.

Current methodologies for text mining the biomedical literature include techniques such as regular expression based pattern matching [6], the development and use of biological concept ontologies [7], and the development of specialized parsers designed to perform Natural Language Processing (NLP) of the biomedical literature [3, 8]. This study seeks to expand upon the current methodological approaches by developing a simplistic yet robust methodology for the identification of genes, proteins, and other biological factors that contribute to a disease of interest

or other biological state. The developed methodology makes use of commonly used computational linguistics techniques, by first requiring the establishment of a corpus of biological literature pertaining to the disease or biological state of interest and then performing a word frequency analysis on the corpus to identify all of the unique words in the corpus. A pruning technique is then applied to the listing of unique words in order to remove all English language and biomedical specific jargon words, leaving a resultant list which primarily contains the names of genes and proteins that contribute to the disease or biological state of interest. The technique was tested via the identification of biological factors that contribute to or are associated with the development and metastasis of lung cancer.

2 Methods

2.1 Establishment of a Lung Cancer Corpus

A corpus of text pertaining to lung cancer was developed via the modification of the PREP.pl perl script [6, 9], which was designed to retrieve Pubmed abstracts and perform regular expression based pattern matching against them. The script was modified to retrieve all Pubmed abstracts pertaining to the keyword query "lung cancer" and save the title and text of each abstract to the corpus. Only the titles and text were added to the corpus since other abstract data such as journal name abbreviations and author names would add additional "words" to the corpus that would not be valid biological factors and would be very difficult to prune out during later processing of the data. Execution of this script resulted in a corpus consisting of 186 MB of text.

2.2 Processing of the Corpus

A word frequency analysis of the corpus was performed that identified every unique word in the system as well as how often each unique word appears in the system. This analysis was performed via a Perl script which identified words as being unique if they were separated by 1 or more non-alphanumeric characters (i.e. \W+). This analysis

resulted in over 177,000 unique words ranging from the most popular word “of”, which occurred 1,338,203 times, to words like “Gp96” (a heat shock protein), which only occurred a single time.

Upon completion of the word identification via the word frequency analysis, the list of unique words was pruned by comparing the identified word to a dictionary of words to remove all words that would not be the names of genes, proteins, or other biological factors that could play a potential role in lung cancer, such as English language words. Initial testing of the technique made use of the words.txt dictionary file that comes standard with any Linux distribution as a basis for identifying English language words, but this led to a high false positive rate, since the dictionary did not contain much of the biomedical jargon and terminology that appears in the biomedical literature, but is not a gene/protein name. Thus, the dictionary was expanded to include such jargon and biomedical terminologies (e.g. “mesenchyme”), in order to better prune the list of biological factor candidates. Other common false positive candidates, include common non-gene/protein name acronyms, such as NSCLC (non-small cell lung carcinoma) or NK (natural killer), cancer drugs being tested within the published literature, such as cisplatin, and tumor cells types, such as A549. Further improvements to the false positive rate are made possible by incorporation of these words into the dictionary as well. Adding in common misspellings and typographical errors would be a way to further prune the list of potential biological factors.

3 Results and Discussion

The newly established methodology does provide a means of successfully identifying genes, proteins, and other biological factors that can contribute to diseases such as lung cancer, as illustrated by the results in Table 1, which demonstrates the top eighteen most frequent biological factors that are associated with the development and progression of lung cancer. Among the listings in Table 1 are many well characterized tumor suppressors (e.g. p53) and oncogenes (e.g. myc), as well as other biological factors crucial to the progression of cancer such as VEGF. In some cases in Table 1, the factors identified may be very broad (e.g. kinase and cyclin), but more specific instances of these categories are usually identified at lower frequencies, such as PKC (occurred 1068 times), a type of kinase identified by the technique.

Table 1: The top 18 most frequently appearing biological factors associated with lung cancer.

Biological Factor	Number of Appearances	Sample Reference
p53	16028	[10]
EGFR	14055	[11]
kinase	10274	[12]
VEGF	6168	[13]
CEA	5800	[14]
IFN	4287	[15]
TGF	4255	[16]
MMP	4002	[17]
RR	3235	[18]
TNF	3182	[19]
COX	2838	[20]
cyclin	2614	[21]
caspase	2498	[22]
NNK	2464	[23]
IGF	2426	[24]
Bcl	2285	[25]
myc	2279	[26]
EGF	2140	[27]

This ability for specificity is also demonstrated when one considers that the technique has the capacity to pick up biological factors that may only have few or even singular occurrences in the corpus, as illustrated in Table 2, which contains a sampling of factor names that only occurred 1-2 times in the corpus. While the entries in Table 2 serve to illustrate the specificity of the technique, however, they do demonstrate one limitation of the current implementation in that the word frequency analysis does not correlate multiple possible spellings of the same name as being identical. For example, even though “IRS1” only occurred twice (as written), there were alternative matches in the corpus such as “IRS-1”. These spelling variants are currently recognized as unique, but future iterations of the word frequency analysis software will be modified to treat them as the same. This spelling issue also applies to ATF6, ELAV3, Dnmt3a, and Gp96 as well. It is notable, however, that relatively few publications currently explore these genes/proteins regardless of spelling. For example, Gp96 is a heat shock protein associated with lung cancer in less than 5 English abstracts [28-30].

Table 2: A sample of biological factor names that appeared as written only 1-2 times.

Biological Factor	Number of Appearances	Sample Reference
IRS1	2	[31]
ELAV3	1	[32]
ATF6	1	[33]
Dnmt3a	1	[34]
Gp96	1	[29]

4 Conclusion

In all, this pilot study demonstrates the potential for using word frequency analysis as a means of identifying the names of genes, proteins, and other biological factors that could play a role in the development of lung cancer or other biological conditions. The utility of the developed methodology, however, is dependent on the use of a robust dictionary of words and terms to be excluded from consideration, although it is hypothesized that the development of such a dictionary is broadly applicable to performing such an analysis across a diversity of biological contexts. It is the contention of the author that the continued expansion of such a dictionary of exclusion terms, would result in a technique that could lead to the rapid identification of candidate genes with a minimal amount of human data curation.

5 References

- [1] J. H. Chiang, H. C. Yu, and H. J. Hsu, "GIS: a biomedical text-mining system for gene information discovery," *Bioinformatics*, vol. 20, pp. 120-1, Jan 1 2004.
- [2] C. M. Frenz and D. A. Frenz, "The application of regular expression-based pattern matching to profiling the developmental factors that contribute to the development of the inner ear," *Adv Exp Med Biol*, vol. 680, pp. 165-71, 2010.
- [3] D. Sahoo, J. Seita, D. Bhattacharya, M. A. Inlay, I. L. Weissman, S. K. Plevritis, and D. L. Dill, "MiDReG: a method of mining developmentally regulated genes using Boolean implications," *Proc Natl Acad Sci U S A*, vol. 107, pp. 5732-7, Mar 30 2011.
- [4] F. Horn, A. L. Lau, and F. E. Cohen, "Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors," *Bioinformatics*, vol. 20, pp. 557-68, Mar 1 2004.
- [5] J. A. Young and C. M. Frenz, "Automated extraction of health resource URLs from biomedical abstracts," in *2010 Long Island Systems Applications and Technology Conference (LISAT) 2010*, pp. 1-3.
- [6] C. M. Frenz, "Deafness mutation mining using regular expression based pattern matching," *BMC Med Inform Decis Mak*, vol. 7, p. 32, 2007.
- [7] H. M. Muller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS Biol*, vol. 2, p. e309, Nov 2004.
- [8] H. Chen and B. M. Sharp, "Content-rich biological network constructed by mining PubMed abstracts," *BMC Bioinformatics*, vol. 5, p. 147, Oct 8 2004.
- [9] C. Frenz, *Pro Perl Parsing*: Apress, 2005.
- [10] X. Wang, D. C. Christiani, J. K. Wiencke, M. Fischbein, X. Xu, T. J. Cheng, E. Mark, J. C. Wain, and K. T. Kelsey, "Mutations in the p53 gene in lung cancer are associated with cigarette smoking and asbestos exposure," *Cancer Epidemiol Biomarkers Prev*, vol. 4, pp. 543-8, Jul-Aug 1995.
- [11] J. G. Paez, P. A. Janne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F. J. Kaye, N. Lindeman, T. J. Boggon, K. Naoki, H. Sasaki, Y. Fujii, M. J. Eck, W. R. Sellers, B. E. Johnson, and M. Meyerson, "EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science*, vol. 304, pp. 1497-500, Jun 4 2004.
- [12] M. G. Kris, R. B. Natale, R. S. Herbst, T. J. Lynch, Jr., D. Prager, C. P. Belani, J. H. Schiller, K. Kelly, H. Spiridonidis, A. Sandler, K. S. Albain, D. Cella, M. K. Wolf, S. D. Averbuch, J. J. Ochs, and A. C. Kay, "Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non-small cell lung cancer: a randomized trial," *JAMA*, vol. 290, pp. 2149-58, Oct 22 2003.
- [13] P. Salven, T. Ruotsalainen, K. Mattson, and H. Joensuu, "High pre-treatment serum level of vascular endothelial growth factor (VEGF) is associated with poor outcome in small-cell lung cancer," *Int J Cancer*, vol. 79, pp. 144-6, Apr 17 1998.
- [14] R. Salgia, D. Harpole, J. E. Herndon, 2nd, E. Pisick, A. Elias, and A. T. Skarin, "Role of serum tumor markers CA 125 and CEA in non-small cell lung cancer," *Anticancer Res*, vol. 21, pp. 1241-6, Mar-Apr 2001.
- [15] K. Mattson, A. Niiranen, T. Ruotsalainen, P. Maasilta, M. Halme, S. Pyrhonen, M. Kajanti, M. Mantyla, K. Tamminen, A. Jekunen, S. Sarna, and K. Cantell, "Interferon maintenance therapy for small cell lung cancer: improvement in long-term

- survival," *J Interferon Cytokine Res*, vol. 17, pp. 103-5, Feb 1997.
- [16] H. J. Baek, S. S. Kim, F. M. da Silva, E. A. Volpe, S. Evans, B. Mishra, L. Mishra, and M. B. Marshall, "Inactivation of TGF-beta signaling in lung cancer results in increased CDK4 activity that can be rescued by ELF," *Biochem Biophys Res Commun*, vol. 346, pp. 1150-7, Aug 11 2006.
- [17] X. Zhang, S. Zhu, G. Luo, L. Zheng, J. Wei, J. Zhu, Q. Mu, and N. Xu, "Expression of MMP-10 in lung cancer," *Anticancer Res*, vol. 27, pp. 2791-5, Jul-Aug 2007.
- [18] H. Uramoto, K. Sugio, T. Oyama, T. Hanagiri, and K. Yasumoto, "P53R2, p53 inducible ribonucleotide reductase gene, correlated with tumor progression of non-small cell lung cancer," *Anticancer Res*, vol. 26, pp. 983-8, Mar-Apr 2006.
- [19] V. Flego, A. Radojicic Badovinac, L. Bulat-Kardum, D. Matanic, M. Crnic-Martinovic, M. Kapovic, and S. Ristic, "Primary lung cancer and TNF-alpha gene polymorphisms: a case-control study in a Croatian population," *Med Sci Monit*, vol. 15, pp. CR361-5, Jul 2009.
- [20] R. E. Harris, J. Beebe-Donk, and G. A. Alshafie, "Reduced risk of human lung cancer by selective cyclooxygenase 2 (COX-2) blockade: results of a case control study," *Int J Biol Sci*, vol. 3, pp. 328-34, 2007.
- [21] O. Gautschi, D. Ratschiller, M. Gugger, D. C. Betticher, and J. Heighway, "Cyclin D1 in non-small cell lung cancer: a key driver of malignant transformation," *Lung Cancer*, vol. 55, pp. 1-14, Jan 2007.
- [22] D. A. Fennell, "Caspase regulation in non-small cell lung cancer and its potential for therapeutic exploitation," *Clin Cancer Res*, vol. 11, pp. 2097-105, Mar 15 2005.
- [23] S. Razani-Boroujerdi and M. L. Sopori, "Early manifestations of NNK-induced lung cancer: role of lung immunity in tumor susceptibility," *Am J Respir Cell Mol Biol*, vol. 36, pp. 13-9, Jan 2007.
- [24] S. J. London, J. M. Yuan, G. S. Travlos, Y. T. Gao, R. E. Wilson, R. K. Ross, and M. C. Yu, "Insulin-like growth factor I, IGF-binding protein 3, and lung cancer risk in a prospective study of men in China," *J Natl Cancer Inst*, vol. 94, pp. 749-54, May 15 2002.
- [25] I. Porebska, E. Wyrodek, M. Kosacka, J. Adamiak, R. Jankowska, and A. Harlozinska-Szmyrka, "Apoptotic markers p53, Bcl-2 and Bax in primary lung cancer," *In Vivo*, vol. 20, pp. 599-604, Sep-Oct 2006.
- [26] J. C. Bergh, "Gene amplification in human lung cancer. The myc family genes and other proto-oncogenes and growth factor genes," *Am Rev Respir Dis*, vol. 142, pp. S20-6, Dec 1990.
- [27] F. Ciardiello and G. Tortora, "Interactions between the epidermal growth factor receptor and type I protein kinase A: biological significance and therapeutic implications," *Clin Cancer Res*, vol. 4, pp. 821-8, Apr 1998.
- [28] T. Kojima, K. Yamazaki, Y. Tamura, S. Ogura, K. Tani, J. Konishi, N. Shinagawa, I. Kinoshita, N. Hizawa, E. Yamaguchi, H. Dosaka-Akita, and M. Nishimura, "Granulocyte-macrophage colony-stimulating factor gene-transduced tumor cells combined with tumor-derived gp96 inhibit tumor growth in mice," *Hum Gene Ther*, vol. 14, pp. 715-28, May 20 2003.
- [29] N. Shinagawa, K. Yamazaki, Y. Tamura, A. Imai, E. Kikuchi, H. Yokouchi, F. Hommura, S. Oizumi, and M. Nishimura, "Immunotherapy with dendritic cells pulsed with tumor-derived gp96 against murine lung cancer is effective through immune response of CD8+ cytotoxic T lymphocytes and natural killer cells," *Cancer Immunol Immunother*, vol. 57, pp. 165-74, Feb 2008.
- [30] S. Singhal, R. Wiewrodt, L. D. Malden, K. M. Amin, K. Matzie, J. Friedberg, J. C. Kucharczuk, L. A. Litzky, S. W. Johnson, L. R. Kaiser, and S. M. Albelda, "Gene expression profiling of malignant mesothelioma," *Clin Cancer Res*, vol. 9, pp. 3080-97, Aug 1 2003.
- [31] Z. Ma, S. L. Gibson, M. A. Byrne, J. Zhang, M. F. White, and L. M. Shaw, "Suppression of insulin receptor substrate 1 (IRS-1) promotes mammary tumor metastasis," *Mol Cell Biol*, vol. 26, pp. 9338-51, Dec 2006.
- [32] V. D'Alessandro, L. A. Muscarella, M. Copetti, L. Zelante, M. Carella, and G. Vendemiale, "Molecular detection of neuron-specific ELAV-like-positive cells in the peripheral blood of patients with small-cell lung cancer," *Cell Oncol*, vol. 30, pp. 291-7, 2008.
- [33] N. Dioufa, E. Kassi, A. G. Papavassiliou, and H. Kiaris, "Atypical induction of the unfolded protein response by mifepristone," *Endocrine*, vol. 38, pp. 167-73, Oct 2011.
- [34] M. Fabbri, R. Garzon, A. Cimmino, Z. Liu, N. Zanesi, E. Callegari, S. Liu, H. Alder, S. Costinean, C. Fernandez-Cymering, S. Volinia, G. Guler, C. D. Morrison, K. K. Chan, G. Marcucci, G. A. Calin, K. Huebner, and C. M. Croce, "MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B," *Proc Natl Acad Sci U S A*, vol. 104, pp. 15805-10, Oct 2 2007.

The living experience of a diabetic adult in India using Fuzzy Relational Maps (FRM)

A.Victor Devadoss¹, V.Susanna Mystica²

¹ Department of Mathematics, Loyola College, Chennai - 600 034, India, hanivictor@gmail.com

² Department Of Bio-Technology, St.Joseph's College Of Engineering, Chennai-119, India, susivictor@gmail.com

Abstract - In this paper we find the relation between the risk factors and the symptoms of diabetes among adults using Fuzzy Relational Maps. Diabetes is a problem with the body's fuel system. It is caused by lack of insulin, a hormone made in the pancreas that is essential for getting energy from food. There are two kinds of diabetes: type 1 and type 2. Type 2 diabetes accounts for 90% of all diabetes cases. In this research paper we examine the adults experiencing diabetes using fuzzy relational maps. We have arrived at interesting conclusions. This paper has four sections. In section one we recall the definition of fuzzy relational maps. Section two is devoted to the description of the problem. Section three is devoted to the adaptation of the fuzzy relation maps to the Diabetic problem. In section four we give the conclusions based on our study.

Keywords: Fuzzy Relational Maps (FRM), Risk factors, Symptoms, Type2 Diabetic, Adult, Urban, rural.

1. Introduction:

The new notion called Fuzzy Relational Maps (FRMs) was introduced by Dr. W.B.Vasantha and Yasmin Sultana in the year 2000. In FRMs we divide the very casual associations into two disjoint units, like for example the relation between a teacher and a student or relation; between an employee and an employer or a relation; between the parent and the child in the case of school dropouts and so on. In these situations we see that we can bring out the casual relations existing between an employee and employer or parent and child and so on. Thus for us to define a FRM we need a domain space and a range space which are disjoint in the sense of concepts. We further assume no intermediate relations exist within the domain and the range space. The number of elements in the range space need not in general be equal to the number of elements in the domain space.

1.1. Fuzzy Relational Maps (FRMs)

In our discussion the elements of the domain space are taken from the real vector space of dimension n and that of the range space are real vectors from the vector space of dimension m (m in general need not be equal to n). We denote by R the set of nodes R_1, \dots, R_m of the range space,

where $R_i = \{(x_1, x_2, \dots, x_m) / x_j = 0 \text{ or } 1\}$ for $i = 1, \dots, m$. If $x_i = 1$ it means that the node R_i is in the ON state and if $x_i = 0$ it means that the node R_i is in the OFF state. Similarly D denotes the nodes D_1, \dots, D_n of the domain space where $D_i = \{(x_1, \dots, x_n) / x_j = 0 \text{ or } 1\}$ for $i = 1, \dots, n$. If $x_i = 1$, it means that the node D_i is in the on state and if $x_i = 0$ it means that the node D_i is in the off state. A FRM is a directed graph or a map from D to R with concepts like policies or events etc. as nodes and causalities as edges. It represents casual relations between spaces D and R . Let D_i and R_j denote the two nodes of an FRM. The directed edge from D to R denotes the causality of D on R , called relations. Every edge in the FRM is weighted with a number in the set $\{0, 1\}$.

Let e_{ij} be the weight of the edge $D_i R_j$, $e_{ij} \in \{0, 1\}$. The weight of the edge $D_i R_j$ is positive if increase in D_i implies increase in R_j or decrease in D_i implies decrease in R_j , i.e. causality of D_i on R_j is 1. If $e_{ij} = 0$ then D_i does not have any effect on R_j . We do not discuss the cases when increase in D_i implies decrease in R_j or decrease in D_i implies increase in R_j . When the nodes of the FRM are fuzzy sets, then they are called fuzzy nodes, FRMs with edge weights $\{0, 1\}$ are called simple FRMs. Let D_1, \dots, D_n be the nodes of the domain space D of an FRM and R_1, \dots, R_m be the nodes of the range space R of an FRM.

Let the matrix E be defined as $E = (e_{ij})$ where $e_{ij} \in \{0, 1\}$; is the weight of the directed edge $D_i R_j$ (or $R_j D_i$), E is called the relational matrix of the FRM. It is pertinent to mention here that unlike the FCMs, the FRMs can be a rectangular matrix; with rows corresponding to the domain space and columns corresponding to the range space. This is one of the marked difference between FRMs and FCMs.

Let D_1, \dots, D_n and R_1, \dots, R_m be the nodes of an FRM. Let $D_i R_j$ (or $R_j D_i$) be the edges of an FRM, $j = 1, \dots, m$, $i = 1, \dots, n$. The edges form a directed cycle if it possesses a directed cycle. An FRM is said to be acycle if it does not possess any directed cycle.

An FRM with cycles is said to have a feed back when there is a feed back in the FRM, i.e. when the casual relations flow through a cycle in a revolutionary manner the FRM is called a dynamical system.

Let $D_i R_j$ (or $R_j D_i$), $1 \leq j \leq m$, $1 \leq i \leq n$. When R_j (or D_i) is switched on and if causality flows through edges of the cycle and if it again causes $R_i(D_j)$, we say that the dynamical system goes round and round. This is true for any node R_i (or D_j) for $1 \leq i \leq m$, (or $1 \leq j \leq n$). The equilibrium state of this dynamical system is called the hidden pattern. If the equilibrium state of the dynamical system is a unique state vector, then it is called a fixed point. Consider an FRM with R_1, \dots, R_m and D_1, \dots, D_n as nodes. For example let us start the dynamical system by switching on R_1 or D_1 . Let us assume that the FRM settles down with R_1 and R_m (or D_1 and D_n) on i.e. the state vector remains as (10...01) in R [or (10...01) in D], this state vector is called the fixed point. If the FRM settles down with a state vector repeating in the form $A_1 \rightarrow A_2 \rightarrow \dots \rightarrow A_i \rightarrow A_1$ or ($B_1 \rightarrow B_2 \rightarrow \dots \rightarrow B_i \rightarrow B_1$) then this equilibrium is called a limit cycle.

Methods of determination of hidden pattern.

Let R_1, \dots, R_m and D_1, \dots, D_n be the nodes of a FRM with feed back. Let E be the $n \times m$ relational matrix. Let us find a hidden pattern when D_1 is switched on i.e. when an input is given as vector $A_1 = (1000\dots 0)$ in D the data should pass through the relational matrix E . This is done by multiplying A_1 with the relational matrix E . Let $A_1 E = (r_1, \dots, r_m)$ after thresholding and updating the resultant vector (say B) belongs to R . Now we pass on B into E^T and obtain BE^T . After thresholding and updating BE^T we see the resultant vector say A_2 belongs to D . This procedure is repeated till we get a limit cycle or a fixed point.

2. Description of the problem

2.1. Introduction

Diabetes is a problem with the body's fuel system. It is caused by lack of insulin, a hormone made in the pancreas (an organ that secretes enzymes needed for digestion) that is essential for getting energy from food. There are two kinds of diabetes:

In type 1 diabetes, which usually starts in children, the body stops making insulin completely.

In type 2 diabetes, also called adult-onset diabetes, the body still making insulin, but cannot use it properly.

Most adults with diabetes have type 2; in fact, type 2 diabetes accounts for 90% of all diabetes cases.

2.2. Facts about Diabetes in Adults

- Diabetes is not contagious disease.

- Diabetes has a genetic component and is greatly influenced by environmental factors related to Lifestyle
- Diabetes contributes to the deaths
- Diabetes often leads to blindness, heart and blood vessel disease, strokes, kidney failure, amputations, and nerve damage.
- Uncontrolled diabetes can complicate pregnancy and put a mother at risk for having a baby with birth defects.
- India has the largest number of people with diabetes, roughly around 35 million. of this approximately 13 million still remain undetected.
- Indians develop diabetes almost one decade earlier than whites. This could be due to the fact that Indians have a low-risk threshold for many of the acquired diabetic factors, like obesity.
- In India, diabetes is more prevalent among males than females (ratio being 1:0.6)
- Amongst diabetics, 4.6% urban and 1.9% of rural population had a direct relation with diabetes.
- Diabetes was twice as frequent amongst vegetarians as non-vegetarians. A higher prevalence of diabetes in urban India
- Expatriate Indians tend to be more overweight, have stronger generic factor, being emigrants, live and marry amongst close relatives and thus have a much higher prevalence of diabetes.
- In India, the average male weight is 55kg and the female, 48.5 kg. Among those detected to be diabetic, 31.5% were overweight. It would seem that leanness does not have negative correlation with diabetes in a country like India.
- The recent World Health Organization report suggests that over 19% of the world's diabetic population currently resides in India. This translates to over 35 million diabetic subjects, and these numbers are projected to increase to nearly 80 million by 2030.
- Obesity raises the risk for diabetes by as much as 93%, and an inactive lifestyle can raise it by as much as 25%.

2.3. Diabetes shifts base in India

Midway through their journey into urbanization, suburbs and small towns are finding themselves in precarious health. Results of a cohort study presented at an international conference recently shows that a higher number of people living in semi-urban areas have diabetes and hypertension when compared to those in cities. Health care experts are concerned that a greater number of people in these areas now run the risk of cardiac arrests, renal failures and strokes. Says Dr. S. Thanikachalam lead investigator of the study and cardiology head at Sri Ramachandra university, who presented the results at an international conference in the city recently: “ We found that nearly 22.2% of people in semi-urban areas have diabetes compared to 17.5% in urban and 14.5% in rural areas. Similarly, the number of people with hypertension was 26.4% in suburban areas compared to 17.3% in urban and 17.9% in the rural population.” The number of people with prediabetic and pre-hypertensive conditions was also found to be higher in semi-urban areas. Here is the logic: Suburbs and small towns have moved away from the routine physical exertions of villagers and neither do they have the awareness and wherewithal for an organized exercise regime like gymnasias.

The study, funded by the department of Science and Technology screened 6,000 people in Chennai, Tiruvallur and Kancheepuram. “We found 43.3% of people with abnormal glucose metabolism, 75.3% with abnormal lipid profiles and 52% with high blood pressure. Though only a person with blood pressure higher than 140/90 is considered hypertensive, people with 135/85 also require intervention. So, at least 50% of our population would require intervention in one form or another,” says Dr. Thanikachalam. State health secretary VK Subburaj says the government is seized of the matter. “We have programmes like door-to-door screening of people. We have been working out new awareness and prevention strategies”, he said.

Another disturbing trend the study revealed was that nearly 80% of the people had shown signs of physiological distress, including anxiety, stress or depression. “It was due to various factors including loss of a family member, financial problems or even other emotional issues. We have adequate studies that prove How lack of good mental health can trigger a series of non-communicable diseases. We think it is necessary to have a series of problems including counseling for such people,” he said. Now ‘rich man’s diseases’ come calling on city slums

Despite the health department’s ambitious project to provide health for all, a study by a city based hospital and research centre shows how poverty has pushed Chennai’s slum dwellers into a series of health problems and chronic disorders. The study has generated interest among healthcare experts particularly because many feel the city’s epidemic pattern of diabetes is beginning to see a change.

A study by MV Hospital for diabetes led by Dr. Vijay Vishwanathan, which screened over 900 people, showed that at least 17.2% of them had respiratory illness and 13.5% had other infections. Anemia was high among women of all age groups and many children were found to be underweight. “In Chennai, more than 25% of the total population are slum dwellers. About 40% of this slum population lives along the rivers and canals and the rest are on the pavements. We saw how slums are largely neglected in terms of provision of healthcare facilities,” Says Dr. Vijay Vishwanathan. His team carried out the study to explore the living conditions and determining the health related problems that affect the underprivileged section of the urban population from all parts of Chennai. The study published in the Indian Journal of Community Medicine got 326 men and 574 women to answer an questionnaire covering socio-demographic details, housing and environmental details, health problems, and behavior, They were then taken to a hospital for clinical examination.

“At least 48% had no access to safe drinking water and 66% had no toilets. About 53% lived in temporary shelter.” Said Shabana Tharkar, Who did the study along with Dr. Vijay. But what makes their condition worse is that in addition to malnutrition and communicable diseases, their modified diet has led to increase in blood sugar. A parallel study by the Madras Diabetes Research Foundation has shown that the incidence of diabetes in Chennai slums has gone up by 134% in the last ten years. The Dr. V. Mohan of Madras Diabetes Research Foundation says the epidemic pattern of lifestyle disorders is beginning to see a change within cities. “I term the causes as influenza or sedentary,” he said. “Diabetes was once called the rich man’s disease. In 10 years, it is likely to become the disease of the poor. And we are seeing differences even within the city. Our study has shown a slowdown in the incidence of diabetes in the middle and upper middle class because they are aware and they can afford exercises. Those in slum dwellers have two-wheelers instead of bicycles. The lack of physical activity and consumption of packaged foods and aerated drinks are showing on their health.” says Dr. Mohan.

2.4. The high risk factors for developing Diabetes

Type 2 diabetes affects all types of people. However, there are factors that can put anyone at higher risk for developing the diabetic are

- Being overweight (body-mass index of 25+)
- Carrying fat around the waist and stomach
- Being sedentary

- Being more than 45 years old (being over 65 increases risk even further)
- Having a family history of type 2 diabetes
- Having gestational diabetes or having a baby that weighed 9 lbs or more
- Being of Indian, or Native Indian descent
- Having a low high-density lipoprotein (HDL) cholesterol level (less than 35)
- Having a high triglyceride level (250 or above)
- Having high blood pressure (140/90 mm/Hg or higher)

Type 2 diabetes used to be quite rare before middle age and people living in the rural areas in India but now affects more and more young people who are overweight. Being overweight, even as a child or teenager is a significant risk factor for developing diabetes as an adult.

2.5. The Symptoms of Diabetes

Diabetes in adults may start slowly. In fact, millions of people don't even know they have it.

They may just feel very tired at first, then later may have these symptoms:

Urinating more than usual, as the body tries to get rid of the extra sugar in the blood, Feeling unusually thirsty, because the body needs to replace the lost fluid, Nausea, Blurred vision, Feeling hungry while losing weight, Frequent infections, Skin sores that won't heal. It's important to remember that diabetes symptoms may not be the same for everyone. The symptoms of type 2 diabetes may come on gradually. Some people may have no symptoms at all. Many people have type 2 diabetes and don't know it. Untreated diabetes can cause serious health problems, such as blindness, heart and blood vessel damage, and permanent nerve damage. In this paper, we give an algebraic approach to the Diabetic problem faced by an adult. This study is significant because most of the adults in India can adopt the same procedure. All South Asians in general and Indians in particular are prone to diabetes. Thus all Indians above the age 25 years ought to be tested for diabetes. By knowing this age group an adult least can take steps to treat himself. This linguistic questionnaire was used to obtain the attributes and using these attributes and the opinion of the experts we have used FRM to analyze the problem.

3. Adoption of FRM model to study about Type2 Diabetic Problem

We have made a sample survey of around 120 people living in Chennai(Patients of M.V. Hospital for Diabetes, Royapettah). They were interviewed using a linguistic questionnaire. The Fuzzy concepts, i.e. attributes are first given in the form of matrix relational equations and then solved. In this paper we use this method to find who the victims of Diabetes are. The following risk factors are the developing condition for diabetes and taken as the attributes of our study

3.1. Attributes Related to the risk factors

The domain space G connected with the risk factors are given by $G = \{G_1, \dots, G_{10}\}$

G_1 : Carrying fat around the waist and stomach

G_2 : Being sedentary

G_3 : Being more than 45 years old

G_4 : Having a family history of type 2 diabetes

G_5 : Having gestational diabetes or having a baby that weighed 9 lbs or more

G_6 : Being of Indian, or Native Indian descent

G_7 : Having a low high-density lipoprotein (HDL) cholesterol level less than 35)

G_8 : Having a high triglyceride level (250 or above)

G_9 : Having high blood pressure (140/90 mm/Hg or higher)

G_{10} : Being overweight (body-mass index of 25+)

3.2. Attributes Related to the Symptoms

The Range space S connected with the symptoms are given by $S = \{S_1, \dots, S_7\}$

S_1 : Frequent urination

S_2 : Excessive thirst

S_3 : Nausea

S_4 : Blurring vision

S_5 : Extreme hunger and losing weight

S₆: Frequent infections

S₇: Skin sores that won't heal

Now using the expert's opinion. We have the following relation matrix by taking Risk factors G₁...G₁₀ as the rows and Symptoms S₁,...,S₇ as the columns.

3.3. First Experts Opinion

The opinion of the first expert is a Diabetic patient from urban and is given vital importance. His opinion is transformed into the Fuzzy Relational matrix P₁ given by

$$P_1 = \begin{matrix} & S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \\ G_6 \\ G_7 \\ G_8 \\ G_9 \\ G_{10} \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

The hidden pattern of the state vector X = (0 0 0 1 0 0 0 0 0 0) is obtained by the following method:

$$\begin{aligned} XP_1 &\hookrightarrow (1\ 1\ 1\ 0\ 1\ 1\ 0) = Y \\ YP_1^T &\hookrightarrow (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1) = X_1 \\ X_1P_1 &\hookrightarrow (1\ 1\ 1\ 1\ 1\ 1\ 1) = Y_1 \end{aligned}$$

(Where \hookrightarrow denotes the resultant vector after thresholding and updating)

When we take G₄ in the ON state (i.e. Having a family history of type 2 diabetes) and all other attributes to be in the off state. We see the effect of X on the dynamical system P₁ is a fixed point given by the binary pair

$$\{(1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1), (1\ 1\ 1\ 1\ 1\ 1\ 1)\}.$$

When we are having a family history of type 2 diabetes node alone in the on state we get say X = (1 1 1 1 1 1 1)

The resultant to be the fixed point given by the binary pair {(1 1 1 1 1 1 1 1 1 1), (1 1 1 1 1 1 1)}.

When the on state is taken as node G₄ we see the hidden pattern is the fixed point which is the same binary pair, which makes all the nodes to be in the on state in the domain space and also makes all the nodes in the range space to be in the on state.

3.4. Second Experts Opinion

The opinion of the second expert who happens to be a Diabetic patient from rural area and his opinion is transformed into the Fuzzy Relational matrix P₂ is given by:

$$P_2 = \begin{matrix} & S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \\ G_6 \\ G_7 \\ G_8 \\ G_9 \\ G_{10} \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

The hidden pattern of the state vector X = (0 0 0 1 0 0 0 0 0 0) is obtained by the following method:

$$\begin{aligned} XP_2 &\hookrightarrow (0\ 0\ 1\ 0\ 0\ 0\ 0) = Y \\ YP_2^T &\hookrightarrow (0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1) = X_1 \\ X_1P_2 &\hookrightarrow (0\ 0\ 1\ 0\ 0\ 0\ 0) = Y_1 \\ Y_1P_2^T &\hookrightarrow (0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1) = X_1 \end{aligned}$$

When we take G₄ in the ON state (i.e. Having a family history of type 2 diabetes) and all other attributes to be in the off state. We see the effect of X on the dynamical system P₁ is a fixed point given by the binary pair {(0 0 0 1 0 0 0 0 0 1), (0 0 1 0 0 0 0)}. Since the working is time consuming, a C program is formulated for finding the hidden pattern.

4. Conclusions and Suggestions

The cause of diabetes continues to be a mystery, although both genetics and environmental factors such as obesity and lack of exercise appear to play roles. The principal reason for

escalating diabetes and regional disparities appears to be rapidly occurring socioeconomic changes and affluence associated with dietary excess and reduced physical activity. Chennai showed a steady increase in the prevalence of diabetes in the urban population. The major observation of the study had been the low amount of physical activity in the urban population in India is the main cause. Increasing urbanisation tends to lead to lower physical activity worldwide. The impact of urbanisation and its influence on life style has been the cause of diabetes.

Early identification of the high risk individuals would help in taking appropriate intervention in the form of dietary changes and increasing physical activity, thus helping to prevent, or at least delay, the onset of diabetes. This means that identification of at risk individuals is extremely important to prevent diabetes in India. The following steps are suggested to prevent diabetes.

- Watch for and treat symptoms of low blood sugar, which may be a medication side effect
- Watch out for early signs of complications such as problems with eyes, feet, skin and kidneys
- It's important to remember that diabetes symptoms may not be the same for everyone
- The symptoms of type 2 diabetes may come on gradually. Some people may have no symptoms at all. Many people have type 2 diabetes and don't know it.
- Untreated diabetes can cause serious health problems, such as blindness, heart and blood vessel damage, and permanent nerve damage.
- Seeing doctor regularly for checkups and a discussion of risk for diabetes is key to staying healthy.
- Eat in a way that keeps blood sugar as steady as possible
- Lose weight if necessary
- Test blood sugar correctly
- Learn to take insulin shots

Start a fitness program

5. References

- [1]. Kosko, Bart (1992). *Neural Networks and Fuzzy Systems*, Prentice – Hall, Englewood Cliffs, New Jersey.
- [2]. Kosko, Bart (!986), Fuzzy Cognitive Maps, *International Journal of Man – Machine Studies*, 34, 65 – 75.
- [3]. Starr M.K (1978) Operations Management, Prentice Hall Inc., USA.
- [4]. M.M.S.AHUJA, Learning to live with DIABETES
- [5]. Tsandiras, A.K.and Mararitis , K.G.(1996). Using Certainty neurons in Fuzzy
- [6]. Shaw, J. E., Sicree, R. A. & Zimmet, P. Z. *Diabetes Res. Clin. Practice* 87, 4–14 (2010).
- [7]. Magliano, D. J. *et al. Diabetes Care* 33, 1983–1989 (2010). Jowett, J. B. *et al. Twin Res. Hum. Genet.* 12, 44– 52 (2009).
- [8]. Ramachandran, A., Ma, R. C. W. & Snehalatha, C. *Lancet* 375, 408–418 (2010). 5. Mohan, V. *et al. Indian J. Med. Res.* 131, 369–372 (2010).
- [9]. Pradeepa, R. *et al. Diabetes Technol. Therapeutics* 12, 755–761 (2010).
- [10]. Dowse, G. K. *et al. Diabetes* 39, 390–396 (1990).
- [11]. Zimmet, P. *IDF Bull.* 36, 29–32 (1996).
- [12]. Mohan, V. *et al. Indian J. Med. Res.* 125, 217–230(2007).
- [12]. Mohan, V. *et al. Diabetes Res. Clin. Practice* 80, 159–168 (2008).
- [13]. Dunstan, D. W. *et al. Circulation* 121, 384–391 (2010).
- [14]. Unnikrishnan, R. *et al. Diabetes Care* 30, 2019–2024 (2007).
- [15]. Sandeep, S., Ganesan, A. & Mohan, V. *Development and Updation of the Diabetes Atlas of India* www.whoindia.org/LinkFiles/NMH_Resources_Diabetes_atlas.pdf (2010).
- [16].Mohan V, Shanthirani CS, Deepa R. Glucose intolerance *J. Assoc.Phys. India.* 2003;51:771–777
- [17]. Mohan V, Deepa M, Deepa R, Shanthirani CS, Farooq S, Ganesan A, et al. CURES-17). *Diabetologia.* 2006;49:1175–1178
- [18]. Mohan V, Shanthirani S, Deepa R, Premalatha G, Sastry NG, Saroja R. (CUPS- 4). *Diabetes Med.* 2001;18:280–287
- [19]. A.Ramachandran, Management of Diabetes an Overview , Hindu 14 Nov 2005
- [20]. K.M.Prasana Kumar, Combating the increasing menace of diabetes in India, Hindu 14 Nov 2005
- [21]. Ramachandran A, Snehalatha C, Latha E, Vijay V, Viswanathan M. *Diabetologia.* 1997;40:232–237
- [22]. Sadikot SM, Nigam A, Das S, Bajaj S, Zargar AH, Prasannakumar KM,et al. (PODIS). *Diabetes Res. Clin. Pract.* 2004;66:301–307
- [23]. Raman Kutty , Soman CR, Joseph A, et al. *Natl. Med. J. India.* 2000;13:287–292
- [24]. Joshi P, Islam S, Pais P, Reddy S, Dorairaj P, Kazmi K, et al. . *JAMA.* 2007;297:286–294
- [25]. Ramachandran A, Snehalatha C, Latha E, Vijay V, Viswanathan M. *Diabetologia.* (2001);44:1094–1101

SleepGaze: A Wireless System for Monitoring and Detection of Sleep Disorders

K.A. Unnikrishna Menon , Davis Jose and Maneesha V. Ramesh

AMRITA Center for Wireless Networks and Applications, Amrita University, Kollam, Kerala, India

Abstract - Sleep disorders are exponentially growing with current statistics as approximately 1 in 6 or 40 million people in USA. This alarming state has to be controlled in its early stage, to achieve physical and mental wellbeing of human beings, contributing to the peace and welfare of whole world. Current sleep monitoring facilities uses dedicated sleep labs at the hospital. However these tests results are error prone since the patient sleep gets disturbed due to the numerous wired sensors attached to their body, new ambience, reduced privacy, and long waiting duration due to the non availability of sleep labs. This research aims to develop a pervasive monitoring system that overcomes these drawbacks and provides the capability to monitor and detect sleep disorders in any place comfortable to the patient such as patients home, hospitals etc thereby collecting the best signals. The real-time data received from the system will be analyzed to detect sleep disorders remotely and issue the alerts to the clinicians. In the first phase of this research, we have designed and implemented the complete system using EMG sensor alone. The initial results are incorporated in this paper.

Keywords: Sleep Monitoring, Polysomnography, Wireless Sensor Networks, Real-time monitoring

1 Introduction

Sleep is a naturally recurring state characterized by reduced or lacking consciousness, relatively suspended sensory activity, and inactivity of nearly all voluntary muscles" [1]. Sleep is important for the restoration and renewal of the body. Inadequate sleep can lead to many disorders like irritability, poor concentration, memory loss, impaired moral judgment, risk of type2 diabetes, decreased reaction time, increase heart rate variability and risk of heart diseases. Sleep disorders actually disturb the sleep cycle and the quality of sleep. According to the statistics of National Heart, Lung, and Blood Institute (NHLBI), 1 out of 6 American are having sleep disorders [2]. Even though a clear statistics about the amount of sleep disorders prevalent in India is unknown, we estimate to have a similar figure equal to that in US. One of the major problems faced in India is the improper treatment & diagnosis available for sleep disorders. This is mainly because of the lack of facilities and the exorbitant cost for the diagnosis and

treatment [3]. The proposed system targets to solve this issue and is aimed at Indian Population.

According to the statistics by World Health Organization [4], it is estimated that 5-10% of the population at any given time is suffering from identifiable depression needing medical attention. By analyzing the sleep pattern, it is possible to detect depression. This can be found by analyzing the time it takes to sleep after going to bed, actual sleep duration, quantitatively measuring whether having deep or shallow sleep, number of awakening during sleep. The proposed system actually can calculate all these parameters to detect depression. Untreated sleep disorders will lead to poor concentration and Excessive daytime sleepiness (EDS). About 22% of the road accidents [5] are caused due to EDS in drivers. Obstructive sleep apnea is also a cause for EDS. Sleep apnea and other sleep disorders can be detected with the proposed system.

The system can be also used for disaster management applications, to monitor the sleep pattern of panic struck population and to provide proper medication to overcome them from the state of trauma.

Polysomnography is used to diagnose sleep disorders like sleep apnea, periodic limb movement disorder (PLMD), Rapid Eye Movement (REM) behavior disorder and narcolepsy. Polysomnography is performed in dedicated sleep labs with all the measuring electrodes positioned on the patient body. With the placement of the electrodes the patient discomfort increases, which in turn affect the sleep leading to the failure of the test. The discomfort is mainly due to the change in the environment and also due to limited mobility since large number of electrodes fixed to the patient body. The proposed system actually takes care to reduce the patient discomfort, since it's a wireless system it overcomes the mobility and environment problems listed above. The system also supports remote monitoring so that the patients can take-up the test from the comfort of their home. The proposed system also helps in data collection from patients to capture the important biomedical signatures before and after an epileptic attack, which can be used for clinical research people to study epileptic attacks in detail.

The remaining portion of the paper is organized as follows, Section II describes the related work. Section III

explains the architecture and design of the proposed wireless remote sleep monitoring system. Section IV outlines the implementation details and Section V concludes the paper and provides the future work.

2 Related Work

In [6], the authors provide a general outline about the measurement of key sleep related biomedical signatures, this can be obtained without wiring or physical contact with the subject. The paper presents a new approach of contact-less measurement of heart rate, physical movement and respiration using Doppler radar mechanism. In the paper the authors illustrate that the Doppler system was able to detect the peaks similar to that of conventional measurements systems. This paper gives a new idea about contact-less sensing of biological signals. The system actually limits the mobility of the patient. The proposed system actually overcomes the mobility limitation.

The paper [7], provides an insight into measuring severity of OSA with the help of a new measure known as the Dynamic Apnea Hypopnea Index time. Normally the severity is measured from the Apnea Hypopnea Index, which is the average of the obstructive sleep events (OAH) during the entire sleep period. According to the authors, the number of OAH events is a random variable with unknown mean and probability distribution. The paper provides details on how to detect apnea from the available data with minimum error. The details regarding what type of physiological signals used to evaluate the algorithm is not specified in the paper. This paper actually helps in the signal analysis part of the proposed system.

In [8], paper describes a system which performs real time monitoring and transmission of physiological data of patients. The data collected from a wireless pulse oximeter is used to detect apnea on a Personal Digital Assistant (PDA) which has a General Packet Radio Service (GPRS)/Universal Mobile Telecommunications System (UMTS) facility. The analysis is based on SpO₂ signals (blood oxygen level). A classifier that runs on the PDA is used for the analysis. The main feature of this classifier is that it works on the limited resources of a PDA. The paper provides more details on the signal processing aspects on how to process the available data to detect apnea from the SpO₂ Signals. The accuracy and reliability is improved in the proposed system by considering multiple parameters for the analysis. The system is limited only for the detection of apnea, but proposed system can be used for the detection of a variety of sleep disorders.

3 Problem Domain : Sleep Disorders and Detection

“Sleep disorders involve any difficulties related to sleeping, including difficulty falling or staying asleep, falling asleep at inappropriate times, excessive total sleep time, or abnormal behaviors associated with sleep”[9]. According to International Classification of Sleep Disorders (ICSD)[10], Sleep disorders are classified into four, Dyssomnias, Parasomnias, Sleep Disorder associated with Medical/Psychiatric disorders, and proposed sleep disorders.

3.1 Dyssomnias

This disorder is characterized by problems in getting sleep or staying asleep or of excessive sleepiness. The three core sub-classification include, Intrinsic sleep disorders, Extrinsic sleep disorders, and Circadian rhythm sleep disorders. Main intrinsic sleep disorders include Psycho-physiological insomnia, Idiopathic insomnia, Narcolepsy, Obstructive Sleep Apnea, Periodic Limb Movement Disorder. Key Extrinsic Sleep Disorders include inadequate sleep hygiene, altitude insomnia, insufficient sleep syndrome, Alcohol-dependent sleep disorder and Sleep Onset association disorder. Few Circadian Rhythm Sleep Disorders include time zone syndrome, shift work sleep syndrome, irregular sleep wake pattern, and non 24-hour sleep-wake disorder.

3.2 Parasomnia

Parasomnias are characterized by undesirable motor, verbal, or experiential phenomenon occurring in association with sleep, specific stages of sleep, or sleep-awake transition phases [11]. Parasomnias are broadly classified into three, Arousal Disorders, Parasomnias associated with REM sleep and other parasomias like Sleep bruxism, Sleep enuresis, Nocturnal paroxysmal dystonia.

3.3 Sleep Disorders associated with Medical/ Psychiatric Disorders

These are classified into Sleep Disorders associated with Mental Disorders, Sleep Disorders associated with Neurological Disorders, and Sleep Disorders associated with other medical disorders like Sleeping sickness, Fibrositis Syndrome.

3.4 Proposed Sleep Disorders

Proposed Sleep Disorders include short sleep, long sleep, subwakefulness syndrome, Sleep hyperhidrosis and Terrifying Hypnagogic Hallucinations. To diagnose and to identify the disorders listed above a detailed multi-parameter sleep study known as Polysomnography (PSG) is carried-out. The test result is known as Polysomnogram which contains a detailed capturing of key biological signals related to brain activity (EEG), Eye Movements (EOG),

Muscle Movements (EMG), cardiac rhythm (ECH), respiration and blood oxygen saturation during sleep. PSG is conducted with an overnight stay in dedicated sleep labs. The main factor that affects the test is the amount of quality sleep the patient gets during the sleep study at sleep labs, which is dependent on the discomfort level the patient faces while placing the electrodes for measuring various parameters which in turn tethers the patients to the bed. Also the new environment of the sleep labs can also affect the quality of the sleep. The proposed system overcomes all the drawbacks with the conventional PSG techniques.

4 Architecture and Design of SLEEPGAZE

The top level overall architecture of the proposed system in depicted in Figure 1. The proposed system has three tier architecture. The base level module has the interface to the patient and the top most module has interface to the clinicians. The modular design and the plug n play features allow the system to be scalable and robust. The proposed system performs the real-time acquisition, wireless transmission and signal analysis & characterization of the signals. The major advantage of the system is that it can detect the sensor failure if the electrode comes out from the patient body and can provide alert to the bystander to fix the sensors properly.

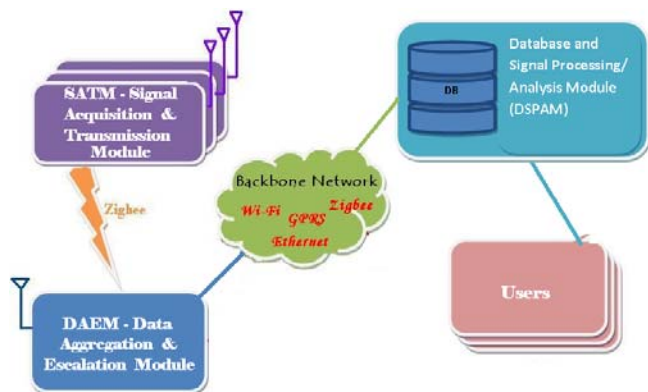


Figure 1. Top Level Architecture

4.1 Signal Acquisition and Transmission Module (SATM)

SATM is the module that has an interface with the patient. This module actually acquires the biomedical signal, performs basic signal conditioning and wirelessly transmits to the base station unit.

The module block diagram is shown in the Figure 2. The electrodes used are Ag/AgCl electrodes. The electrodes pick the bio-potentials and generate a corresponding voltage output. The micro volt level of the electrode output needs to be amplified. The instrumentation amplifier amplifies the

electrode output and generates an output that is sufficient for further signal conditioning. The band pass filter performs the required filtering of the signal and processed signal is sampled and wirelessly transmitted using MicaZ notes.

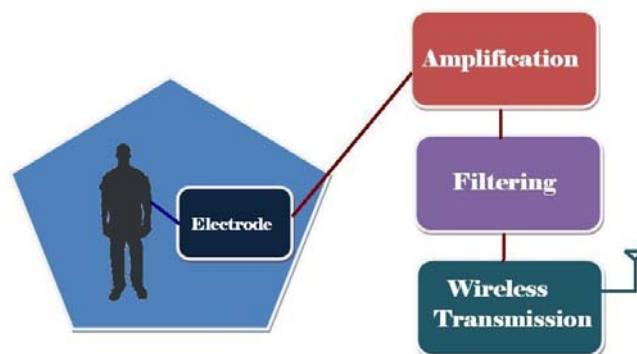


Figure 2. Block Diagram of SATM

The sampling rate of the signal can be changed real-time depending upon the application and the user requirements, if required. The key feature is of this module is the plug n play capability of the module. Different modules can be interface to the system depending on the requirement and the type of parameter that is to be monitored.

4.2 Data Aggregation & Escalation Module (DAEM)

The functionality of DAEM is to receive the continuous real-time signals from the SATM. The received signals will be aggregated, and transmitted to the server through heterogeneous wireless networks. If the congestion experienced in the wireless networks is high, then the signals have to be temporarily stored in DAEM, and later transmitted to the server.

The DAEM hardware architecture is designed to achieve the above mentioned functionalities. The architecture of DAEM is shown in the Figure 3.

The Zigbee module will be receiving the data wirelessly from the SAT module. The DAE module will aggregate the received over a time and will be uploading the data to a remote server, using the available wired backbone or internet. This module will perform an initial level of analysis to detect sensor failure and has an alerting mechanism to notify the user. The actual signal processing takes place at the server. The memory facilitates the temporary storage of received data. The keypad and the LCD Display provide the required user interface. The Ethernet/USB interface provides the base station with the Ethernet and USB interface for external connectivity.

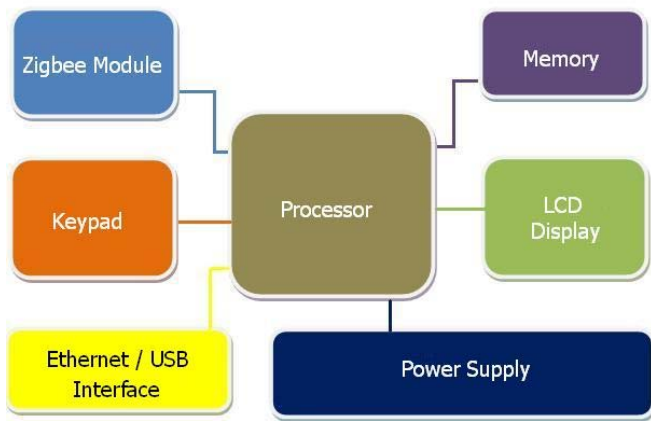


Figure 3. Block Diagram of DAEM

4.3 Database and Signal Processing/Analysis Module(DSPAM)

The functionality of DSPAM module is to store all the transmitted data, perform the essential signal processing and analyze the data.

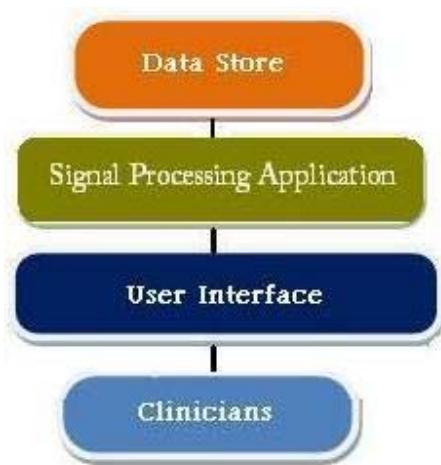


Figure 4. Block Diagram of Server and Signal Processing Application

The analysis results and the raw data will be accessible to the clinicians once they authenticate and login to the application. The signal processing application has access to all the transmitted data and will perform a basic and application specific signal processing depending upon the user requirement. The user interface will display the waveforms. The block diagram is shown in Figure 4. The server can also perform real-time data dissemination depending upon the test scenario and user requirement. i.e.: If the doctor requires an alert to his mobile phone if the measured parameters crosses a predefined threshold, so that he can login and check the real-time data only when there is a necessity.

5 Implementation

As a first phase to the development and implementation of the system, the SAT Module for EMG was developed and tested.



Figure 5. Network Flow Diagram

The acquired signal was wirelessly transmitted to a remote PC using wireless sensor network. The network flow diagram of the test scenario is shown in Figure 5 and the transmitted EMG signal is shown in Figure 6. Since it is an ongoing research activity, the other modules for the entire sleep suite are under development.



Figure 6. Captured EMG Signal from Electrodes.

The IC used for pre-amplification is INA122 from Texas Instruments. INA122 has a variable gain from 1 to 10000, high CMRR, low noise and low quiescent current which makes it suitable for physiological amplification

applications. The filtering is performed in the range of 10 to 500Hz. Active fifth order Butterworth high pass and fifth order Bessel low pass filters were designed, developed and tested. Maximally flat response in both magnitude and phase and nearly linear-phase response in the pass band makes the Bessel filter ideal for this applications. The software tools used for the design and simulation of the above circuits were TinaTI from Texas Instruments, FilterLab from Microchip and LabVIEW & MultiSIM from National Instruments.



Figure 7: Packet Structure

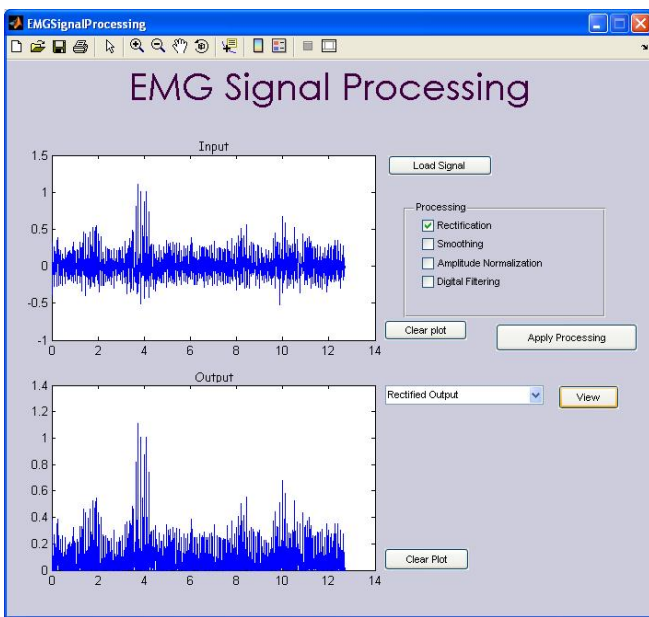


Figure 8 : Signal Processing GUI

The conditioned EMG signals were sampled satisfying the Nyquist criteria and was wirelessly transmitted to a remote location using Wireless sensor Networks having MicaZ nodes. The coding was done using the nesC programming language and TinyOS platform. The packet structure used is shown in Figure 7. The transmitted packet is divided into three main segments, the time stamp, sourceID and data segments. The time stamp helps in synchronizing and reorganize the data and the source ID helps in identifying the source at the receiver. In order to increase the battery life of the node, the data was transmitted only when there was some signal activity.

A GUI was also developed to perform basic signal processing on the EMG signal is shown in Figure 8.

The functionality and integration testing of the modules will be completed shortly. Tests are planned to validate the functionality of the system with multiple nodes and all the modules integrated. The final system will be deployed at Amrita Institute of Medical Science, Kochi, India.

6 Conclusions

The paper provides a novel design for the monitoring and detection of sleep disorders. The main advantage of the system is that remote monitoring and diagnosis is made possible with the proposed system. The system can be manufactured at a very low price compared to the commercially available products in the market. Alerting mechanism provide a feed back to the bystander, if the sensors are not working properly. Since the design has three tier architecture, the system is scalable and robust. The system actually reduces the discomfort level in patients since they can take the test from the comfort of their home and by improving their mobility. The future works envisaged include securing the wireless transmission, QoS analysis considering multiple system implementations and system commercialization. The main aim of the system is to target the rural population of India thereby making them accessible to clinicians and providing better remote healthcare and reliable diagnostic opportunities at low cost. This system can be used for determining different other ailments such as depression, post traumatic stress and relief work.

7 References

- [1] Sleep – Wikipedia <http://en.wikipedia.org/wiki/Sleep>.
- [2] National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health, USA. Available from: http://www.wrongdiagnosis.com/s/sleep_disorders/basics.htm.
- [3] V.M. Kumar, "Sleep Disorders: Current Understanding", The Indian Journal of Chest Diseases & Allied Sciences, 2008, Vol 50 pp 129-135.
- [4] Conquering Depression :Facts and Figures,Mental Health and Substance Abuse, World Health Organization. http://www.searo.who.int/en/Section1174/Section1199/Section1567/Section1826_8101.htm
- [5] Garbarino S., "Sleep disorders and road accidents in truck drivers", G Ital Med Lav Ergon, 2008 Jul-Sep, Vol30(3), pp:291-6.
- [6] V.M.Lubecke, O. Boric-Lubecke, "Wireless technologies in sleep monitoring", Radio and Wireless Symposium, 2009. RWS '09. IEEE. pp.135 - 138.

[7] A. S. Karunajeewa, U. R. Abeyratne, S. I. Rathnayake, V. Swarnkar, "Dynamic Data Analysis in Obstructive Sleep Apnea" , Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE, pp.4510.

[8] A. Burgos, A. Goi, A. Illarramendi, J. Bermudez, "Real-Time Detection of Apneas on a PDA", Information Technology in Biomedicine, IEEE Transactions, Volume 14, Issue 4, pp.995.

[9] PubMed Health, National Center for Biotechnology Information, U.S. National Library of Medicine, <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001803/>

[10] American Academy of Sleep Medicine; European Sleep Research Society, Japanese Society of Sleep Research, Latin American Sleep Society (2001). "The International Classification of Sleep Disorders, Revised: Diagnostic and Coding Manual". 2001 edition, ISBN 0-9657220-1-5, PDF-complete, Library of Congress Catalog No. 97-71405. Retrieved 2010-08-08.

[11] Thorpy MJ, Glovinsky PB. Parasomnias. Psychiatric Clinics of North America, Dec1987, Vol10, Issue 4, pp. 623-39.

Doppler Ultrasound Blood Flow Measurement System for Assessing Coronary Revascularization

J. Solano¹, M. Fuentes¹, A. Villar², J. Prohias², F. García-Nocetti¹

¹Universidad Nacional Autónoma de México, IIMAS, México D.F., 04510, México

²Hospital Hermanos Ameijeiras, La Habana, 10400, Cuba

Abstract - *This work describes a Doppler ultrasound system for measuring blood flow. The system is intended to be used for assessing coronary implants and bypass operations. Quantifying the blood flow through these implants/bypasses is an important task to ensure the surgical process, thus, reducing both the post-surgical and death risks. The system is based on an open architecture that is portable and low-cost, incorporating the advantages of expensive systems with dedicated hardware. It incorporates a pulsed-wave bi-directional Doppler ultrasound flow detector working at 8 MHz. Signal conditioning, detection of direction, signal processing, spectrogram displaying, parameters calculation, and a database handling subsystem complete the system. A graphical user interface is provided for controlling and monitoring the whole system. Doppler signal is processed using both Fourier Transform-based and Parametric Model-based algorithms, having the facility to incorporate alternative higher-resolution spectral estimation methods. The system is being assessed in coronary revascularization.*

Keywords: Blood flow measurement, Doppler ultrasound, signal processing, spectral analysis.

1 Introduction

Ultrasonic techniques have been successfully used in the development of medical diagnostic equipment in obstetrics, cardiology and the peripheral vascular system among others. This equipment may generate the image of an internal structure or the associated spectrogram of an artery's blood flow using external ultrasonic transducers [1,2,3]. Ultrasonic diagnostic is a well-established and widely used technique in almost all medical areas. Although initially its development was focused to obstetrics, very soon several applications were found in cardiology [4].

The use of instruments based on the Doppler principle has allowed extracting phase information from the echoes of body moving structures producing images and sonograms which are used to estimate pressure and flow parameters [5]. Development of pulsed Doppler techniques in conjunction with the signal and image processing methods have generated a notorious increment in the use of ultrasound, opening new

options and displacing other invasive methods used up to nowadays.

This work describes a Doppler ultrasound system for measuring blood flow. The system is intended to be used for assessing coronary implants and bypasses. Quantifying the blood flow through these implants/bypasses is an important task to ensure the surgical process, thus, reducing both the post-surgical and death risks. The system is based on an PC architecture that is portable and low-cost, incorporating the advantages of expensive systems with dedicated hardware. It incorporates a pulsed-wave bi-directional Doppler ultrasound flow detector working at 8 MHz. Signal conditioning, detection of direction, signal processing, spectrogram displaying, parameters calculation, and a database handling subsystem complete the system. A graphical user interface is provided for controlling and monitoring the whole system. Doppler signal is processed using both Fourier Transform-based and parametric model-based algorithms, having the facility to incorporate alternative higher-resolution spectral estimation methods based on time-frequency distributions. The system is being assessed in a number of coronary implant and bypass surgical operations.

2 Doppler ultrasound

Doppler ultrasound systems either continuous or pulsed are used as a non-invasive method for detection and evaluation of the blood flow [6]. Doppler frequency is proportional to blood velocity in the sampled volume and as the arterial blood flow is pulsed the Doppler signal has a spectrum that constantly varies in the time domain.

In ideal conditions the Doppler power spectrum has a similar form to a blood flow histogram in the sampled volume. This is depicted in figure 1a. The analysis of the Doppler signal gives relative information to the evolution of the distribution of the blood particle velocity in the artery [7]. An increment in the Doppler frequency range as a result of some type of turbulence in the blood flow is typically used to detect artery occlusions and other vascular problems, see figure 1b.

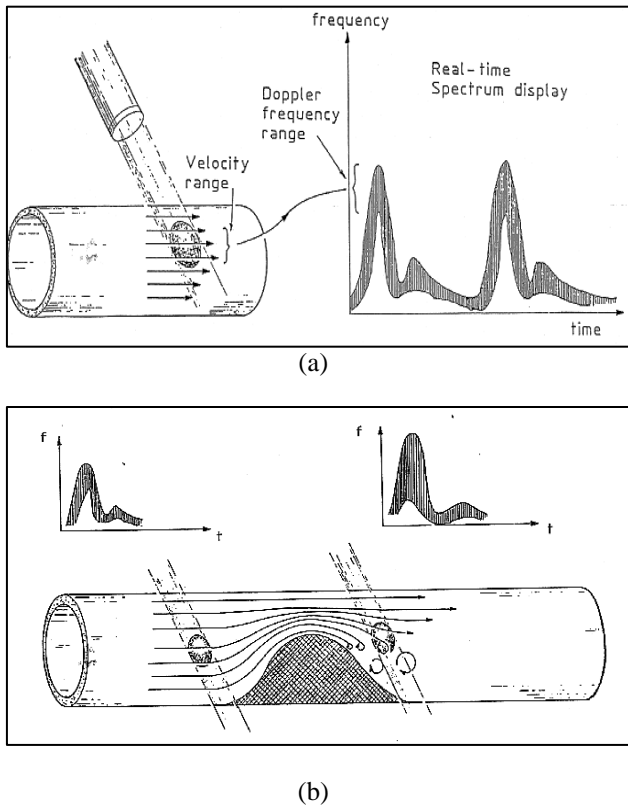
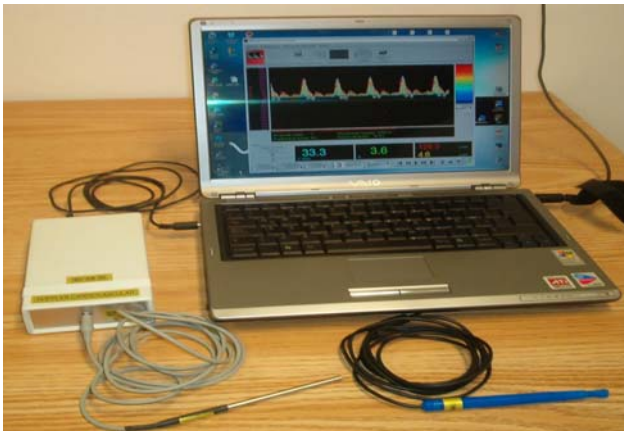


Figure 1.- Doppler ultrasound measurement

3 System description

The system is based on an PC architecture that is portable and low-cost, incorporating the advantages of expensive systems with dedicated hardware. It incorporates a pulsed-wave bi-directional Doppler ultrasound flow detector working at 8 MHz. Flow direction, signal processing, spectrogram displaying, parameters calculation and a database handling subsystem complete the system. A graphical user interface is provided for controlling and monitoring the whole system. Figure 2 shows the complete system. The system described in this work introduces some



modifications in order to optimize its size, cost and operation.
Figure 2.- Doppler ultrasound blood flow system

4 Pulsed wave flow detector

The design of a pulsed wave bi-directional Doppler Ultrasound blood flow detector is presented. The system includes a piezoelectric transducer operating in pulsed wave mode at 8 MHz of frequency. It uses a quadrature phase demodulation for detecting the Doppler signal produced by the blood flow. The Doppler detector generates audio signals I (in phase) and Q (in quadrature). These audio signals in quadrature are used as an input for further processing.

4.1 Sensing probe

The system described in this work incorporates in a sensing probe, the transducer and the detector of the ultrasonic Doppler signal. Figure 3 shows a diagram of this sensing probe. This device has two piezoelectric ceramics, which are excited in a continuous mode, using demodulation in quadrature to detect the ultrasonic Doppler signal and giving as output the I and Q signals. The oscillator-transmitter and the detector-demodulator circuits are integrated in a printed circuit board. These PZT-5 ceramics with a ‘D’ shape are connected 1 cm away from the circuit for noise reduction and a higher sensitivity. System includes an ultrasound 8 MHz probe, however the circuit design allows the use of 4, 5, 8 and 10 MHz piezoelectric ceramics.

4.2 Filters

Considering that the blood flow velocity profile in humans is within the 20–750 mm/s range and the ultrasound velocity in tissue is 1540–1600 m/s [8,9], we may estimate the resulting Doppler signal bandwidth (F_d), and use ultrasonic transducers in the 2–10 MHz range. This Doppler signal may be calculated using the expression; $F_d = (2v/c) f_0$, where v is the blood velocity [m/s], c is the ultrasound velocity in the medium and f_0 is the transducer frequency, using this expression and the values given, the Doppler signal is within the 200–10,000 Hz range. Quadrature signals (I, Q outputs) from the blood flow-sensing probe are connected to a two channel amplifying and filtering module, a schematic diagram of this module is shown in Figure 4. Filters are dynamic analogue, fifth order band-pass and with 300 and 8000 Hz cut frequencies, and a 40–50 dB amplifier per channel.

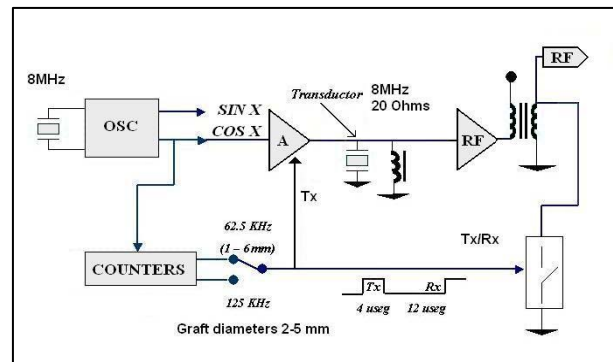


Figure 3.- Sensing probe diagram

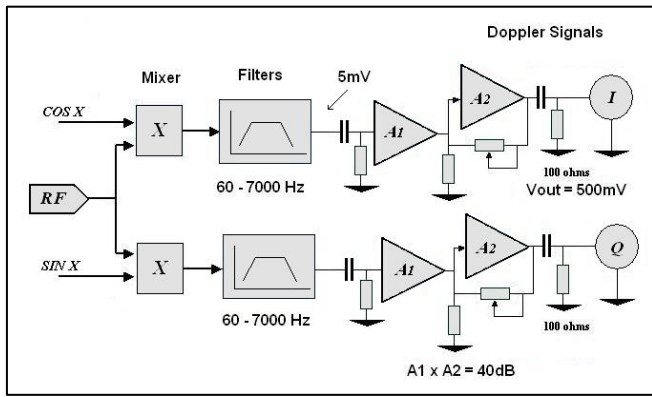


Figure 4.- Amplifying and filtering module

4.3 Flow direction

Blood flow signals $I(t)$ and $Q(t)$ are filtered and amplified giving as a result signals $I(t)$ and $Q(t)$. These signals are input and then are transform into quadrature signals $d(n)$ and $q(n)$ to be digitally processed. There are several methods to transform quadrature signals $d(n)$ and $q(n)$ into flow directional signals $f(n)$ (forward flow) and $r(n)$ (inverse flow). The phasing filter [1] was selected to transform the signals. This has the advantage that the processing time is around milliseconds. Figure 5 shows the block diagram of the algorithm. Here Hilbert transform was implemented using FFT in order to achieve efficiency.

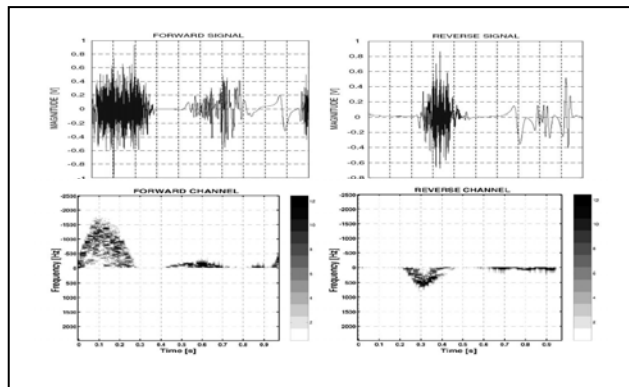
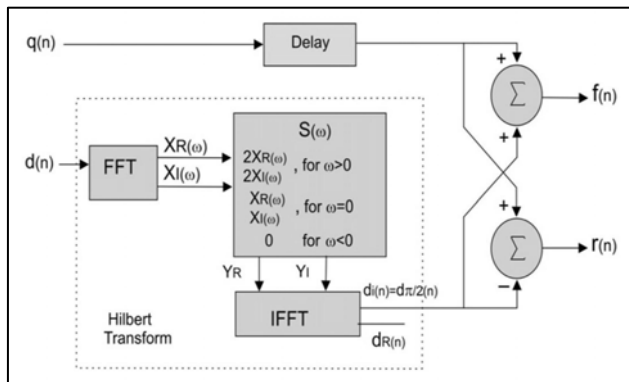


Figure 5.- Phasing filter and flow direction case study

5 Doppler signal processing

In order to measure blood velocity and to monitor its flow, it is necessary to estimate the Doppler signal spectrum. A conventional method to determine and display the spectral information is real time spectral analyzer, see Figure 6. The frequency information of the signal may be display as an amplitude graphic of the signal spectral components versus frequency (frequency spectrum) for each sample interval. Due to the blood velocity in arteries is periodic, the Doppler signal is cycle-stationary, therefore, the Doppler spectrum of each sample interval show variations in the mean frequency and shape along the cardiac cycle. Then, it is necessary to use very short intervals (5–10 ms) where the Doppler signal may be considered as a stationary signal. Spectral power density estimation of a Doppler signal is achieved using methods based on the Fourier transform (FT). However, several research studies present spectral estimation alternative methods such as parametric methods [10-16]. Processing module includes different processing capabilities and calculates automatically the Pulsatility Index, Resistance Index and volumetric flow. The software can also process the Doppler signal using a CFFT (Complex Fast Fourier Transform) algorithm [3,4] or an AR-Modified Covariance algorithm [10] in order to visualize the spectral broadening due to possible stenosis. Doppler blood flow signal is typically represented by a spectrogram where the horizontal axis is time [s], the vertical axis is frequency [Hz] or Volumetric Flow [ml/min] and the Amplitude is represented with a color proportional to its magnitude. The software was developed using C++ programming language and Open GL for graphics display. The Graphical User Interface (GUI) has been developed using GTK. Figure 6 shows examples of spectrograms displaying 512 point windows at 11025 S/seg sampling rate. Hanning windows are used with a 5 ms overlap to reduce the numeric noise due to windowing. The complete spectrogram is build with all the consecutives spectra, scaling the amplitude to a dynamic range 1 – 12 (1 being at 25 dB and 12 at 37 dB). Doppler signal was divided into 2-20 ms overlapped windows and processed.



(a)



(b)

Figure 6.- Spectrograms corresponding to 6 cardiac cycles using (a) FT and (b) AR-modified covariance based methods (over zero values-direct flow, below zero-inverse flow)

6 Tests and results

The testing of the detection device in the laboratory was conducted using a blood flow “phantom” system which includes an electronic controlled pump that emulates different flows and heart rates through 2–4 mm diameter vessels as is shown in figure 7. A mimic blood fluid was used to produce the Doppler effect in the fluid passing through the vessels. The system was also tested in real open-heart surgeries in 10 patients that had coronary implanted grafts, see figure 8.

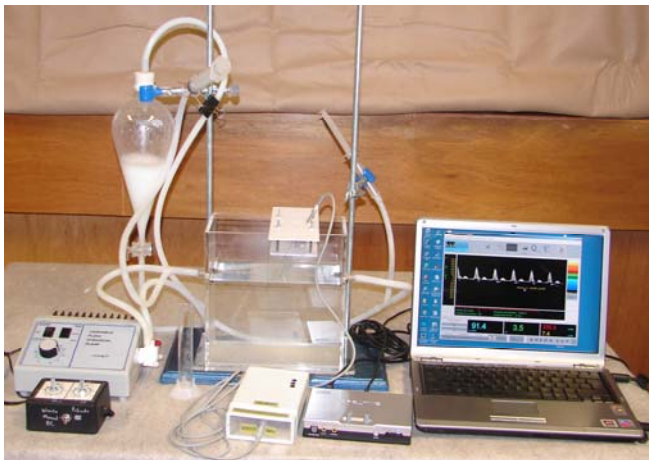


Figure 7.- Doppler ultrasound system *in vitro*

The application software allows the user to select the diameter of the artery, the frequency and angle of the ultrasound probe. It also allows de user to select the amplifier gain, threshold, dynamic range and processing approach (CFFT or Modified Covariance) so the surgeon can visualize the spectrogram according to predefined patterns of the signal. The software incorporates a stand-alone data base that will capture all single or sequential grafts done in each

operation and that can be uploaded or downloaded from a general distributed database system connected via internet.



Figure 8.- Doppler ultrasound system *in vivo*

7 Conclusions

A Doppler ultrasound system for measuring blood flow has been presented. The system is intended to be used in coronary implants and bypasses, aiming to verify the quality of flow in coronary grafts which is essential for the success of a heart surgery and the recovery of a patient with heart disease. Quantifying the blood flow in these implants/bypasses is an important task to ensure the surgical process, thus, reducing both the post-surgical and death risks. The spectrogram output and estimated parameters generated by the system provides important quantitative and qualitative information of the blood flow and can even detect possible errors during surgery or even internal stenosis or “flaps” in the new implanted grafts.

The system is based on an architecture that is portable and low-cost, incorporating the advantages of expensive systems with dedicated hardware. A graphical user interface has been provided for controlling and monitoring the whole system. Doppler signals are processed using both Fourier Transform-based and Parametric Model-based algorithms, having the facility to incorporate alternative higher-resolution spectral estimation methods.

The system has been tested successfully in the laboratory (with synthetic signals in a “phantom”) and during real surgery, separating effectively the direct and inverse flow components of the Doppler signal and giving important information about the quality of blood flow, providing the cardiovascular surgeon with an suitable tool for detecting anomalies during the coronary graft surgery. Further work is being carried out, aiming to provide higher-resolution spectral estimation methods together a number of software tools that can help in the interpretation of the Doppler grafts signals database.

Acknowledgements

Authors acknowledge project DGAPA-UNAM-PAPIIT (IN114710), project CYTED (P506PIC0295) by the financial support. Also we want to acknowledge to A. Hernandez, J.A. Contreras, M. Vazquez and I. Sanchez for their technical support in the development of this work.

8 References

- [1] D.H. Evans, W.N. McDicken, Doppler Ultrasound, Physics, Instrumentation, and Signal Processing. Second edition, John Wiley & Sons Ltd., 2000.
- [2] P. Atkinson, J.P. Woodcock, Doppler Ultrasound and its use in Clinical Measurement, Academic Press Inc. London Ltd., 1982.
- [3] J.A. Jensen, Estimation of Blood Velocities Using Ultrasound, Cambridge Univ. Press, UK, 1996.
- [4] M.D. Cavaye, R.A. White, Arterial Imaging – Modern and Developing Technology, Chapman & Hall Medical, London, 1993.
- [5] R.L. Powis, W.J. Powis, A Thinker's Guide to Ultrasonic Imaging, Urban and Schwarzenberg, 1984.
- [6] S. Marple Lawrence, Computing the discrete-time "analytic" signal via FFT, IEEE Trans. Signal Process. 47 No. 9 (1999) 2600–2603.
- [7] P.J. Fish, Physics and Instrumentation of Diagnostic Medical Ultrasound, John Wiley & Sons, Chichester, UK, 1990.
- [8] P.J. Fish, Non-stationary broadening in pulsed Doppler spectrum measurements, Ultrasound Med. Biol. 17 (1991) 147–155.
- [9] P. Atkinson, A fundamental interpretation of ultrasonic Doppler velocimeter, Ultrasound Med. Biol. 2 (1975) 107–111.
- [10] M.G. Ruano, D.F. Garcia Nocetti, P.J. Fish, P.J. Fleming, Alternative parallel implementations of an AR-modified covariance spectral estimator for diagnostic ultrasonic blood flow studies, Parallel Comput. 19 (1993) 463–476.
- [11] J. Solano, D.F. Garcia Nocetti, M.G. Ruano, High performance parallel-DSP computing in model-based spectral estimation, Microprocess. Microsyst. 23 No. 6 (1999) 337–344.
- [12] J. Solano, K. Rodriguez, D.F. Garcia Nocetti, Model-based spectral estimation of Doppler signals using parallel genetic algorithms, J. Artif. Intell. Med. 19 No.1 (2000) 75–89.
- [13] M.M. Madeira, S.J. Bellis, L.A. Beltran, J. Solano, D.F. Garcia Nocetti, W.P. Marnane, M.O. Tokhi, M.G. Ruano, High performance computing for real time spectral estimation, IFAC J. Control Eng. Pract. 7 No.5 (1999) 679–686.
- [14] J.Y. David, S.A. Jones, D.P. Giddens, Modern spectral analysis techniques for blood flow velocity and spectral measurements with pulsed Doppler ultrasound, IEEE Trans. Biomed. Eng. 38 (1991) 589–596.
- [15] F. Garcia Nocetti, J. Solano Gonzalez, E. Rubio Acosta, E. Moreno Hernandez, Parallel computing in time–frequency distributions for Doppler ultrasound blood flow instrumentation, Mex. J. Biomed. Eng. XXII, 1 (2001) 12–19.
- [16] J. Solano, M. Vazquez, E. Rubio, I. Sanchez, M. Fuentes, F. García-Nocetti, Doppler ultrasound signal spectral response in the measurement of blood flow turbulence caused by stenosis, Physics Procedia, 3 (2010) 605-613.

Analysis of the Electrogastrograms of Elderly Subjects by using Maximum Lyapunov Exponent

Matsuura Yasuyuki, Miyao Masaru, and Takada Hiroki

Abstract— Not much data are available regarding the electrical activity in the stomachs and intestines of elderly gastrectomized patients. The purpose of this study was to determine the feasibility of using a complex dynamic method to analyze the electrogastrograms (EGGs) of healthy young, healthy elderly, and gastrectomized elderly male individuals. We analyzed the EGGs by using the maximum Lyapunov exponent (MLE), which is one of the indices of the chaotic characteristics of time series. Significant differences were observed between the MLEs estimated from the EGGs of the young and elderly individuals for most of the temporal intervals. Our data indicate that the EGGs of elderly gastrectomized subjects might be distinguished from the EGGs of healthy elderly individuals on the basis of the MLE distribution.

I. INTRODUCTION

MANY young women suffer from gastrointestinal diseases such as constipation and functional dyspepsia including gastroesophageal reflux disease (reflux esophagitis). Percutaneous electrogastrograms (EGGs) unrestrainedly and easily measure gastrointestinal activities.

The first electric activity record on the body surface was performed by Alvarez in 1921, and he named it an electrogastrogram (EGG) [1]. EGGs were easily affected by electrocardiograms (ECGs) and electromyograms (EMGs) of the diaphragm during breathing due to the law induced potential from the abdominal wall. There was also no clear association with gastric activity and data analysis methods, and, therefore, they did not achieve clinical application like ECGs and electroencephalograms.

There is regular electrical activity in the stomach and small intestine, like the heart, and electric depolarization and repolarization are repeated. A pacemaker for gastric electrical activity exists in 1/3 of the greater curve of the gastric body,

This work was supported in part by scientific research fund grants (Grant-in-Aid for Scientific Research) and the Young Project "The influence of hot fomentation in the stomach on chronic constipation in the middle-aged and elderly" (2007) by the Japanese Society of Public Health Active Research Fund, and the Hori Art and Science Promotion Foundation.

Matsuura Yasuyuki is with University of Fukui Headquarters for Innovative Society-Academia Cooperation, 3-9-1 Bunkyo, Fukui City, Fukui JAPAN (e-mail: matsuura@nagoya-u.jp).

Miyao Masaru is with Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya JAPAN (e-mail: mmiyao@is.nagoya-u.ac.jp).

Takada Hiroki is with Graduate School of Engineering, University of Fukui, 3-9-1 Bunkyo, Fukui City, Fukui JAPAN (corresponding author to provide phone: +81-776-27-8795; fax: +81-776-27-8955; e-mail: takada@u-fukui.ac.jp).

and electrical activity travels to the pyloric part 3 times per minute (3 cpm, cycle per minute) in humans. The pacemaker triggers periodical electric activities controlled by the vagus nerve. This involves a cell group network called the interstitial cells of Cajal (ICCs) [2 – 4].

The advantages of EGGs were their utility to measure the above-mentioned periodical electric activity and evaluate the (gastrointestinal) autonomic nerve function. In the stomach of resting healthy individuals, peristalsis occurs 3 times per minute when a certain period of time has passed after meals [2 – 4]. The normal range of the EGG fluctuation cycle is between 2.4 – 3.6 cpm, but there is no clear standard except for a frequency close to 3 cpm [5, 6].

EGG studies have made progress with the recent improvement of measurement technology. However, the common EGG analysis method is a spectral analysis technique such as Fast Fourier Transform (FFT), and few reports are available on non-linear analysis. However, considering complex organic activity, non-linear analysis methods including chaos analysis and evaluation based on stochastic process analysis are considered inevitable for the modelization of dynamic movement, an accurate diagnostic index, and extraction of a body assessment index.

Maximum Lyapunov exponent (MLE) is a common index of non-linear analysis [7, 8], and has been widely used in various fields including economic model and sound analysis [9 – 11]. In biosignal analysis, biosignals are considered to be generated based on the non-linear dynamic systems with a few degrees of freedom in the pulse and brain waves, and R – R interval of ECG. Therefore, chaos analysis is used [12 – 13]. In contrast, few reports are available on the chaos analysis of EGGs using the Lyapunov exponent.

A previous study showed that there were groups with and without gastric electrical activities in subtotal gastrectomy cases, although no EGG was recorded in total gastrectomy cases [14]. Therefore, it is difficult to diagnose and judge gastrectomized EGGs of healthy individuals and gastrectomized patients whose intestinal electrical activities and digestive functions decline with age solely with spectral analysis.

The purpose of the present study was to perform a basic examination of non-linear analysis application in EGG. The EGGs of healthy young males, healthy elderly males, and elderly gastrectomized males were analyzed using MLE, which analyzes the chaos of time series signals, and compared.

II. MATERIALS AND METHODS

A. Method

Subjects were 7 healthy young males aged between 21 and 25, 7 healthy elderly males aged between 65 and 76, and 3 elderly gastrectomized males aged between 67 and 76 whose stomach had been resected by more than 2/3. A full explanation was given to the subjects prior to the experiment, and signed consent forms were obtained. The research on young individuals was approved by the Ethics Committee, Nagoya City University Graduate School of Natural Sciences, and the research on healthy elderly and elderly gastrectomized males was approved by the Ethics Committee of Aichi Medical University.

EGG in a supine position was conducted for 90 minutes. Measurement was performed in a sound-insulated (40 dB) experimental room without windows. The room temperature was between 20 – 24°C, humidity was 40 – 55%, and the air current was below 0.1 m/s. Subjects were told to finish meals 2 hours before measurement so that it was not affected by the meals. Measurement was started between 14:00 and 15:00 for all subjects to avoid the influence of circadian rhythm (circadian change).

EGG measurement was performed using unipolar induction. The measurement was amplified by a biomedical amplifier (MT11: NEC Medical), and recorded in a data recorder (PC216 Ax, Sony Precision Technology).

Several methods have been proposed for EGG measurement methods [5], and the number of electrodes and pasting position vary. All the measurement and pasting methods include measurements involving the area closest to the stomach pacemaker. Therefore, measurement was performed in the area closest to the stomach pacemaker in the present study.

EGG electrodes were pasted as shown in Fig. 1, using 2 disposable ECG electrodes (Vitrode Bs, Nihon Kohden). Pasting was performed after confirming a sufficient reduction of skin resistance using Skin Pure (Nihon Kohden).

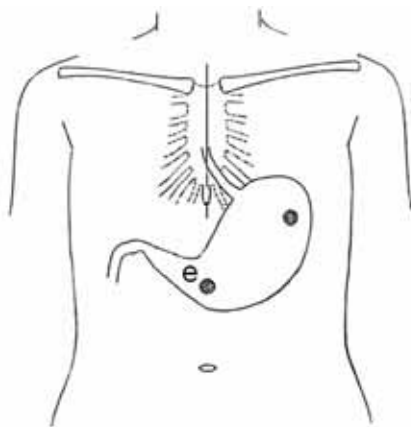


Fig. 1. Pasting position of EGG electrodes.

B. Time-series extraction

The recorded EGG was A/D converted at 1 kHz to obtain time-series data. A low-pass filter for a 0.15-Hz treble cutoff frequency was applied to the obtained data to remove electronic noise from the incorporated EMG and electronic devices, and resampling was performed at 1 Hz to remove noise.

The EGG time series with removed noise was moved at a 300-point (5-minute) interval in a 1,200-second (20 minutes) time window to divide data. EGG time series for a total of 255 subjects (supine position: 15 cases x 17 subjects (7 young healthy, 7 elderly healthy, and 3 gastrectomized individuals)) were developed for analysis.

C. Analysis method

The Lyapunov exponent is a quantity that characterizes the rate of separation of two trajectories on an attractor with time, and demonstrates the enlarged distance of the behavioral gap caused by a minute initial gap [7, 8]. The maximum exponent is called MLE. This exponent quantitatively evaluates the complexity of the attractor that formulates the time series $\{x_t\}_{t=0}^{1199}$.

The Rosenstein analysis method was used in the present study [15, 16]. The attractor was constructed using the obtained data. An infinitesimally close point \mathbf{x}_j from the point \mathbf{x}_i on the attractor was created, and the ratio of the distance with d intervals was assessed by the changes with time as shown in the following Eq.(1).

$$\Delta_i(t, d) = \frac{|\mathbf{x}_i(t+d) - \mathbf{x}_j(t+d)|}{|\mathbf{x}_i(t) - \mathbf{x}_j(t)|}, \quad (1)$$

where the interval d expresses an embedded dimension. The calculations are made for multiple pairs, and uniform operation is performed using the following formula:

$$\langle \log \Delta(t, d) \rangle = \frac{1}{N} \sum_{i=1}^N \log \Delta_i(t, d) \quad (2)$$

The Lyapunov exponent λ is estimated using the following formula in which τ means the embedding delay:

$$\lambda(d) = \frac{1}{\tau} \langle \log \Delta(t, d) \rangle \quad (3)$$

There is a potential for the time series to show chaos when MLE is positive [7, 8]. The bigger the value, the more irregular the wave becomes, suggesting a complex orbit [7, 8]. In the present study, numbers were fixed including the data length for 1,200 points, d for 3 (dimension), and τ for 3 to estimate the MLE.

III. RESULTS

Fig. 2 shows EGGs of healthy young (a), healthy elderly (b), and gastrectomized individuals (c) 10 minutes after measurement initiation for 5 minutes. Normal fluctuation cycles are observed in EGGs of Figs.2 (a), (b), and (c). However, EGGs of the healthy young individuals (Fig. 2 (a)) showed a large amplitude and unstable fluctuation cycle. In contrast, EGGs of the healthy elderly (Fig. 2 (b)) showed a regular pattern, and that of gastrectomized individuals showed wavelengths with a shorter cycle compared to the two other groups, suggesting a different fluctuation pattern.

Figs. 3 (a), (b), and (c) show the two-dimensional attractors ($\tau=3$) formed based on EGGs of the healthy young (Fig.2 (a)), the healthy elderly (Fig.2 (b)), and gastrectomized individual (Fig.2 (c)), respectively.

Fig. 4 shows fluctuation of the average and the standard deviation of the MLE estimated from EGGs of the healthy young, healthy elderly, and gastrectomized individuals. Fig. 5 shows the frequency distribution of the MLE estimated from EGGs of the healthy young, healthy elderly, and gastrectomized individuals.

MLEs of EGGs in the healthy young ranged from 0.69 – 0.79, with an average of 0.75, standard deviation of 0.018, and standard error of 0.069. MLEs of EGGs in the healthy

elderly ranged from 0.59 – 0.76, with an average of 0.72, standard deviation of 0.031, and standard error of 0.012. MLEs of EGGs in the gastrectomized individuals ranged from 0.68 – 0.78 with an average of 0.73, standard deviation of 0.028, and standard error of 0.016. The MLE was positive in all subjects and all analysis intervals. The results suggested that sensitivity to initial conditions was seen in EGGs of the healthy young, healthy elderly, and gastrectomized individuals. There was a significant difference in MLEs estimated from EGGs of the healthy young and healthy elderly according to the time.

IV. DISCUSSION

MLEs were estimated based on EGGs of healthy young, healthy elderly, and gastrectomized individuals. The results showed that MLEs of healthy young individuals were around 0.74. In contrast, MLEs of healthy elderly individuals were around 0.72. The results suggested that EGGs of healthy young individuals are more irregular in wavelength and

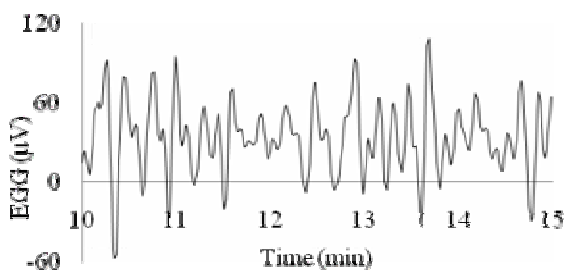


Fig. 2(a). Example of an EGG in a healthy young individual

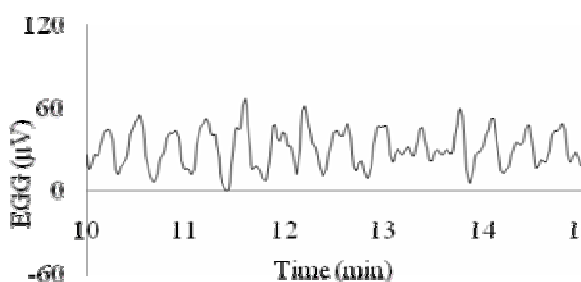


Fig. 2(b). Example of an EGG in a healthy elderly individual

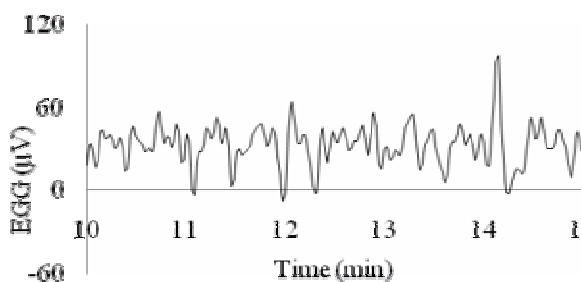


Fig. 2(c). Example of an EGG in a gastrectomized individual

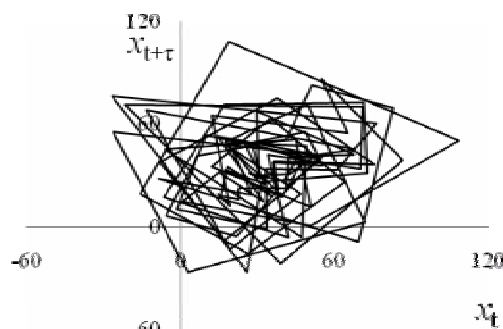


Fig. 3(a). Attractor of the healthy young EGG (Fig. 2(a))

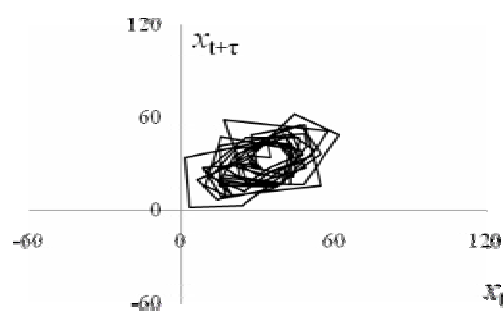


Fig. 3(b). Attractor of the healthy elderly EGG (Fig. 2(b))

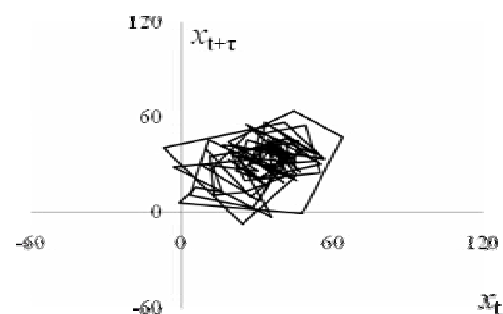


Fig. 3(c). Attractor of the gastrectomized EGG (Fig. 2(c))

complex in the orbit compared to those of the healthy elderly. Although MLEs of the healthy young and healthy elderly EGGs generally continued to be flat, MLEs of gastrectomized individuals' EGGs changed over time. This is considered to be due to the fact that part of the gastric pacemaker cell group was lost on subtotal gastrectomy, and the electrical activity-derived component of the intestine was dominant in the EGG wavelength.

The shape of the MLE frequency distribution showed a one-peak distribution in the healthy young, strained-floor distribution in the healthy elderly, and multiple-peak distribution in the gastrectomized individuals. The elderly show larger individual differences compare to the young. This causes a strained frequency distribution and greater variance in the healthy elderly. Multiple-peak distribution of MLEs in the gastrectomized individuals is considered to be caused by transmission fluctuation of gastric electrical activity due to gastrectomy.

In this study, EGGs of the healthy young, healthy elderly, and gastrectomized individuals were compared using an index in the non-linear analysis. The indices in the non-linear analysis are expected to apply to the evaluation of motion sickness induced by stereoscopic movies.

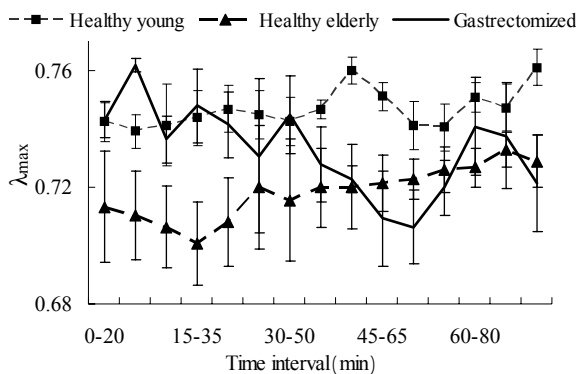


Fig. 4. Average and standard error of MLE (λ_{max})

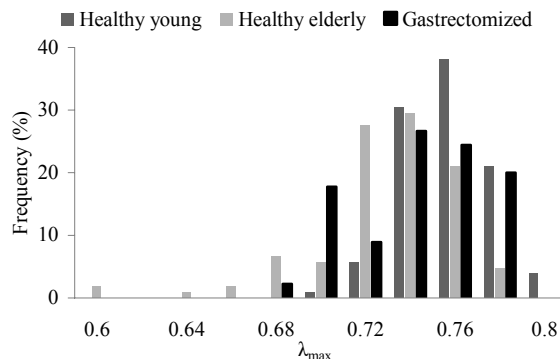


Fig.5. Frequency distribution of MLE (λ_{max})

V. CONCLUSION

EGGs of the healthy young, healthy elderly, and gastrectomized individuals were compared using the MLE, which is an evaluation method of time-series chaos as a basic examination of non-linear analysis method application to EGG.

There was a significant difference in MLEs estimated from EGGs of the healthy young, healthy elderly, and gastrectomized individuals according to the time. There is a potential for EGG classification of gastrectomized individuals based on the MLE distribution.

The MLE was used for analysis in the present study, and further basic examinations are planned employing other non-linear analysis methods.

EGGs of 3 gastrectomized individuals were used in the present study. Further studies are planned with an increasing number of cases.

REFERENCES

- [1] W. C. Alvarez, "The electrogastrogram and what it shows," *Journal of the American Medical Association*, vol.78, pp.1116-1119, 1922.
- [2] S. Homma, "Isopower mapping of the electrogastrogram (EGG)," *Journal of the Autonomic Nervous System*, vol.62, pp.163-166, 1997.
- [3] L. K. Kenneth, M. Robert, *Handbook of Electrogastrography*, Oxford University Press, Oxford, 2004.
- [4] J Z. Chen, R. W. McCallum, *Electrogastrography Principles and Applications*, Raven Press, New York, 1994.
- [5] J. Z. Chen, R. W. McCallum, "Clinical applications of electrogastrography", *American Journal of Gastroenterology*, vol.88, no. 9, pp.1324-1336, 1993.
- [6] M. Nagai, M. Wada, Y. Kobayashi, S. Togawa, "Effects of lumbar skin warming on gastric motility and blood pressure in humans", *Japanese Journal of Physiology*, vol.53, no. 1, pp.45-51, 2003.
- [7] A. M. Lyapunov, *The general problem of the stability of motion*, Comm. Soc. Math., Kharkow, 1892 (in Russian) (reprinted in English, A. M. Lyapunov, "The general problem of the stability of motion," *International Journal of Control*, vol.55, no.3, pp.531-534, 1992)
- [8] C. Sato , *Theory of nonlinear oscillation*. Asakura Shoten, Tokyo, 1970.
- [9] M. Tanaka-Yamawaki, M. Tabuse, "A Dynamical Model of the Economic System : Simulation and its Time Series Analysis," *Technical report of IEICE*, NLP98-4, pp.23-29, 1998.
- [10] E. Hojin, Y. Shiraishi, N. Furuse, "A Comparison among Lyapunov Exponents Calculation Methods in Human Voice Analysis," *Technical report of IEICE*, CAS2003-3, pp.13-18, 2003.
- [11] T. Suzuki, M. Nakagawa, "Fluctuation of the vocal sound and its chaotic and fractal analyses," *Technical report of IEICE*, NLP2004-55 , pp.7-12, 2004.
- [12] T. Sugiura, T. Iokibe, S. Murata, M. Koyama, "A Method for Discrimination of Arrhythmia by Chaotic Approach," *Journal of Japan Society for Fuzzy*, vol.8, no. 3, pp.541-546, 1996.
- [13] Y. Fujiwara, H. Genno, K. Matsumoto, R. Suzuki, K. Fukushima, "Estimating Human Sensations Using Chaos Analysis of Nose Skin Temperature," *Journal of Japan Society for Fuzzy* , vol.8, no.1 , pp.95-104 , 1996.
- [14] K. Imai, M. Sakita, "Pre- and postoperative electrogastrography in patients with gastric cancer," *Hepatogastroenterology*, vol.52, pp.639-644, 2005
- [15] M. T. Rosenstein, J. J. Collius, C. J. De Luca, "A practical method for

calculating largest Lyapunov exponents from small data series,”
Physica D, vol.65, pp.117-134, 1993.

- [16] S. Sato, M. Sano, Y. Sawada, “Practical methods of measuring the generalized dimension and the largest Lyapunov exponent in high dimensional chaotic systems,” *Progress of theoretical physics*, vol.77, no.1, pp.1-5, 1987.

A Personalized Health Information System to foster Preventive Medicine

Sebastian Klenk, Julia Möhrmann, Andre Burkovski, Jürgen Dippon, Peter Fritz and Gunther Heidemann

Abstract—The first and foremost task of all health information systems is to inform the users about their current health level. Good systems give information on what action might change their status quo for the better. A excellent system would engage the user in these actions to improve their health in a sustainable way.

In the course of this paper we will demonstrate how such a system might look like. Our main emphasis will be on its attainability with currently available data sources.

I. INTRODUCTION

People are traveling in their cars along a highway, only to find that the road heads directly off a cliff. Not surprisingly, this creates a pileup at the cliffs bottom with all sorts of injuries and fatalities. So, where do you put the hospital?

This quote taken from Goetz's book "The decision tree" [5] illustrates the dilemma we have with our current health care system: We treat people only after they fell off the cliff and do not prevent them from falling. In this paper we will present a system that is intended to make people think about performing a u-turn before they reach the cliff. We will start with a small example which will serve as a guidance through this paper.

Example 1: You, a health concerned user, decide to do something about your health. You consult your doctor and buy a lot of books and magazines. The information you get from your doctor is rather medical as well as detailed and you get the advice to exercise more and perform a healthy diet. The facts and information you get from books, magazines or health portals are also rather generic and in no way personalized. After a number of days with healthy food and irregular exercise your motivation drops. You neither see any progress nor does the abstract idea of better health allow you to further remain obedient.

What is clearly lacking in this example is the connection between your current situation, its outcome and the improved outcome after exercise and a better diet. The same example with feedback on the actions taken would result in a dramatically different outcome.

Example 2: You, a health concerned user, decide to do something about your health. You consult your doctor and buy a lot of books and magazines. The information you get

is very specific to your current situation. You get detailed information on how more exercise and a healthier diet improve your health and reduce the risk of fatal diseases. After each of your exercise you can see how your life expectancy changes and after each healthy meal you see the risk reduction for different diseases. This kind of continuous feedback keeps you on track for the next few years.

Goetz [5] argues in his book that personal monitoring and direct feedback allows for more conscious decisions. In the next few sections we will propose a system that can support such a process. It monitors personal lifestyle data, compares the data with epidemiological data, estimates probable outcomes and proposes alternatives.

II. THE DATA

Changing people's behavior requires them to be knowledgeable about their current actions and what consequences of these actions are. Further it is necessary to demonstrate how a change in action positively influences their health. At first sight, this seems to be rather trivial and the, therefore necessary, data easy to obtain. Basically we would need data on

- 1) the current health condition, the current lifestyle and
- 2) on the expected progress or decline of the person's health.

But obtaining objective, quantitative data on people's lifestyle is still a question of active research. How do we measure the average stress level, or the overall amount of exercises people actually perform? How to measure people's diet, alcohol or smoking habits? For obvious reasons, having them record each of these factors by hand does not work. We need a (semi-) automatic way of determining these aspects.

A. Personal lifestyle information

As lifestyle data we will consider any data that is in any way connected to the personal lifestyle. This could be dietary facts, the current stress level, the amount of exercise, relationship status or parental status. Even though this data is all around us, it is not easy to access. We believe that a semi-automated approach is most promising. We will now discuss how the raw data can be obtained and we will present methods how to calculate actual information from the different sources.

1) *Movement data:* Lifestyle data, which is relatively easy to obtain, is movement data. How much does the person move, or at least how much does the smart phone of a person

Sebastian Klenk, Julia Möhrmann, Andre Burkovski, Jürgen Dippon and Gunther Heidemann are with the University of Stuttgart, Stuttgart, 70569, Germany (phone: +49 (0)711 685 88-241; email: klenksn@vis.uni-stuttgart.de).

Peter Fritz is with the Institut für Digitale Medizin, Stuttgart, 70192, Germany

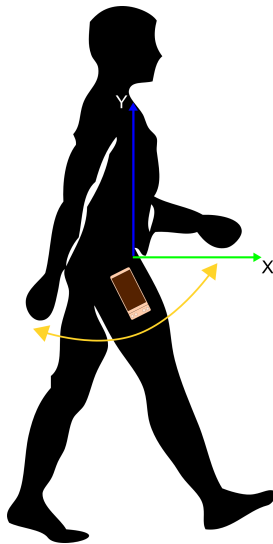


Fig. 1. The structure of the personalized information system. On the left side of the brick wall is the medical information system. On the right side is the patient information portal.

move? The difficulty with this kind of data is that, as long as it is not processed, it is of little use.

Figure 1 shows how the data is collected. Most of today's smart phones are capable of collecting acceleration data. There are three sensors, one for each axis in a three-dimensional space. Each of these sensors collects information on the acceleration along its axis. Once the data is inside the phone it has to be classified. This is necessary because not all movement is actually due to self-induced body movement. Some movement might be caused by the movement of a car the user is in or because he or she is taking the elevator instead of walking up the stairs. Figure 2 shows the acceleration curves for these two cases. There is already some promising research being done in this area [3], [1]. These approaches use time-frequency domain features and let users label their data themselves. The result is a time series of different discrete blocks of activities. These include movement-related aspects such as walking, running or climbing stairs, but also things like driving a car or taking an elevator.

With information about different activities performed by a person and the duration of each of these activities it is easy to obtain a measure for the level of exercise performed by that person. This is already a good health indicator, but focusing solely on the amount of movement might also lead into wrong directions. A high stress level could result in a high degree of personal movement which would lead to the conclusion of a healthy life. We therefore have to include further information.

2) *Stress level*: Obtaining information on psychological aspects of a person's health is a difficult task. Stress for example is perceived and handled differently by different people. There are however physical manifestations of stress such as in speech [6]. The voice and the articulation changes significantly when people are under stress. This is used, for

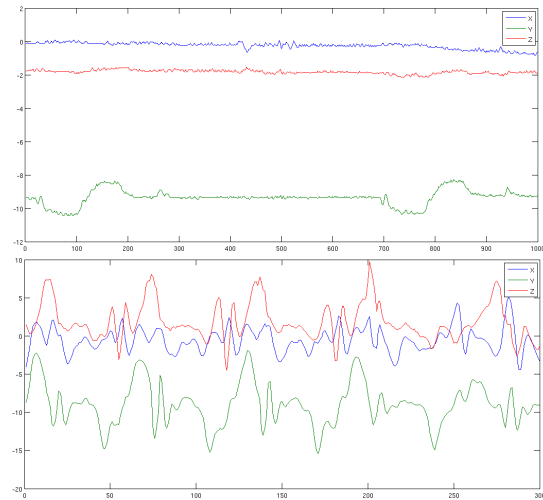


Fig. 2. Acceleration curves for an elevator ride (top) and a walk upstairs (bottom).

example, in the area of driver safety to detect the stress and distraction level of a driver [2]. A similar approach could be used to detect the stress level of a person answering the phone. Given the speech data from phone calls, the same algorithms can be used to determine whether the speaker is currently under pressure or not. With this information we could determine the stress level at the moment of the phone call. With the data on all phone calls a user performs it would be possible to calculate a stress measure that could be used as a health indicator.

3) *Eating habits*: Exercise and Stress are just two important aspects of personal health. Others are eating habits. Again, modern smart phones can help us gather information on this lifestyle fact. Because, especially in urban areas, people eat out most of the time, eating habits can be obtained from their geographical location.

Most restaurants are visited for the food they are famous for. Therefore concluding from the restaurants people visit to the food they eat is not too far of a reach. People visit fast food restaurants for fast food, not for salad. They go to a sportsbar for beer and wings and not for juice and vegetarian food. If we accumulate data from several restaurant visits we might get a quite good idea of eating habits.

The difficulty with this kind of information is the classification of restaurant. Not all restaurants are already labeled with a suitable class label. We would therefore require the user to provide some initial information on the type of restaurant he is visiting and on the kind of they serve.

4) *Semi automatic classification*: So far we assumed that given the data and a smart enough algorithm we can deduce almost any information. Movement can be classified by looking at different frequencies, listening closely gives us information on the stress level and the restaurant visits reveal the kind of dish a person likes. Unfortunately it probably will not work that smoothly. The algorithms will most likely need further information. Such input could be a verification of a classification result: was the classification the algorithm

performed right or should it make adjustments. "Was it right to deduce that you just had steak after you visited this steak house? No this place is famous for its fresh salad bar, I had salad". Semi automatic approaches to machine learning come in different flavors and are long known in the pattern recognition community [7].

B. Medical data

Besides personal lifestyle data we will also need data on the outcome of the current lifestyle of people. This data can be obtained from different clinical or epidemiological sources.

From the early days of medicine data on the success or failure of treatments has been gathered. Evidence based medicine has increased the importance of such data collections. They have become the foundation of treatment decisions. Especially for chronic diseases, such as heart diseases, stroke, cancer or diabetes there are large collections of data that cover numerous aspects of a person as well as follow-up information. These data bases have been used as rich source for epidemiologist and should now be opened up to all people.

1) *Epidemiological data*: From the point of view of those people involved, these large collections of data are a good thing because all the relevant data is already available. The drawback is that, in its current complexity, most likely, people will be overwhelmed by the amount of information available. Fortunately, given the information about peoples current condition, a small number of key diagnosis dates is sufficient to calculate all necessary probabilities. Therefore it is easy to calculate the personalized expected development of a chronic disease of a person, given only his current age and as little as five to ten other variables. The obtained data can then be compared to the expected development under a different condition. All the data which is necessary for such a calculation is already contained in different research data sets. The computations required to analyze them are mostly known for a quarter of a century. Almost all statistics packages are therefore perfectly capable of analyzing them. The resulting plots, Kaplan-Meier diagrams or hazard curves are easily interpretable, even by non experts.

2) *Further information*: So far, we only discussed data that can be statistically analyzed. Such data is an important source of information when it comes to determining what the expected outcome of a persons current lifestyle is. Besides this quantitative data there is also qualitative data that is of interest for people. Such information might be data on different diets or exercises, the side effects of a special treatment or whether any of this is covered by insurance. Most of this information is already publicly available in different sources and different qualities.

One major source of qualitative information are medical publishers such as Thieme or Springer. Most of them provide some form of online service. These services provide information on diseases, treatments, drugs and other health care related topics. Depending on the targeted audience the treatment of the subject ranges from coarse to very fine and

detailed. Besides services that stem from print products there are a number of native online services such as WebMD, MedicineNet or Healthline.

Other important sources of information can be found in user generated content such as the open encyclopedia Wikipedia or public health portals and forums such as iMedix or eHealthForum. Of course the degree of quality in these sources varies extremely depending on the person who contributed the content.

All these sources of health data provide information that can be searched for, and found, with simple key word based queries. Such a search benefits mostly from the vast amount of data available online.

III. THE INFORMATION SYSTEM

When you decide to change your lifestyle towards a more sustainable way of life and you are confronted with a system that bombards you with all the data described above, you will soon stop using such a system. Even worse, instead of having reassuring guidance you will feel confused and insecure.

Data is probably the most important aspect of an information system, but the extraction of the valuable information from the vast amount of data available is also the most difficult aspect of such a system.

When presenting personalized information, focusing on the right information is a major challenge. In this section we will show how system architecture and information flow can be constructed such that only relevant and suitable information will be presented to the user.

The personalized information system we are proposing consists of three parts. Each of these parts performs its own data processing. The structure of the system can be seen in Figure 3. The first part is the, already existing, medical information system (left to the brick wall). Its main objective is the statistical analysis of large patient data corpora. The second part (right to the brick wall in Figure 3) is the personal health information portal which gathers and aggregates data from different sources. The third part, and one of the data sources, is the mobile application that collects the personal lifestyle information.

A. Medical information system

The statistical analysis of patient data has a long and fruitful history in medicine. Most medical research work is based on thorough numerical evaluation. The results of these calculations are mostly left to the physician to interpret. Patients usually don't come in contact with this kind of data. There is a good reason for this: The statistically valid interpretation of subtle statistical differences is neither easy nor obvious.

In a health information system the software decides which are the different options that can be compared. This way only statistically valid queries can be formulated. Interpreting the resulting plots from these queries, is also made much more easy as only a very small number of intuitive plots are presented.

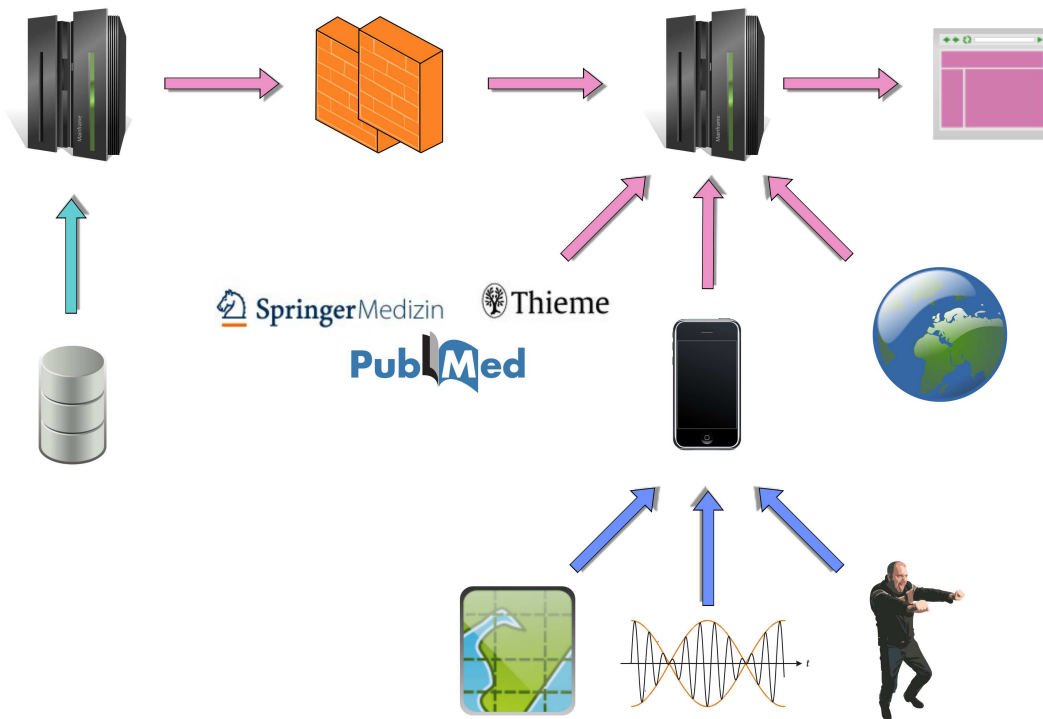


Fig. 3. The structure of the personalized information system. On the left side of the brick wall is the medical information system. On the right side is the health information portal with the mobile data collector.

B. Personal health information portal

Portals are places where data, people and services are integrated into one seamless presentation of information. The proposed personal health information portal integrates clinical and statistical data from the medical information system with personal health information, lifestyle information other publicly available data sources.

The clinical data will be obtained directly from medical information systems. Here the physical separation (represented by the brick wall in Figure 3) serves two purposes: First, it guarantees that only data will be presented to the user which is suitable for his needs. Second, it also protects the privacy rights of other patients in the database underlying the medical information system.

The publicly available data will be obtained from different sources. First of all there are medical publishers that provide information services. Some of them also provide standardized access methods such as SOAP, REST or the, currently under development, InfoButton Specification [4]. The main difficulty when connecting to these services is the standardized nomenclature. Different services require different name spaces. The InfoButton for example requires ICD-10 codes whereas other sources allow for ordinary key word search.

Accessing open sources such as Wikipedia is usually rather easy. Here XML-APIs are provided. Other social web sites have to be included into the portal on a individual basis.

C. Personal lifestyle app

Applications running on smart phones, usually called apps, have become extremely popular especially because of their simple interfaces and their easy handling. This should also be the driving force when developing such an application for health information collection: simplicity and ease of use [8].

The app in this system serves as a data entry method only. This means that its only purpose is to collect data, query information from the user and send this data to the information system. Therefore not much user interaction is required. The user has to be able to respond to queries and alter the information the app has generated. This might be the case when the app falsely assumes the user is performing some exercise or is at a certain type of restaurant. In such a situation the user has to be able to change the current data that is generated by the app.

D. Related work

There is, of course, already a number of methods and systems that try to tackle the kinds of problems discussed in this paper. There are, for example numerous videos or questionnaires that serve as health guidances or decision aids. Most of them are either non interactive, such as videos, or in form of questionnaires that present answers to predefined questions. Especially the interactive and explorative character of the system proposed in this paper is aimed at fostering a change in lifestyle that people can identify themselves with and have trust in.

Besides several approaches to foster a healthier lifestyle there are several web sites that server a more educational purpose (such as WebMD – <http://www.webmd.com> or heart.org – <http://www.heart.org>). These lack the interactivity and the personalization that is achieved through integrating the patients clinical data. The same holds true for web sites for people with chronic diseases, where they can organize their drug regiments and symptomatology [9].

IV. CONCLUSION

A sustainable change towards a healthy lifestyle requires a thorough commitment. Such a commitment is only possible if the person truly believes that such a change is necessary. We have demonstrated, in the course of this paper, how different data sources and information on the personal lifestyle can be combined in a personalized information system that provides a user with enough information to come to such a conclusion.

We argue that the current smart phone generation is well equipped to collect all data necessary to form a sufficiently clear picture of a users lifestyle. As we discussed whether state of the art technology is capable of extraction information on the users stress level, eating habits and amount of exercises performed.

Medical data and online sources on health questions are an other cornerstone of the system proposed. They augment the personal lifestyle data with life expectancy data or further information such as dietary facts or suitable exercises.

Personal health information is of concern to everyone. Making well informed decisions on our personal health should be the rule and not the exception. Public and personalized access to health information, as we propose it will benefit this cause.

REFERENCES

- [1] BAO, L., AND INTILLE, S. S. Activity recognition from user-annotated acceleration data. *Pervasive 2004* (April 2004), 1–17.
- [2] BOŘIL, H., BOYRAZ, P., AND HANSEN, J. H. L. Towards multi-modal driver's stress detection. In *Proc. of 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety* (Dallas, TX, 2009).
- [3] BREZMES, T., GORRICO, J.-L., AND COTRINA, J. Activity recognition from accelerometer data on a mobile phone. In *IWANN '09: Proceedings of the 10th International Work-Conference on Artificial Neural Networks* (Berlin, Heidelberg, 2009), Springer-Verlag, pp. 796–799.
- [4] DEL FIOLE, G. Context-aware knowledge retrieval (infobutton) decision support service implementation guide. Tech. rep., HL7, 2010.
- [5] GOETZ, T. *The decision tree*. Rodale, 2010.
- [6] HANSEN, J., AND PATIL, S. Speech under stress: Analysis, modeling and recognition. In *Speaker Classification I*, C. Miller, Ed., vol. 4343 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2007, pp. 108–137.
- [7] HASTIE, T. J., TIBSHIRANI, R. J., AND FRIEDMAN, J. H. *The elements of statistical learning*, corrected print. ed. Springer, 2002.
- [8] LINGHAO, Z., AND YING, L. On methods of designing smartphone interface. In *Software Engineering and Service Sciences (ICSESS), 2010 IEEE International Conference on* (2010), pp. 584 –587.
- [9] MOLPUS, J., Ed. *The Impact of Personalized Medicine Today*. Breakthroughs Reports. HealthLeaders Media, April 2010.

A High Throughput Computational Analysis of Claudin Gene Family in Human Ovarian Cancer

Dr. Shaukat I. Malik¹, S. Sameen², and Z. Khalid²

¹Department of Bioinformatics, Mohammad Ali Jinnah University (MAJU) New Campus, Islamabad, Pakistan

²Department of Bioinformatics & Biotechnology, International Islamic University (IIU), Islamabad, Pakistan

Abstract- *Gene expression analysis is very instrumental in understanding the pathogenesis of the disease. To enhance the understanding of the molecular basis of the disease there is a need to extract the buried patterns in gene expression profiles. This paper is intended to provide a computational approach for the analysis of claudin gene family association with the pathogenesis of ovarian cancer. Our analysis verified some major members of claudin family as either up regulated or down regulated. It shows the differential expression of cldn3, cldn4 and cldn7 in ovarian cancer. In addition to that the up regulation of cldn16 and down regulation of cldn5 in human ovarian cancer has also been observed.*

Keywords: Ovarian cancer, Claudin, Expression analysis.

1 Introduction

Ovarian cancer is the sixth most common cause of cancer death among woman worldwide ^[1]. Environmental and genetic factors are both important in ovarian carcinogenesis. This disease predominantly affects postmenopausal women causing approximately 13,300 deaths each year and for over half of all deaths from genital cancer. The highly lethal nature of this tumor is related to the absence of symptoms in the majority of women with early stages of the disease and it is the leading cause of mortality due to gynecological malignancy. In the past two decades, much progress has been made in identifying genes involved in the development of ovarian cancer. These identified genes are useful in understanding the pathogenesis of ovarian cancer and defining its molecular signature. They can also serve as biomarkers for early diagnosis and targets for drug development. Claudin gene family is implicated with various types of cancers ^{[2] [3] [4] [5]}. This family consists of 23 tight junction proteins ^[6]. The correct arrangement of all claudin genes is very necessary to perform its function which is the formation of tight junctions. Any problem in its gene arrangement causes cancers. Association of ovarian cancer with some members of the claudin family has already been reported before e.g. cldn3 ^[7], cldn4 ^[8] and cldn7 ^{[9] [10] [11] [12]}. The function of claudins is highly tissue specific because claudin3 and claudin4 was observed in ovarian cancer but not in ovarian cystadenomas ^[13]. Here we will computationally

analyze the whole claudin gene family association with ovarian carcinoma.

2 Methods

2.1 Gene Finder tool

The Gene Finder is a tool that identifies one gene or list of genes, based on selected search criteria. This tool is available at Cancer Genome Anatomy Project (CGAP) official website <http://cgap.nci.nih.gov/Genes/GeneFinder>. By choosing the search criteria as ovarian cancer and claudin gene family it showed all of the ovarian cancer related genes of claudin family.

2.2 SAGE Genie

The SAGE Genie is a gene expression database that reliably matches SAGE tags, 10 or 17 nucleotides in length, to known genes. It not only produces the list of tags but also provide the frequency of occurrence of these tags in each normal and cancerous tissues by scanning all of the given expression libraries. All publicly available data to date was used for the analysis of gene expression of claudin gene family. SAGE anatomic viewer <http://cgap.nci.nih.gov/SAGE/AnatomicViewer> ^[14] was used for collection of tags. Both NlaIII and Sau3A tags were mapped to UniGene clusters <http://www.ncbi.nlm.nih.gov/unigene/>. The reliable UniGene clusters matched to claudin tags were adopted to determine the levels of expression of claudin gene family in both normal and ovarian cancer libraries. The list of tags and matched unigene clusters is provided in table 2.

2.3 Virtual Northern

Virtual northern available at CGAP allows the user to observe the expression of a specific gene in all SAGE and EST libraries. Five libraries of ovarian carcinoma and two of normal ovarian expression were studied in northern blot analysis for the expression patterns of all of the gene

Symbol	Name	Sequence ID
CLDN1	Claudin 1	NM_021101
CLDN10	Claudin 10	NM_182848 NM_001160100 NM_006984
CLDN15	Claudin 15	NM_001185080 NM_014343
CLDN16	Claudin 16	NM_006580
CLDN3	Claudin 3	NM_001306
CLDN4	Claudin 4	NM_001305
CLDN5	Claudin 5	NM_001130861 NM_003277
CLDN6	Claudin 6	NM_021195
CLDN7	Claudin 7	NM_001307 NM_001185023 NM_001185022
CLDND1	Claudin domain containing 1	NM_001040199 NM_019895 NM_001040182 NM_001040181 NM_001040183 NM_001040200

members of claudin gene family. The difference of greater than two fold was considered significant.

2.4 Microarray analysis

Two microarray datasets having normal and cancerous ovarian cancer tissues available on Gene Expression Omnibus (GEO) <http://www.ncbi.nlm.nih.gov/geo/> was used. The dataset GSE6008 contain 99 individual ovarian tumors and 4 individual normal ovary samples contributed by Hendrix ND ^[15] and the second dataset GSE4122 is contributed by [Tate DL](#) and co workers with 32 cancerous tumors and 14 controls. These datasets was used for further verification of SAGE and northern blot analysis results. Statistical analysis was done to analyze the microarray expression. Two major statistics applied to available data was t-test and significance analysis of microarrays (SAM).

3 Results

3.1 Gene Finder Results

Gene finder provided the list of only those claudin gene family members which found to be more frequently involved in the expression of ovarian cancer on the basis of number of tags available in CGAP libraries of ovarian cancer as compared to normal ovarian libraries. These selected gene members from claudin gene family are shown in table 1. Only these selected genes were chosen for further analysis.

Table 1: Gene finder results of claudin gene family members involved in the expression of ovarian cancer

3.2 SAGE and virtual Northern blot analysis of Claudin genes expression in ovarian cancer

There are two normal ovarian libraries and five SAGE libraries of ovarian cancer tissues available in GEO. This was also observed by using tool SAGE Absolute Level Lister (SALL) <http://cgap.nci.nih.gov/SAGE/SALL?ORG=Hs> available at CGAP website. The reliable tags of all selected claudin gene family members were then extracted from SAGE Genie by as shown in table 2. Only those genes were picked which have atleast > 2 fold difference. From 10 genes 7 genes were found to have greater than 2 fold difference. The cldn1, cldn 10 and cldnd1 have no significant differences or they found to have almost same result that's why they were excluded from the list. The virtual northern results confirm the involvement of cldn3, cldn7, cldn4, cldn15, cldn16, cldn5 and cldn6 in the ovarian cancer.

GENE ID	UNIGENE CLUSTER	SAGE TAG	NORMAL (TPM)	OVARIAN CANCER (TPM)
CLDN3	Hs.647023	CTCGCGCTGG	0.0	77
CLDN7	Hs.513915	TATAGTCCTC	0.0	37
CLDN4	Hs.647036	ATCGTGGCGG	0.0	91
CLDN15	Hs.38738	GCCCCTCCAG	4	9
CLDN16	Hs.251391	TTGCCATCCT	0.0	4
CLDN6	Hs.533779	TTTGTTAGT	0.0	28
CLDN5	Hs.505337	GACCGCGGCT	0.0	14

Table 2: SAGE Anatomic Viewer and Northern blot analysis results:

3.3 Microarray analysis of Claudin genes expression in ovarian cancer

The involvement of above selected genes of claudin family was then verified by Microarray analysis. Two datasets GSE6008 and GSE4122 from GEO contains the gene expression information related to ovarian cancer. All of the above mentioned genes can be located in these data sets. The results obtained from both datasets (table 4) shows that cldn4, cldn7, cldn16 and cldn3 are highly over-expressed in ovarian cancer while cldn5 is down regulated. Cldn6 and cldn15 showed a very different behavior, as cldn6 is found to have over expression and down regulation of cldn15 in dataset GSE4122 while in GSE6008 no significant difference is detected. These findings through fold change analysis were further verified through t-test and SAM. The t-test confirmed the cldn3, cldn4, cldn7, cldn16, cldn15 and cldn5 as significant genes and cldn6 was the only non significant gene so it was excluded from further analysis. The differential expression of these significant genes was also detected in a volcano plot in fig 1. Further mining of selected members of

claudin family was done through SAM which separated cldn4, cldn3, and cldn7 as positive significant genes and cldn5 as negative significant. This is also shown in SAM graph fig 2.

Table 4: Microarray results of dataset GSE6008 & GSE4122

CLAUDIN GENES	REFERENCE ID	FOLD CHANGE IN GSE 6008	FOLD CHANGE IN GSE4122	P-VALUE	FALSE DISCOVERY RATE
CLDN4	201428_AT	> 2	> 5	0.0018636247	0.003261343
CLDN7	202790_AT	> 2	> 3	2.9067648E-6	6.7824512E-6
CLDN3	203953_S_AT	> 3	> 6	0.0	0.0
CLDN5	204482_AT	< 2	< 3	0.009433004	0.011005172
CLDN15	219640_AT	NO CHANGE	NO CHANGE	1.7732685E-8	6.2064395E-8
CLDN16	220332_AT	> 1.5	> 3	0.004344086	0.0060817203

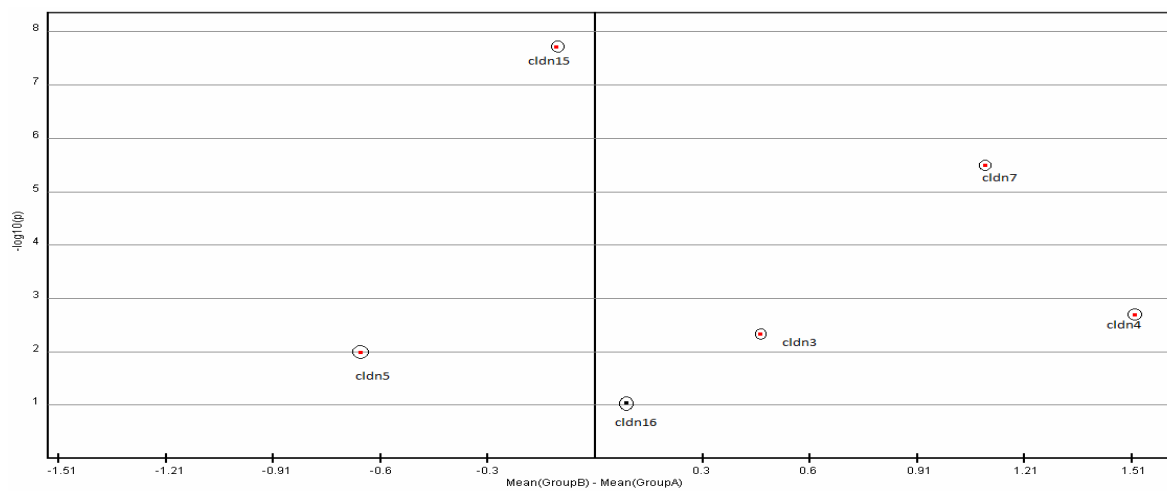


FIGURE 1: VOLCANO PLOT

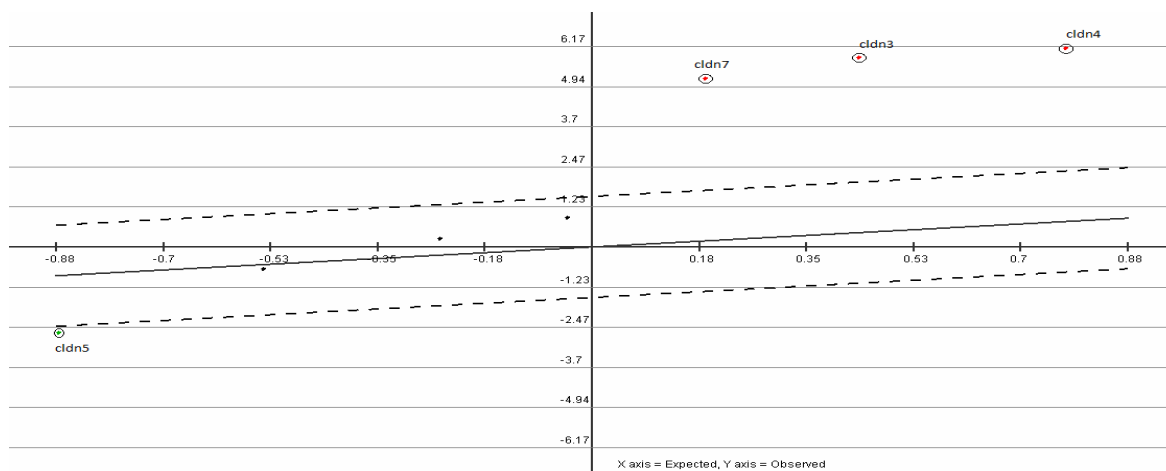


FIGURE 2: SAM GRAPH

4 Discussion

Claudins are tight junction proteins and their involvement in several cancers implicated their role in tumor development. Some of the claudin family members association with the ovarian cancer has already been identified in vitro but this is the first in silico analysis of complete claudin gene family's association specifically with the ovarian cancer. The approach used in this paper is very reliable because the in silico methods of detecting SAGE tags and northern blot analysis are very reliable gene expression methods because they are based upon DNA sequencing. Secondly the microarray analysis was done on the data available on GEO and all this data is experimentally produced data so the chances of error are minimized.

As it is described earlier that three members of claudin family has already been reported for their involvement in expression of ovarian cancer, our results not only verified the significant up regulation of these genes but we also observed the over expression of cldn16 in cancer state. Cldn16 showed up regulation in both SAGE and Microarray analysis. Although the expression of cldn16 is less than the cldn4, cldn3 and cldn7 but its involvement in cancer can never be neglected.

Another interesting finding is the down regulation of cldn5 in microarray data sets, which is very amazing as far as we observe the role of claudins. The tight junction formation ability of claudins makes us to believe their up regulating role in tumor formation but here the association of cldn5 as significant downregulated gene in ovarian cancer is the most surprising thing which reveals the fact that may be our knowledge about the claudins is still very limited and there are many other unraveled truths about the claudins that have to be identified yet. But the SAGE analysis of cldn5 revealed the totally different results by showing its up regulation in cancerous ovary. These findings about the cldn5 makes its role suspicious in ovarian cancer which must be analyzed in vitro.

5 Conclusion

Our work verified the previously known association of claudin members with ovarian cancer. In addition to that our systematic methodology has disclosed the high gene expression level of cldn16 in ovarian cancer which might play an important role in the

cancer cell differentiation and proliferation. Furthermore cldn5 shows down regulated behavior verified by microarray analysis. This needs to be further confirmed, there may be a chance that cldn5 would serve as a novel biomarker for the treatment of ovarian cancer.

6 References

- [1] Cannistra SA. "Cancer of the ovary"; *N Engl J Med*, Vol 351, Page number 2519–29, 2004.
- [2] Morin PJ. "Claudin proteins in human cancer: promising new targets for diagnosis and therapy"; *Cancer Res*, Vol 65, Page numbers 9603-9606, 2000.
- [3] Swisshelm K, Macek R, Kubbies M. "Role of claudins in tumorigenesis"; *Adv Drug Deliv Rev*, 57, 919-928, 2005.
- [4] Nakanishi K, Ogata S, Hiroi S, et al. "Expression of Occludin and Claudins 1, 3, 4, and 7 in Urothelial Carcinoma of the Upper Urinary Tract"; *Am J Clin Pathol*, Vol 130, Page numbers 43-49, 2008.
- [5] Dhawan P, Ahmad R, Chaturvedi R, et al. "Claudin-2 expression increases tumorigenicity of colon cancer cells: role of epidermal growth factor receptor activation"; *Oncogene*, (7 March 2011) | doi:10.1038/onc.2011.43.
- [6] sukita S, Furuse M. "Pores in the wall: claudins constitute tight junction strands containing aqueous pores"; *J Cell Biol*, Vol 149, Page numbers 13-16, 2000.
- [7] Heinzelmann-Schwarz VA, Gardiner-Garden M, Henshall SM., et al. "Overexpression of the Cell Adhesion Molecules DDR1 and Ovarian Cancer Claudin 3, and Ep-CAM in Metaplastic Ovarian Epithelium"; *Clin Cancer Res*, Volume 10, Page numbers 4427-4436, July 6, 2004.
- [8] Boylan K.L.M., Misemer B, DeRycke MS., Andersen JD, et al. "Claudin 4 is differentially expressed between ovarian cancer subtypes and plays a role in spheroid formation"; *International Journal of Molecular Sciences*, Volume 12, Issue 2, Pages 1334-1358, February 2011.
- [9] Hough CD, Sherman-Baust CA, Pizer ES, Montz FJ, Im DD, Rosenshein NB, Cho KR, Riggins GJ, Morin PJ. "Large-Scale Serial Analysis of Gene Expression Reveals Genes Differentially Expressed in Ovarian Cancer"; *Cancer Research*, Vol 60, Page numbers 6281-6287, 2000.
- [10] Rangel LBA, Agarwal R, D'Souza T, Pizer ES, Alò PL, Lancaster WD, Gregoire L, Schwartz DR, Cho KR, Morin PJ. "Tight Junction Proteins Claudin-

3 and Claudin-4 Are Frequently Overexpressed in Ovarian Cancer but Not in Ovarian Cystadenomas”; *Clin Cancer Res*, Vol 9, Page number 2567-2575, 2003.

[11] Hewitt KJ, Agarwal R and Morin PJ. “The claudin gene family: expression in normal and neoplastic tissues”; *BMC Cancer*, 6:186, 2006.

[12] Bignotti E, Tassi RA., Calza S, et al. “Differential gene expression profiles between tumor biopsies and short-term primary cultures of ovarian serous carcinomas: Identification of novel molecular biomarkers for early diagnosis and therapy”.; *Gynecologic Oncology*, Volume 103, Issue 2, Pages 405-416, November 2006.

[13] Rangel LBA, Agarwal R, D'Souza T, et al. “Tight junction proteins claudin-3 and claudin-4 are frequently overexpressed in ovarian cancer but not in ovarian cystadenomas”; *Clin Cancer Res*, Vol 9, Page numbers 2567–75, 2003.

[14] Boon K, Osorio EC, Greenhut SF et al. “An anatomy of normal and malignant gene expression”; *PROC NATL ACAD SCI*, Volume 99, Issue 17, Page numbers 11287-92, Aug 20, 2002.

[15] Hendrix ND, Wu R, Kuick R, Schwartz DR et al. “Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas”; *Cancer Res*, Vol. no. 66, Issue no. 3, Page numbers (1354-62), Fe

SESSION

DRUG TARGETS, IMMUNOTHERAPY, AND COMPUTATIONAL METHODS FOR SYSTEM + MICRO BIOLOGY

Chair(s)

TBA

Classifying HIV-1 Circulating Recombinant Forms

A. Steven Eliuk¹, B. Keith Ruiters², and C. Pierre Boulanger¹

¹Department of Computing Science,

Advanced Man Machine Interface Laboratory (AMMi), University of Alberta, Edmonton, Alberta, Canada

²University of Alberta, Edmonton, Alberta, Canada

Abstract—*The intent of the following paper is to expound on new algorithmic ideas that show marked improvement over formerly state-of-the-art functions in HIV-1 subtyping, such as those found in Wu et al. and NCBI. The paper identifies deficiencies in these older conceptions and sets forth, in a clear and simplistic manner, our improved methodology. The two main boons to the new method described below are the development and utilization of reference profiles and the increased recombination prediction accuracy due to increased branching options and redesigned replacement policies. There is also a new importance placed on absolute prediction accuracy, thus making room for a multitude of real-world possibilities.*

Keywords: Recombination, HIV-1 subtyping, statistical classification

1. Introduction

Human Immunodeficiency Virus type-1 (HIV-1) is incredibly adaptive and diverse. This diversity is caused by a high error rate during transcriptase and a likelihood of recombination [4]. Recombination is the process by which different pure subtypes recombine to form a new strain, in terms of HIV-1, a new circulating recombinant form (CRF) is generated. Understanding recombination, and correctly classifying the pure subtypes that define a CRF, gives the research community the means by which to correctly define the phylogeny of the virus. By understanding the evolution of the virus, the development of effective drug treatments and control vaccines could be possible. Lastly, by correctly classifying an HIV-1 CRF in a host, correct drug treatment could be established, if available for the CRF in question.

Techniques from [11] and [12] and those from NCBI [10], and others [5], [4] using sequence alignment have been very good at predicting the genetic subtypes for an HIV-1 strain, with Wu *et al.* obtaining 100-percent prediction accuracy. However, detection and classification of an HIV-1 CRF is very difficult [11] to attain. Algorithms, such as construction of top strings from *relative entropy*, in order to determine the subtypes of a CRF test sequence and that is proposed by [10], which uses NCBI sliding window to create BLAST similarity scores between reference and testing sequences, have performed reasonably well (obtaining \approx 87-percent and 77-percent prediction accuracy respectively). However one should note, in [11], that the prediction accuracy refers not

to the number of correctly predicted pairs, but the number of correctly predicted subtypes. For example, take testing sequences CRF1-A1F1 and CRF1-A2G1. These two sequences have four subtypes, mainly A1, F1, A2, and G1, [11] only gives the accuracy in terms of correctly defined singles. In this case, a 50-percent prediction accuracy would represent classifying 2/4. Even though it is possible that CRF1-A1F1 was classified as A1 and D, likewise CRF1-A2G1 could be classified as B and G1. Results being 2/4 correctly identified (50-percent) but zero pairs correctly classified. In this paper, *absolute* prediction accuracy will refer to the metric of correctly classified *pairs*, and *relative* prediction accuracy will refer to the metric used in [11] of correctly classified subtypes. Obviously a correctly classified pair provides more information, but for comparative purposes with [11] we will list both *relative* and *absolute* prediction accuracy.

[11] shows great results in terms of *relative* prediction accuracy, achieving the noted 87-percent; however, testing for *absolute* accuracy (complete *pair* subtype match) results in 70-percent prediction accuracy, a remarkable difference. The novelties of our algorithm stem from a quicker implementation of [11] along with changes and improvements in both *relative* and, most importantly, *absolute* prediction accuracy. There are three new techniques implemented, all obtaining improvements in runtime and accuracy; however, all are based on the generation of top strings T , *relative entropy*, and Euclidean distance between reference sequences and test sequences, formally found in Wu *et al.*

The information below will: give a formal description of the methodologies used, the underlying algorithm, and the three subsections defining the main novelty of each algorithm; provide a small section describing the 42-reference sequences used for generation of top strings and the 91 CRF test sequences; a results section, showing results of T on accuracy; and lastly *relative* and *absolute* prediction accuracy of the baseline algorithm from [11] and the three new refinements.

2. Review

2.1 Nucleotide composition string selection in HIV-1 subtyping using whole genomes [11]

The techniques from Wu *et al.* are based on nucleotide composition string selection. This methodology was chosen by Wu *et al.* for a number of reasons. First, it requires

no foreknowledge of the genes being tested. Second, no compression is undertaken which results in fewer errors. Due to the fact that every composition string provides unequal amounts of information to the evolutionary distance calculation, Wu *et al.* noted that by selecting the most important composition strings, those that contribute the most evolutionary data, analysis of thousands of strings can be done in a very affordable manner. This nucleotide composition string selection is a highly effective way to assess HIV-1 recombination and evolution. By selecting the genes that contribute most information to the evolutionary process Wu *et al.* met with impressive results in predicting HIV-1 subtyping. The dataset utilized by Wu *et al.* was composed of 867 pure subtype HIV-1 strains and 331 recombinants. By setting the maximum number of strings at 500 and ensuring string length did not exceed 21, Wu *et al.* attained 100% leave-one-out subtyping accuracy while maintaining computational efficiency. To further test this methodology, Wu *et al.* blindly compared their results to three HIV-1 subtyping programs, again meeting with impressive results.

2.2 Top Strings

A string of nucleotides is generated from a reference sequence in an incremental fashion up to length- K . For example, take the nucleotide sequence AAGC, and length- $K = 3$, the strings constructed would be A, AA, AAG, A, AG, AGC, G, and GC. Notice that the maximum length string is three equaling length- K .

Each string generated from the reference sequences is scored based on *relative entropy* (or Killback-Leibler distance), Equation 1.

$$s(\alpha) = \sum_{i=1}^n |\pi(\alpha, i)| \ln \left| \frac{\pi(\alpha, i)}{\Pi(\alpha)} \right|, \quad (1)$$

where $s(\alpha)$ = *relative entropy* of string α , i = genome i , n = number of whole genomes, $\pi(\alpha, i)$ = absolute composition value for string α in a given genome i , defined in Equation 2, and $\Pi(\alpha)$ = *unnormalized* background probability.

$$\pi(\alpha) = \frac{p(\alpha) - q(\alpha)}{q(\alpha)} \quad (2)$$

where $\pi(\alpha)$ = absolute composition value, $p(\alpha)$ = probability of string α in a given genome, and $q(\alpha)$ = expected appearance of string α defined in Equation 3.

$$q(a_1 a_2 \dots a_k) = \frac{p(a_1 a_2 \dots a_{k-1}) * p(a_2 a_3 \dots a_k)}{p(a_2 a_3 \dots a_{k-1})}, \quad (3)$$

where $p(a_1 a_2 \dots a_{k-1})$ = probability of sub pattern a_1 to a_{k-1} , $p(a_2 a_3 \dots a_k)$ = probability of sub pattern a_2 to a_k , and $p(a_2 a_3 \dots a_{k-1})$ = probability of sub pattern a_2 to a_{k-1} .

2.3 Complete Composition Vector (CCV)

After the scoring and ranking of strings, the top T strings are used to compute a CCV. The vector always has T values and represents the composition values of the top strings in a given genome. Where the vector index i would represent the composition value of the i^{th} top ranked string. String selection and scoring is very important to this technique, with higher scores seeming to contain richer information [11], [12]. Generating the selected string composition vector is rather simple. If there are less than 500 strings, add the current string in question. If not, and the current string has a higher score, a larger absolute relative entropy, then the lowest score is replaced. This technique of only storing the richest 500 strings basically resolves all memory issues according to [11]. Once all the strings have been examined a 500-dimensional composition vector is built. For example, testing in [11] included the use of 42 reference whole genomes, 331 recombinant, and 825 pure subtype whole genomes. 500 top ranked strings were used, in turn producing a 500-dimensional composition vector. The technique was 100% successful in the subtyping of the 825 pure subtypes. Most importantly, the technique does not rely on prior knowledge about the genomic sequences.

2.4 Pair-wise distance

Given a pair of genomes, a and b , the distance between them can be represented as the Euclidean *distance* between their respective CCV's as seen in Equation 4.

$$distance = \left(\sum_{l=1}^m (a_l - b_l)^2 \right)^{1/2} \quad (4)$$

2.5 Basic Local Alignment Search Tool (BLAST)

BLAST is a widely used method for comparison between nucleotide and protein sequences. It is used to determine relative relationships between test and reference strains [6]. BLAST is such an effective tool because of its speed and ease of use; however, it is victim to one downfall, namely that, because of its high speed, its optimality cannot be guaranteed in alignment. This large speedup, approximately fifty times faster than conventional optimal algorithms, is made possible by a simple heuristic. Using this heuristic ensures high computing speed while maintaining quality results and high accuracy. More information about the specifics of BLAST can be found at [6]. BLAST is a useful tool in the analysis of recombination in HIV, such as being able to compare a test strain against known reference strains using BLAST, in order to classify the test strain. After utilizing BLAST, the results can show a high probability of belonging to a certain clade, being recombinant, or being a pure subtype. If the results show the test strain belongs to a certain clade, a drug treatment that is specific to this clade can be administered for a more effective treatment.

2.6 National Centre for Biotechnology (NCBI) algorithm using Scored BLAST

NCBI, being considered a state-of-the-art institution [10] in recombination detection prior to 2007, utilizes a technique that uses a score based BLAST [6] pairwise alignment between overlapping segments. This alignment is carried out between a query sequence and a known reference sequence. The algorithm moves a window along the query sequence, processing each window segment separately while comparing each against the reference sequences using BLAST. BLAST returns a similarity score for each local alignment [10]. The reference sequence that matches with the highest similarity score is assigned for the local alignment. The process is repeated for each window and recorded. Once the comparisons are completed, if a single genotype is assigned to most segments, the query segment is considered a single genotype and classified accordingly. If multiple genotypes were recorded during local alignment and the percentage belonging to each genotype is higher than a predefined threshold, the query sequence is deemed recombinant. This process could easily be used to speculate the most probable breakpoint for recombination [10] because the location of divergence is easily seen when local alignment produces a new reference sequence and they match continually. The three parameters that govern the NCBI method are: the choice of window size, often experiment specific; the incremental step, defined as the amount the window is shifted along the sequence; and the similarity threshold, defined as the percentage of non-primary genotypes that can be recorded before recombination is considered, for a match. The NCBI method is impressively simple and the results it yields are among the best when detecting recombination. Tests of 48 reference sequences [10] were used to predict recombinant deterministic forms. NCBI was able to obtain a 73.4% prediction accuracy where later CCV tests only yielded 66.2% prediction accuracy using the same reference sequences. This method was able to accurately predict all but two CRF12BF strains, namely AY771588 and AY771589. The techniques of [11] were tested on the 91 strains that have deterministic recombinant forms and was able to determine 87.3% accuracy. Likewise, the 42 known reference sequences were used; however, 5000 top ranked strings were used vs. the 500 top ranked strings used in pure subtyping. The results were a substantial increase over those of NCBI.

2.7 Detecting subtypes in CRF

Difficulty arises when trying to compute the pure subtypes that make up a CRF. There is no guarantee the breakpoint is consistent and it likely varies. Therefore, Wu *et al.* suggests breaking a sequence into equal parts. At each testing, a consecutive number of parts are removed and the remaining concatenated together. For example, take a partitioning factor $P = 50$, a CRF genome would be broken into ≈ 180

nucleotides (9000 nucleotides / 50 = 180). A maximum l parts can be removed, $l \approx P/2$ seems to work well in empirical testing.

Given a partitioning factor of $P = 50$, and if $1 \leq l \leq 25$ parts can be removed, we would construct the following test strings.

$$\begin{aligned}
 & l = 1, \\
 & s_1 = (p_2 \dots p_{50}), \\
 & s_2 = (p_1, p_3 \dots p_{50}), \\
 & \dots \\
 & s_{49} = (p_1 \dots p_{48}, p_{50}), \\
 & s_{50} = (p_1 \dots p_{49}). \\
 & l = 2, \\
 & s_1 = (p_3 \dots p_{50}), \\
 & s_2 = (p_1, p_4 \dots p_{50}), \\
 & \dots \\
 & s_{47} = (p_1 \dots p_{47}, p_{50}), \\
 & s_{48} = (p_1 \dots p_{48}). \\
 & \vdots \\
 & \vdots \\
 & l = 25, \\
 & s_1 = (p_{26} \dots p_{50}), \\
 & s_2 = (p_1, p_{27} \dots p_{50}), \\
 & \dots \\
 & s_{24} = (p_1 \dots p_{24}, p_{50}), \\
 & s_{25} = (p_1 \dots p_{25}).
 \end{aligned}$$

In all, 950 strings are constructed. For each test string the CCV is generated and the Euclidean distance between the test string and the reference CCVs are calculated, see Equation 4. The two reference sequences, that, when compared against the test sequence, produced the lowest scores are recorded. In all, 1900 reference sequences would be stored. The frequency of a reference sequence can be thought of as the amount of the test genome that belongs to a specific reference sequence; in turn, a specific pure subtype. The two reference sequences with the greatest frequency are reported as the two predicted pure subtypes of the test CRF sequence.

2.8 Conclusion

The base technique from Wu *et al.* is seen in many different areas of computer-based learning. The algorithm breaks down into a learning stage, a metric between learned top ranked strings and reference sequences; distance is then computed between test and reference sequences before the minimum distances between reference and test data is finally associated with the most probable match. Many enhancements are possible, such as using the ordering of top ranked strings as a weight metric. Giving a higher weight to the very best strings and decreasing accordingly as lower ranked strings are used. Likewise, different distance metrics can be used when comparing test to reference sequences and; furthermore, the metric used to score a string can be replaced with a variety of other metrics. As with most

Replacement Policy	
R	[A or G]
Y	[T or C]
K	[G or T]
M	[A or C]
S	[G or C]
W	[A or T]

Fig. 1: Nucleotide Replacement Policy, see [1]

Replacement Policy	
B	[C or G or T]
D	[A or G or T]
H	[A or C or T]
V	[A or C or G]
N	[A or C or G or T]

Fig. 2: Complex Nucleotide Replacement Policy, see [1]

learning techniques, the metrics or kernels used are often application or class-of-problem specific – more testing in this area is needed and enhancement in predicting CRFs is probable.

3. Methodologies

3.1 Nucleotide Replacement Policy - Alg. 1

The reference sequences used to construct the top strings T often contain questionable nucleotides. Frequently these nucleotides are ignored, as in Wu *et al.* However, by ignoring these nucleotides, it is possible that important strings or patterns could be lost. Algorithm 1 focuses on replacing these questionable nucleotides as seen in Figure 1, based on internationally agreed standards outlined in [1]. During string generation, when one of these questionable nucleotides is seen, it is replaced with two possible occurrences. Most importantly, because we are not incrementing the occurrence of substrings for the newly generated strings, the probability calculations are still accurate.

3.2 Complex Nucleotide Replacement Policy - Alg. 2

The reference sequences used to construct the top strings T often contain complex questionable nucleotides. These are nucleotides that have > 2 possible replacements. Likewise, we are never incrementing subpatterns of the newly formed strings so the probability distributions are still accurate. The replacement policy used can be seen in Figure 2, which are also based on internationally agreed standards [1]. For testing purposes, algorithm 2 also uses the simple replacement policy seen in the previous section.

Reference Distribution		
6	subtype A	4 A1 and 2 A2
4	subtype B	4 B1
4	subtype C	4 C1
3	subtype D	4 D1
8	subtype F	4 F1 and 4 F2
3	subtype H	3 H1
3	subtype G	3 G1
2	subtype K	2 K1
3	subtype N	3 N1
2	subtype J	2 J1
4	subtype O	4 O1

Fig. 3: Pure subtype distribution in 42 reference sequence database

3.3 Reference Profiles - Alg. 3

Creating the top strings T has a small disadvantage to strings or patterns seen in the same subtypes. For instance, say a string was seen in four pure subtype reference sequences. We would like a way to emphasize this occurrence, rather than the marginal increment it would get using the standard *relative entropy* calculation. In the simplest form, we combined the reference sequences into pure subtype profiles. In all, 13 reference profiles were constructed, representing 13 pure subtypes. This provided an increased *relative entropy* score for regularly seen strings/patterns in the same subtype. Reference profiles use both simple and complex nucleotide replacement policies as described in the previous sections.

4. Pure Subtype and CRF Databases

Although many techniques use simulated data, we believe using actual data is more realistic regarding the natural diversity found in HIV-1, in terms of recombination and pure subtype reference sequences. With this consideration in mind, we focus testing entirely on the datasets used in [11]. This makes comparison between algorithms easier and prior results from [11] can be examined directly. Lastly, the generation of good testing data is difficult to achieve. The issues surrounding data acquisition are mainly the complex nature of naturally occurring recombinant forms and how to simulate them. For instance, there is often multiple breakpoints in a strain and non-reciprocal exchange [7], [8], [9], which is very hard to reproduce. Therefore, we focus on test data previously classified and internationally used for recombinant form classification, mainly those found in [11].

4.1 Reference Pure Subtype Sequences

42 pure subtype reference sequences are used to construct the top ranked strings. The distribution of the 42 reference sequences can be seen in Figure 3.

Test CRF Distribution	
52	subtype A1 and G1
3	subtype A1 and B1
3	subtype D1 and F1
11	subtype B1 and C1
3	subtype C1 and D1
10	subtype B1 and F1
7	subtype B1 and G1
2	subtype A2 and D1

Fig. 4: 91 unique test sequences and respective compositions

4.2 CRF Test Sequences

91 deterministic CRF test sequences are used. These test sequences are well-documented and the respective pure subtypes are well-defined and accepted. The distribution of the 91 test sequences can be seen in Figure 4. Most importantly, not all pure subtypes are seen in the 91 test sequences; however, all pure subtypes are used during the training stage of the algorithm. Better results can be obtained by narrowing the training stage to only those reference sequences that are present in the CRF test sequences. However, the goal of the research is to construct a method to reliably predict recombinant forms (pure subtypes that define the CRF) from test sequences where there is no knowledge of the phylogeny of the sequence. Therefore, all pure subtype reference sequences are always used regardless of the specifics that may be known about the test data. Lastly, throughout all testing, the knowledge of what pure subtypes make up a given CRF is never used, only during verification of the prediction.

5. Results

Overall, some notably important results were obtained. Chiefly, a quicker runtime was realized, a limit has been found for top string count, and better prediction accuracy in terms of both *relative* and *absolute* accuracy for all algorithms was achieved.

5.1 Limits on number of Top Strings

Figures 5 and 6 clearly show that, as T grows from 0 to 5000, prediction accuracy steadily improves. As T grows larger than 5000 one can see accuracy, conversely, drops. These results counter suggestions in [11] that the greater the size of T the greater the knowledge contained in T.

5.2 Relative prediction accuracy

Previous results from [11] show 87-percent prediction accuracy and using [10] NCBI with a sliding window and BLAST comparative scores, obtained 77-percent prediction accuracy. Simple nucleotide replacement policy resulted in 88-percent prediction accuracy and complex nucleotide replacement policy resulted in 90-percent prediction accuracy. These results are rather impressive on their own and should

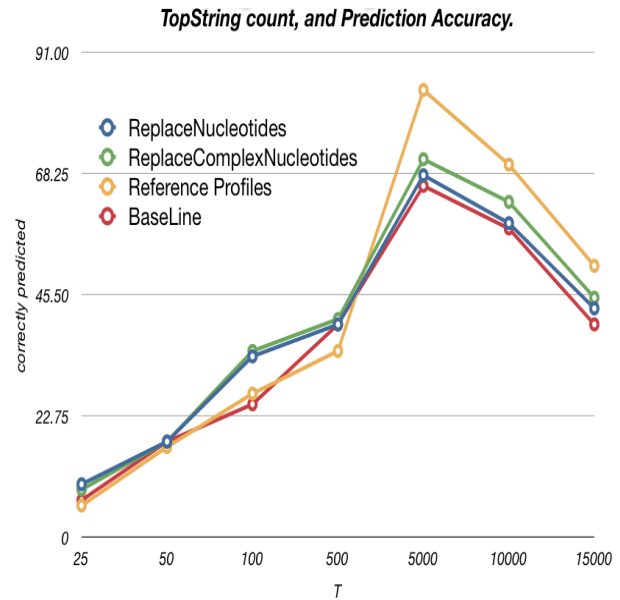


Fig. 5: Accuracy improves up to $T \approx 5000$, decreases steadily as $T > 5000$.

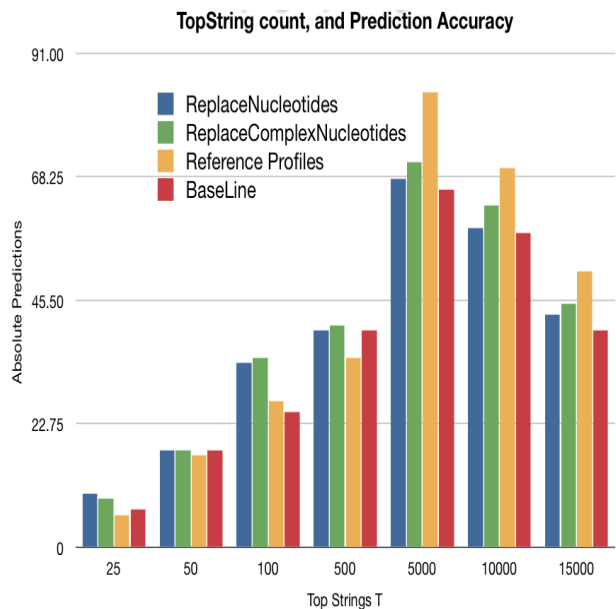


Fig. 6: Accuracy improves up to $T \approx 5000$, decreases steadily as $T > 5000$.

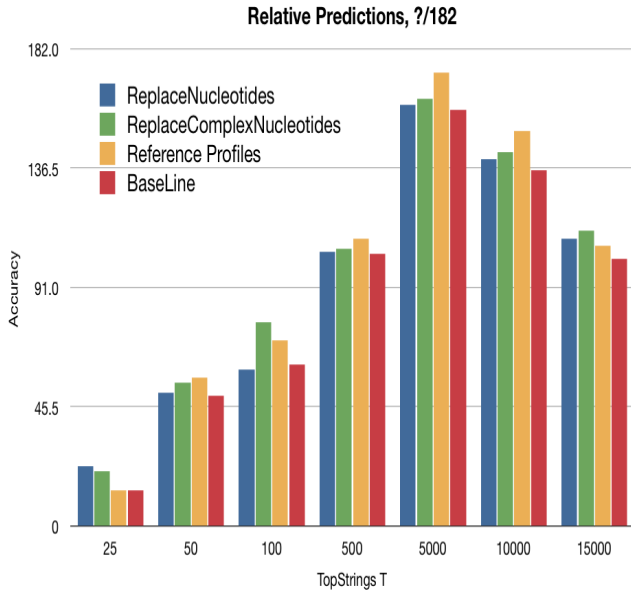


Fig. 7: *Relative* prediction accuracy in terms of 182 subtypes

not be overshadowed by the results from the third algorithm reference profiles. Obviously, there is knowledge contained in these areas of questionable nucleotides, the fact the top strings changed dramatically depending on the use of simple and complex nucleotide replacement policies show us this. However, the third algorithm shows extraordinary results, achieving 95-percent accuracy, see Figure 8. This is likely because when the *relative entropy* is calculated for a string, the strings are given a slight boost because they are seen in the foreground distribution more than the background. The boost is only slight, but works well experimentally.

5.3 Absolute prediction accuracy

Absolute prediction accuracy is an important metric because it not only tells us how many pure subtypes we predicted correctly, but it also reports many correctly predicted pairs were obtained. This is ultimately the goal: predict the makeup of a CRF with high precision. Previous algorithms demonstrated only marginal accuracy, as in Wu et al., where even our simple and complex nucleotide replacement policies show a respectable gain in terms of *relative* accuracy, fare much better in terms of *absolute* accuracy. For instance, looking at Figures 9 and 10, we see the simple nucleotide replacement policy predicts three more pairs correctly, and complex nucleotide replacement predicts five more pairs correctly, when compared to the baseline algorithm that predicts only 66/91. These results show clearly, like that shown in *relative* prediction accuracy, that information is gained when using the replacement policies. This information results in new strings in our top strings list that were never available previously. *Absolute* prediction accuracy was never included

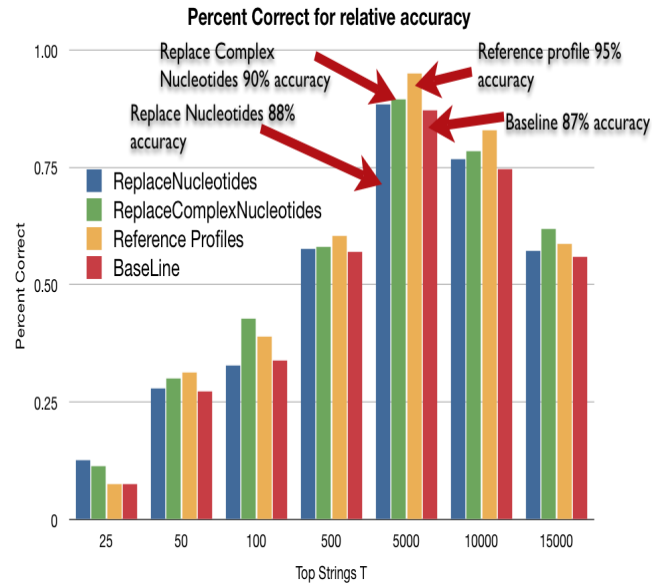


Fig. 8: *Relative* prediction accuracy percent correct.

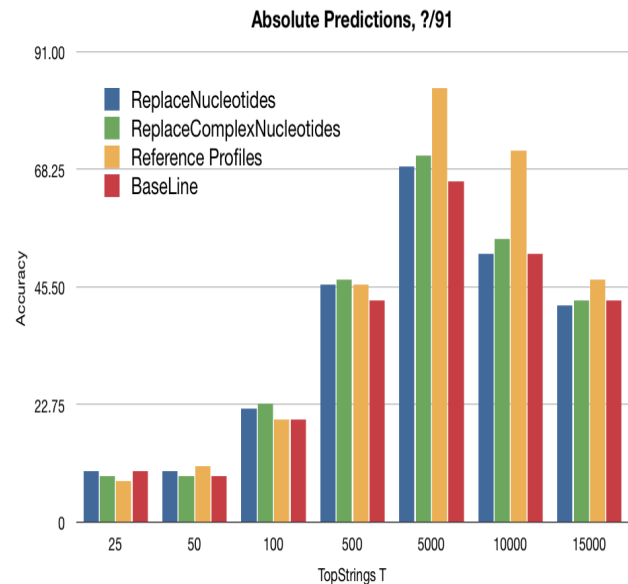


Fig. 9: *Absolute* prediction accuracy in terms of pairs correctly labelled.

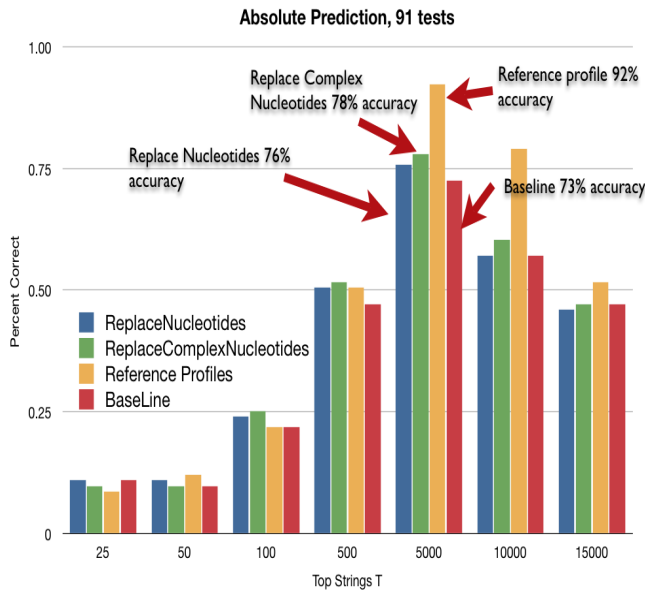


Fig. 10: *Absolute* prediction accuracy percent correct.

in [11] for NCBI tool and, in turn, are not included.

Reference profiles show a clear advantage and rather impressive results, with 84/91 correctly predicted pairs and 92-percent overall *absolute* prediction accuracy. Clearly this technique works well on the 91 deterministic CRF database.

5.4 Considerations

These results show our techniques perform well under the 42 reference sequences used and the 91 deterministic CRFs; however, there really is not much deterministic CRF data available. Even the database of 91 CRFs likely has some error and that could be in our benefit or not. In the future we hope to obtain more reliable datasets, not simulated data, but real-world CRFs that have been carefully examined to define the composition of pure subtypes. We hope the real-world data will further validate our method. Likewise, it is interesting to see that there is an upper bound to $T \approx 5000$ and counters previous thoughts that more strings would contain more information [11]. Some other results that were not included in this paper are, chiefly, the results of $T > 5000$ can be improved marginally if we restrict the length- k of strings to ≈ 14 . However, this only resulted in a small improvement and therefore, was not formally displayed. One can infer that shorter length strings are more important in classification, even though longer strings often can show more information.

The ability to calculate the composition of the test sequence is very important, and because we are not limited to only two results per test sequence we can easily give composition based certainty that our algorithm uses for classification. We have seen many examples that show five or

more base type ancestry. There is also the ability to do inter-clade analysis after the initial classification is done using the same algorithm. This is very important because one often wishes to know the composition outside of the reference profiles that were created by joining pure subtypes.

Lastly, all results are available online, see [2]. We invite any and all suggestions and also look forward to testing other research groups' data, whether HIV-1 specific or not.

Conclusion and FutureWork

[11] provides a novel starting point based on a general machine learning framework used in bioinformatics. We have shown a substantial increase in terms of both *relative* and *absolute* prediction accuracy in all of our algorithms. The goal of our research is to build a tool that gives high certainty results concerning the makeup of a CRF. These results can lead to more accurate HIV-1 phylogeny and the development of widely applicable treatments that are more adaptive to recombinant forms of HIV-1. Further testing is needed to validate the results in this paper, we will continue to refine our algorithm and as more deterministic and reliable data becomes available we hope to have a sound method for detection of recombination and classification or pure subtypes in the sequence.

References

- [1] J. Biochem. Nomenclature for incompletely specified bases in nucleic acid sequences. *Biochem Journal*, 229:281D286, 1985.
- [2] S. Eliuk. <http://www.cs.ualberta.ca/~eliuk/recombinationresults/>.
- [3] S. Eliuk, P. Boulanger, and K. Kabin. Sunviz: A real-time visualization environment for space physics applications. In *ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II*, pages 1–11, Berlin, Heidelberg, 2008. Springer-Verlag.
- [4] D. P. Martin, C. Williamson, and D. Posada. Rdp2: recombination detection and analysis from sequence alignments. *Bioinformatics*, 21(2):260–262, 2005.
- [5] I. Milne, F. Wright, G. Rowe, D. Marshall, D. Husmeier, and G. McGuire. Topali: software for automatic identification of recombinant sequences within dna multiple alignments. *PubMed*, 20(11):1806–7, 2004.
- [6] E Myers, S Altschul, W Gish, D Lipman, and W Miller. <http://blast.ncbi.nlm.nih.gov/blast.cgi>.
- [7] D. Posada. Evaluation of methods for detecting recombination from dna sequences: Empirical data. *Mol Biol Evol*, 19(5):708–717, 2002.
- [8] D. Posada and KA. Crandall. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, 54(3):396–402, 2002.
- [9] D. Posada and KA. Crandall. Evaluation of methods for detecting recombination from dna sequences: Computer simulations. *Proc Natl Acad Sci USA*, 98(24):13757–62, 2002.
- [10] Mikhail Rozanov, Uwe Plikat, Colombe Chappey, Andrey Kochergin, and Tatiana A. Tatusova. A web-based genotyping resource for viral sequences. *Nucleic Acids Research*, 32(Web-Server-Issue):654–659, 2004.
- [11] Xiaomeng Wu, Zhipeng Cai, Xiu-Feng Wan, Tin Hoang, Randy Goebel, and Guohui Lin. Nucleotide composition string selection in hiv-1 subtyping using whole genomes. *Bioinformatics*, 23(14):1744–1752, 2007.
- [12] Xiaomeng Wu, Xiu feng Wan, Gang Wu, Dong Xu, and Guohui Lin. Whole genome phylogeny via complete composition vectors, 2004.

Self-Regulating Physiologically Based Pharmacokinetic Model and Creation of Drug Concentration Profiles in Plasma and Tissues

Stanislav Polinkevych

Recherche et Education en Mathematique,
101 Meaney Rue, Kirkland, Quebec, H9J 3B9, Canada

Abstract – *A physiologically based pharmacokinetic model is build to determine the dynamics of drug (compound) concentration in the human body. The model consists of two major subsystems. The first subsystem simulates the diffusion of the drug(s) and respiratory gases between plasma and the tissues. Second subsystem controls the processes of the drug and gas delivery to the tissues. The system of control is based on the principles of optimal control theory and the mechanisms of self-regulation. The model allows simulation of a combined influence of multiple clearance factors. The drug is administered intravenously into the human body and goes through phases of Absorption, Distribution, Metabolism, and Excretion (ADME). The results of numerical calculations of drug concentration profiles under renal and hepatic clearance are reported. The model can be tailored to suit the experimental needs in the fields of pharmacological and medical research.*

Keywords: pharmacokinetics, drug, dynamics, ADME, model, profile

1 Introduction

Drug development is a costly and time consuming process [1]. To reduce expenditure of this process, multiple analytical methods and tools are currently used. Attempts were made to integrate the compartments and build physiologically based models [2]. A variety of mathematical models and software tools was created and is available on market today [2]. Most of them are based on traditional concepts of pharmacokinetic modeling [2, 3]. However, physiological regulation mechanisms of the internal state of the system (organism) and the mechanisms of the drug transportation were not modeled.

The complexity of the mechanisms of regulation in the living organism required more sophisticated models (tools) which could represent the organism as a single system. The most important physiological mechanisms and principles of regulation of the drug transport system had to be properly incorporated into the model.

A new approach to the drug kinetics model creation with the use of Functional Respiratory System as a major system of

drug transportation was offered [4]. A brief overview of the model simulating renal clearance was given.

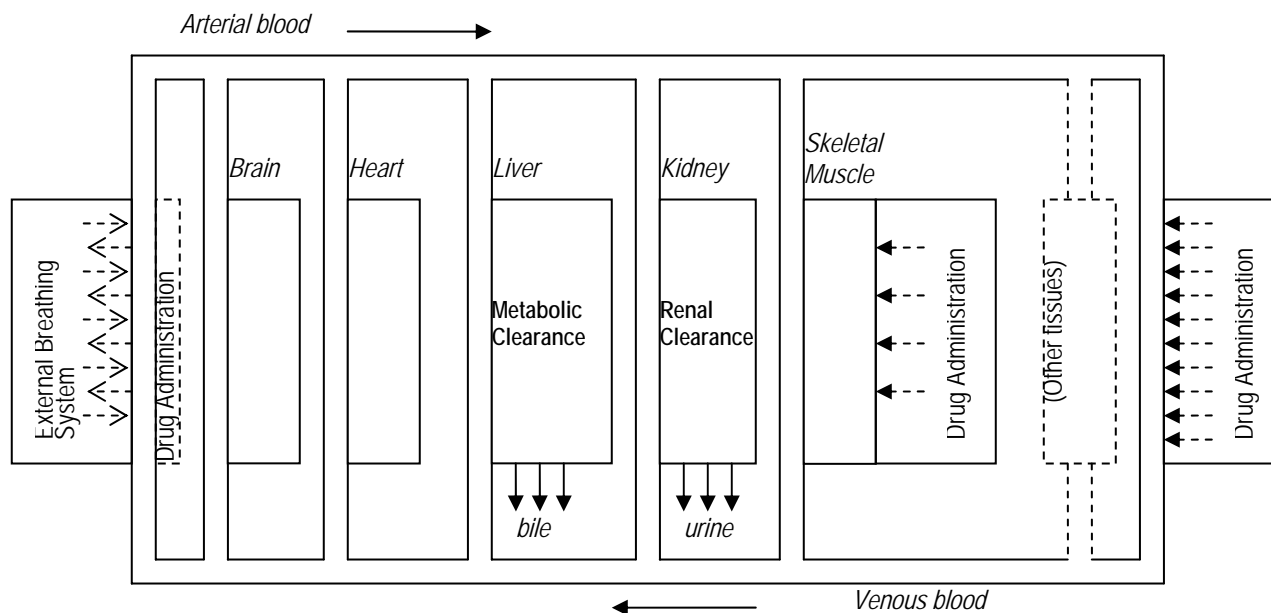
In this paper a new model for drug kinetics simulation is introduced. This model is built on the principles described in [4] and implements the simultaneous influence of two drug elimination factors – renal and hepatic clearance. The control of the model is based on the principles of optimal control theory and the mechanisms of self-regulation.

The model allows calculating drug concentration dynamics in plasma and the tissues of the organism, as well the distribution of the dose among the tissues and the re-distribution and elimination of the drug. The model represents a self-regulating system, and is built on modern knowledge of the cardiovascular and respiratory systems. As such, the simulated drug dynamics clearly reflect changes in the internal and external environment of the organism. Note, that traditional pharmacokinetics models attempt to predict the drug concentration under steady-state conditions. However, the newly developed model presented in this paper allows simulating the drug concentration dynamics in the organism in both – steady and disturbed states. A disturbed state is induced by modeling of internal and external stress factors exerted on a steady state of the system (organism).

This model is proven to be a dynamic and adaptable tool to build accurate as well as extensive drug concentration profiles in both plasma and tissues. It allows simulating multiple dosage regimens; determining effective and viable dosage gradients; rates and methods of administration and to simulate various routes of the drug elimination, including renal clearance and multiple schemes of metabolism.

2 General description of the multi-compartmental model

Organ tissues of the body are represented by m compartments, between which the drug d is transported via the closed circulatory system. The compartments represent the following tissues: cardiac, cerebral, hepatic, renal, skeletal muscle, and skin. An additional compartment is allocated for the remaining tissues. The distribution of the drug and respiratory gases through tissue capillaries is represented by a network of pipelines through which arterial and venous blood is pumped.



Scheme 1: Joined multi-compartmental system

The general scheme of the circulatory system is shown in Scheme 1. The sites of administration and routes of drug distribution as well as the compartments chiefly responsible for drug elimination are shown.

3 The system of drug and gas transport and exchange in the compartments

The system consists of a set of differential equations governing drug concentrations in the capillaries and the tissues of the compartments, and a separate set of differential equations that governs the tensions of respiratory gases – O₂ and CO₂.

In this paper, intravenous blood infusion is considered. The model can be adapted to other routes of administration.

Concentration of the drug *d* in the plasma (*c_{dct_i}*) of tissue capillaries is represented by the following set of differential equations:

$$V_{ct_i} \frac{dc_{dct_i}}{d\tau} = Q_{t_i} c_{fa} - G_{dt_i} - Q_{t_i} c_{dct_i}, \quad (1)$$

$$i = \overline{1, m},$$

Index *i* is assigned to a corresponding tissue, *V_{ct_i}*

 - the volume of the blood in the capillary of the tissue *i*.

Modeling of renal clearance is performed through modification of equation (1).

Concentration of the drug in the tissues (*c_{dt_i}*) is calculated by:

$$V_{t_i} \frac{dc_{dt_i}}{d\tau} = G_{dt_i} - Q_d c_{dt_i} \quad i = \overline{1, m}, \quad (2)$$

The multi-compartmental system is governed by the self-regulating control system built on the principles of Optimal Control Theory [9].

Where *V_{t_i}*

 - the volume of the tissue *i*, *G_{dt_i}* - the flow of the drug from the capillary of the tissue *i* into the tissue, *Q_{t_i}* - the blood flow through the capillary of the tissue *i*, *Q_d* - the rate of the drug clearance.

G_{dt_i}

 is calculated by the formula:

$$G_{dt_i} = D_{dt_i} S_{t_i} (c_{dct_i} - c_{dt_i}) \quad (3)$$

D_{dt_i}

 is the value of the diffusion coefficient, *S_{t_i}* - the area of the surface of the diffusion.

The equation of the drug concentration in venous blood is calculated by:

$$V_v \frac{dc_{d\bar{v}}}{d\tau} = \sum_{t_i} (Q_{t_i} c_{dct_i} + Q_{dt_i}) - Q c_{d\bar{v}} \quad (4)$$

where $Q = \sum_i Q_{t_i}, i = \overline{1, m}, V_v$ - the volume of the venous blood.

The equations for the tension of oxygen (*p_{1ct_i}*) and carbon dioxide (*p_{2ct_i}*) in the blood of tissue (compartment) capillaries are [5]:

$$\frac{dp_{1ct_i}}{d\tau} = \frac{1}{\alpha_1 V_{ct_i}} (\alpha_1 Q_i (p_{1a} - p_{1ct_i}) \quad (5)$$

$$+ \gamma Hb Q_i (\eta_a - \eta_{ct_i}) - G_{1t_i} - \frac{d\eta_{ct_i}}{d\tau} \gamma Hb \cdot V_{ct_i} \Big),$$

$$\frac{dp_{2ct_i}}{d\tau} = \frac{1}{\alpha_2 V_{ct_i}} (\alpha_2 Q_i (p_{2a} - p_{2ct_i}) + \gamma_{BH} BH Q_i (z_a - z_{ct_i}) - G_{2t_i} + \gamma Hb Q_i (z_a (1 - \eta_a) - z_{ct_i} (1 - \eta_{ct_i}))) + \gamma Hb V_{ct_i} \frac{d\eta_{ct_i}}{d\tau} + \gamma_{BH} BH V_{ct_i} \frac{dz_{ct_i}}{d\tau} \Big). \quad (6)$$

Where

$$\eta_{ct_i} = 1 - 1,75 \exp(-0,048 m_{cL} p_{1cL}) + 0,75 \exp(-0,12 m_{cL} p_{1cL}) \quad (7)$$

$$m_{ct_i} = \delta (pH_{cL} - 7,4) + 1, \quad (8)$$

$$pH_{ct_i} = 6,1 + \lg \frac{BH}{\alpha_2 p_{2cL}}, \quad (9)$$

$$z_{ct_i} = \frac{p_{2ct_i}}{p_{2ct_i} + 30}, \quad (10)$$

In equations (5)-(10) α - solubility of corresponding gas, pH is the acidity of blood, BH is the concentration of enzyme carbonic anhydrates, p_{1a} - tension of oxygen (p_{2a} - carbon dioxide) in arterial blood, η_{ct_i} is the degree of saturation of hemoglobin with oxygen, S - the area of the surface of alveoli capillaries in a compartment, q - consumption of oxygen (index 1) or production of carbon dioxide (index 2).

In the compartments (tissues) the dynamics of the parameters of the model are defined by equations:

$$\frac{dp_{1t_i}}{d\tau} = \frac{1}{V_{t_i} \left(\alpha_{1t_i} + \gamma_{Mb} Mb V_{t_i} \frac{\partial \eta_{Mb_i}}{\partial P_{1t_i}} \right)} (G_{1t_i} - \dot{q}_{1t_i}) \quad (11)$$

$$\frac{dp_{2t_i}}{d\tau} = \frac{1}{\alpha_{2t_i} V_{t_i}} (G_{2t_i} + \dot{q}_{2t_i}) \quad (12)$$

where

$$\eta_{Mb_i} = 1 - \exp(-0,12 p_{1t_i}) \quad (13)$$

The flows of gases through alveoli-capillary membranes are calculated based on

$$G_{t_i} = D_{t_i} S_{t_i} (P_{ct_i} - P_{t_i}), \quad (14)$$

D_{t_i} - diffusion coefficient of the corresponding gas.

Scheme 1 and correspondingly, the model can be modified for other possible ways of drug administration.

4 Description of the self-regulating control system

Volumetric blood flows in tissue capillaries Q_i are considered as control parameters. Then, the general criterion of control of the system (1)-(13) is given as the cost functional:

$$I = \int_{\tau_0}^{\infty} \sum_{i=1}^m K_i \left\{ (G_{1t_i} - q_{1t_i})^2 + (G_{2t_i} + q_{2t_i})^2 \right\} d\tau. \quad (15)$$

K_i - coefficients dependent on the size and the type of the corresponding compartment.

The task of control of the system (3)-(15) is formulated as the transformation of disturbed trajectories of the system (1)-(12) into the area of attraction of the stationary solution (equilibrium point - if a drug administration is not modeled by a periodic function), defined by inequalities:

$$|G_{1t_i} - q_{1t_i}| \leq \varepsilon_1, \quad (16)$$

$$|q_{2t_i} + G_{2t_i}| \leq \varepsilon_2. \quad (17)$$

The process of control of the system (3)-(15) is provided by the changes of parameters Q_i in order to minimize the functional (13). The values of Q_i are calculated using the methods of Optimal Control Theory [9].

If system (1)-(12) is disturbed by the changes in external or internal environments, then the new homeostatic state is determined, and the trajectories of the system (1)-(12) are transferred in the area defined by conditions (16)-(17).

5 Drug concentration profiles calculated by the model for intravenous drug administration

The model was adapted to reflect the characteristics of an average human being of 75 kg weight, including the surface of drug and gas exchange. The steady state of the system was characterized by the oxygen consumption of 4.3 ml/sec. Tensions of respiratory gases in arterial and venous blood were kept constant; O_2 arterial tension was equal to 95 mmHg, CO_2 arterial tension was equal to 42 mmHg.

To conduct the numerical experiments with the model (1)-(17), 200mg of the drug d were introduced intravenously every 5 hours within a 25 hour period. Several series of experiments were conducted with the model. First, the drug distribution and its diffusion to the compartments was simulated under the conditions of no clearance and no physical stress. Thus, benchmark values of the drug dynamics were set, upon which the trajectories of the system with renal and hepatic clearance would be evaluated.

Second, the calculations with model were performed under the effects of renal clearance exclusively.

The last series of experiments simulated both renal and hepatic methods of drug clearance.

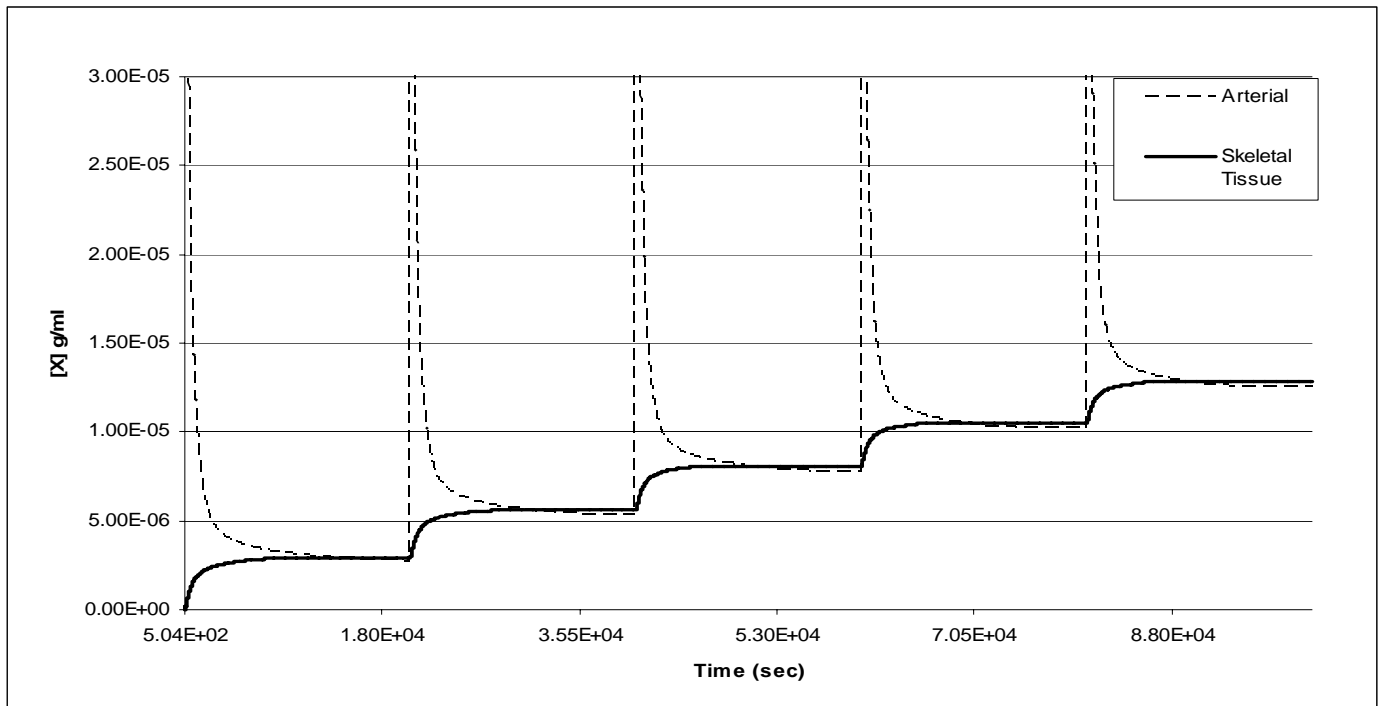


Figure 1A: Drug d Dynamics in Arterial Blood and Skeletal Muscle Tissue

Figures 1A, 1B and 1C display the trajectories of the drug concentration in skeletal, cardiac muscles, cerebral tissue, and arterial blood. The evident peaks on the trajectory that correspond to the drug concentration in arterial blood, show the moments of drug infusion into venous blood. Soon after infusion, the trajectory descends to a much lower level due to the rapid dose distribution to the tissue capillaries and

to the tissues themselves. The trajectories of the drug concentration in skeletal muscle (Figure 1A), cardiac muscle (Figure 1B) and cerebral tissue compartments (Figure 1C) show the dynamics of drug accumulation in the corresponding compartment.

The rate of the infusion of the drug can be regulated in the model (1)-(17) according to the dosage requirements.

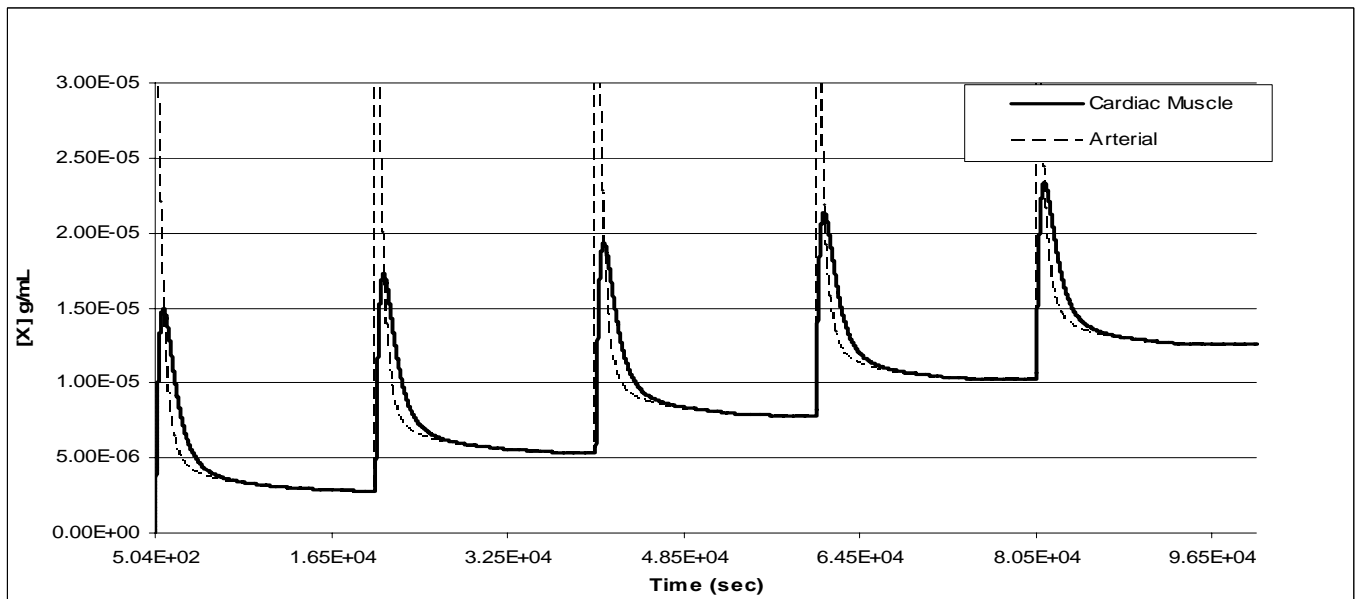


Figure 1B: Drug d Dynamics in Arterial Blood and Cardiac Muscle

The evident peaks on the trajectory of the cardiac muscle are observed due to a higher ratio between the volume of the capillaries ($V_{c(i)}$) in the cardiac muscle and the volume of the

tissue compartment in comparison to the same ratio in skeletal muscle or cerebral tissue. Once the concentration of the drug in the capillaries becomes lower than the concentration of the

drug in the tissue, the diffusion of the drug back to the capillaries from the tissue begins.

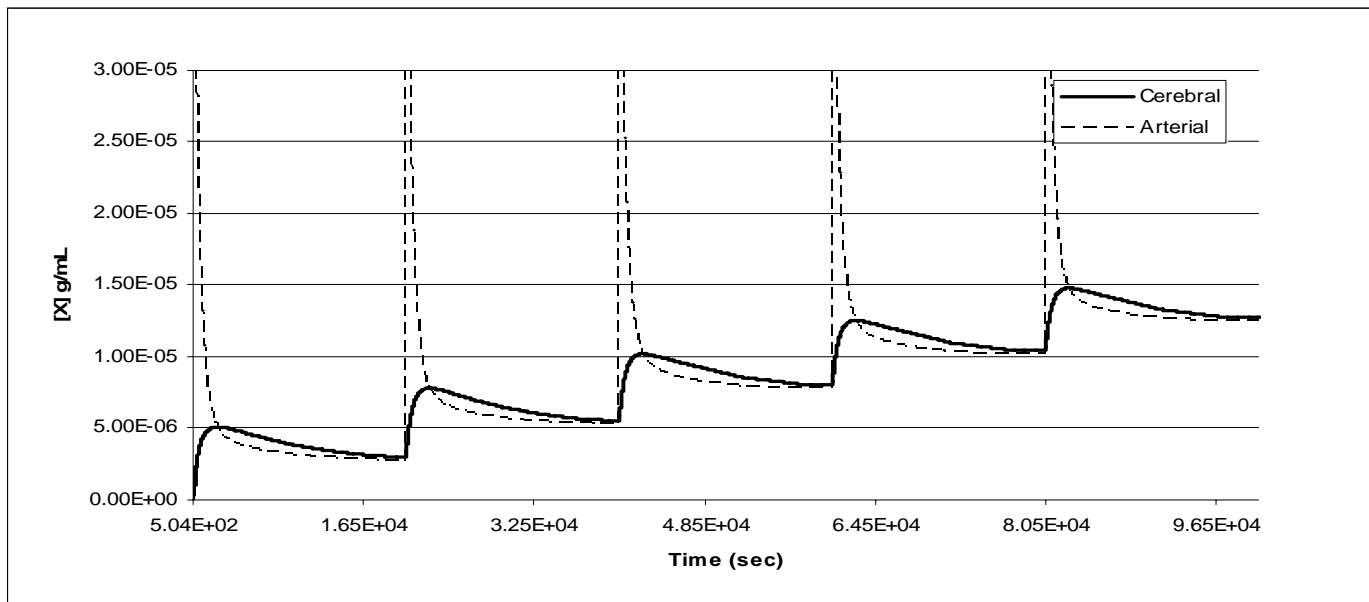


Figure 1C: Drug d Dynamics in Arterial Blood and Cerebral Tissue

Figure (2) presents the drug dynamics within skeletal muscle tissue under varying degrees of renal clearance in a steady state. It is apparent, that the dynamics of the drug concentration trajectory changes significantly. The calculations show significant decrease in the levels of the drug

concentration in the skeletal muscle. Under the same rate of infusion, lower clearance rates show more rapid drug accumulation in the skeletal muscle, than under higher clearance rates.

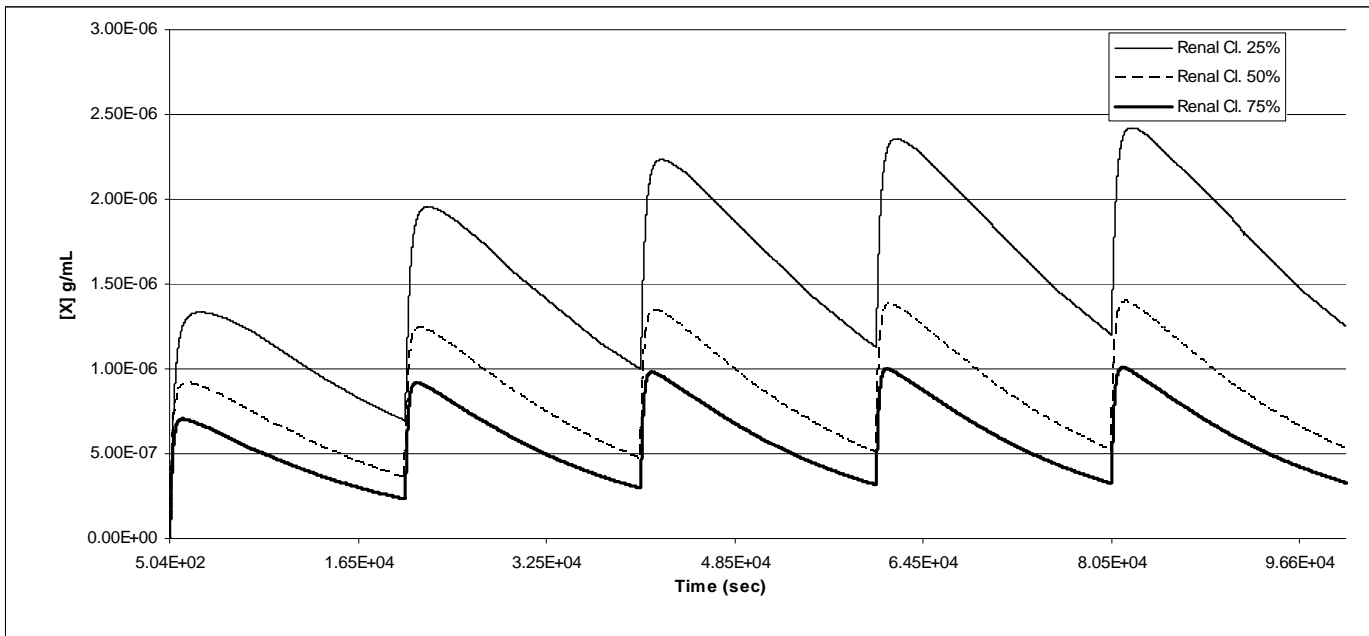


Figure 2: Effect of Renal Clearance on Drug d Dynamics in Skeletal Tissue.

Figure (3) presents the effects of renal and renal and hepatic (metabolic) clearance of the drug concentration in skeletal muscle. Two trajectories of drug concentration in skeletal muscle were simulated. One was simulated solely under renal

clearance. The second was simulated under a combination of renal and hepatic methods of clearance. The dynamics clearly demonstrate the difference in the levels of the drug concentration between two trajectories. The trajectory that

was calculated in the result of simulation of the joined effect of both ways of clearance displays significantly lower values of the drug concentration.

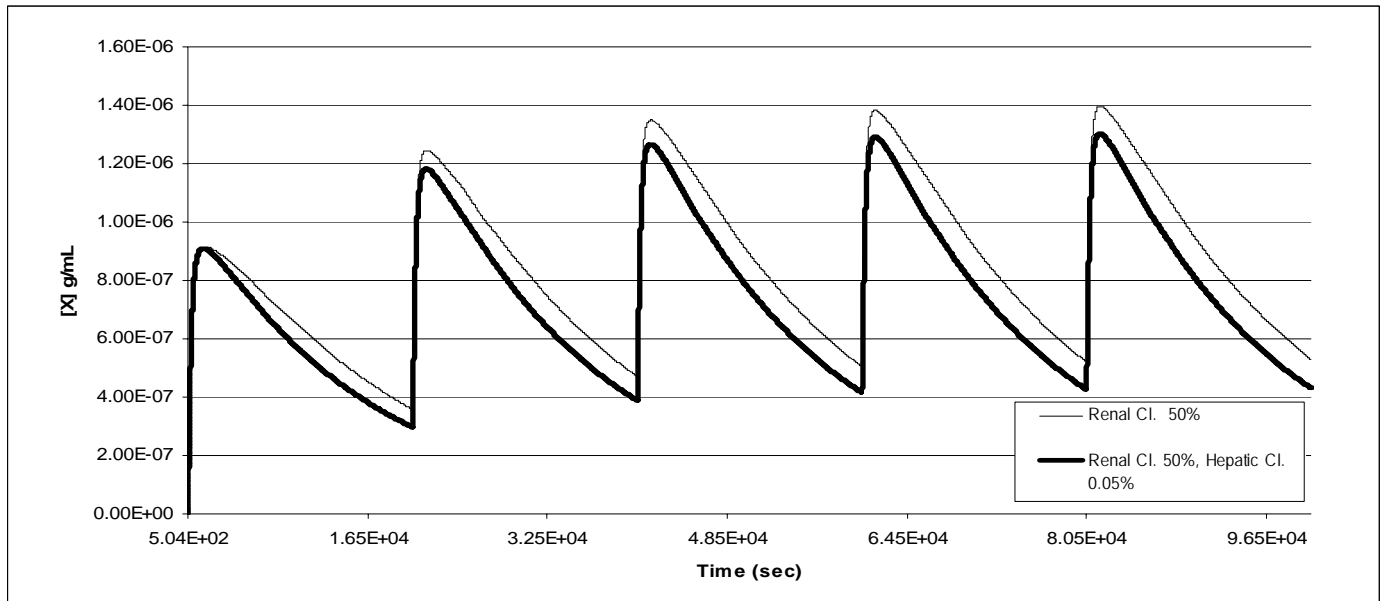


Figure 3: Effect of Hepatic Clearance on Drug d in Skeletal Tissue.

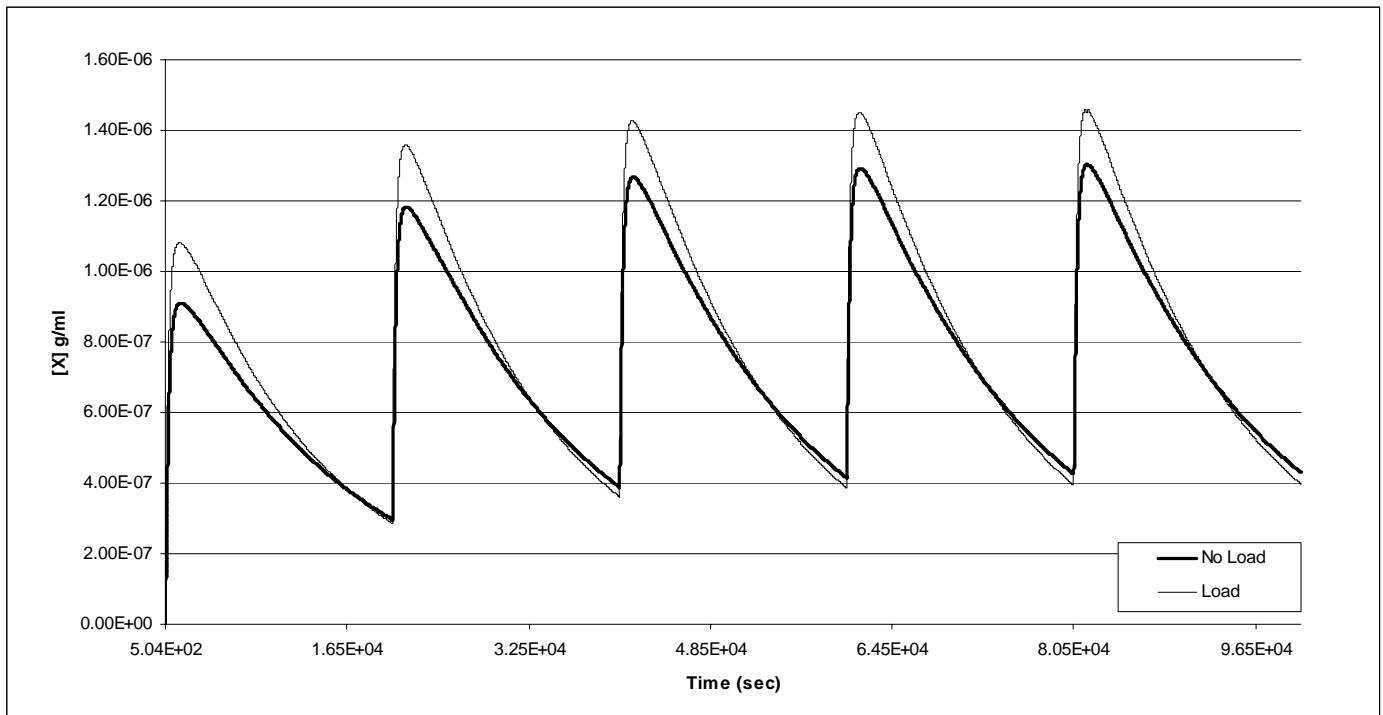


Figure 4: Effect of physical load on the drug concentration dynamics in skeletal muscle along with renal clearance ($K_R = 0.5$) and hepatic clearance ($K_H = 0.0005$).

Figure 4 represents the dynamics of the drug concentration in the skeletal muscle in a steady state (“No load”) and under physical load (“Load”, oxygen consumption

was equal to 15 ml/sec). The dynamics of the trajectory under the load shows higher rate of accumulation of the drug in the muscle as well as higher rate of elimination. These changes

are explained by the increase in the blood flow in the skeletal muscle under the load from 19 ml/sec to approximately 100

ml/sec. Under physical load the blood flow grows due to increased oxygen demand in the skeletal muscle.

6 Conclusion

The results of the experiments with the model (1)-(17) proved that the model can be successfully used for the calculations of drug dynamics in plasma and tissues influenced by several factors of drug clearance. The dynamics of the drug concentration trajectories prove that the combined influence of two clearance factors accelerates elimination of the drug from the system (organism).

The model (1)-(17) allowed to build the accurate pictures of the drug dynamics in every (each) tissue/compartiment in a steady and disturbed states. It was proven, that the model allows simulating the influence of several concurrent factors of drug clearance.

The model permits to calculate the dosage and the regimens of the drug for the different routes of its administration and ways of clearance. It can be readily adapted to a specific drug with the defined physical, physiological and chemical properties. Different mechanisms of clearance including multiple schemes of metabolism can be introduced into the model.

The model (1)-(17) is a versatile tool of the simulation of drug kinetics parameters, and can be tailored to suit the experimental needs in the fields of pharmacological and medical research; it considerably reduces the time and the cost of the laboratory studies.

7 References

[1] New Drug Development: Regulatory Paradigms for Clinical Pharmacology and Biopharmaceutics (Drugs and the Pharmaceutical Sciences), Informa Healthcare, 2004

[2] Applied Biopharmaceutics And Pharmacokinetics, 5th Edition, McGraw-Hill Companies Inc., 2005

[3] Dennis A. Smith, Han van de Waterbeemd, K. Don, Pharmacokinetics and Metabolism in Drug Design, second ed., Wiley-VCH, Weinheim, 2006.

[4] K.Polinkevych; G.Onopchuk; S.Polinkevych. Simulation of drug dynamic parameters and optimization models of the functional respiratory system in pharmacokinetic studies. Nonlinear Analysis (December 2009), 71 (12), pg. e936-e941

[5] G. Onopchuk, K. Polinkevich, Mathematical models of the respiratory system and practical problems of medicine, METMBS '04, June 21_24, 2004, CSREA Press, Las Vegas, 2004, pp. 69_74.

[6] G. Onopchuk, Mathematical modeling of hypoxic states of human organism and their pharmacological correction in case of ischemia and atherosclerosis, METMBS 2005, Las Vegas, Nevada, USA, June 20_23, 2005, CSREA Press, Las Vegas, 2005, pp. 201_204.

[7] G. Onopchuk, K. Polinkevych, Integrated approach to system researches of complex problems of medicine and biology, ISFR 2005 - The New Roles of Systems Sciences for a Knowledge-Based Society, Nov. 14_17, 2005, Kobe, Jaist Press, Komatsu, 2005, pp. 287_288.

[8] G. Onopchuk, K. Polinkevych, Mathematical approach to validation of the material balance type simulation models for pharmacokinetics studies, SERP 2007, June 25_28, 2007, CSREA Press, Las Vegas, 2007, pp. 210_216.

[9] Ross, I. M. *A Primer on Pontryagin's Principle in Optimal Control*, Collegiate Publishers, 2009

Identifying Co-targets to Fight Drug Resistance Based on a Random Walk Model

Liang-Chun Chen², Hsiang-Yuan Yeh¹, Cheng-Yu Yeh², Carlos Roberto Arias², Von-Wun Soo^{1,2}

¹Department of Computer Science, National Tsing Hua University, HsinChu 300, Taiwan

²Institute of Information Systems and Applications, National Tsing Hua University, HsinChu 300, Taiwan

Abstract- Drug resistance has now posed more severe and emergent threats to human health and infectious disease treatment. However, the wet-lab approaches alone to counter drug resistance have so far achieved limited success in understanding the underlying mechanisms and pathways of drug resistance. Our approach applied A heuristic search algorithm in order to extract drug response pathways from protein-protein interaction networks and to identify the co-target for effective antibacterial drugs. In this paper, we chose one of the killer infectious diseases, Mycobacterium Tuberculosis as our test bed. The results showed that the acetyl-CoA carboxylase is believed to be involved in fatty acid and mycolic acid biosynthesis and is strongly associated with the drug resistance mechanisms. Our analysis are consistent with the recent experimental results and also found alanine and glycine rich membrane and cell wall-associated lipoproteins to be potential co-targets for countering drug resistance.*

keywords : Drug resistance, Co-target, Random walk, Mycobacterium Tuberculosis

1 Introduction

Drug resistance has been posing an emergent threat to human health and infectious disease treatment. Several wet-lab experiments like rotation of antibiotic combinations, identification of new targets and chemical entities that may be less mutable are being explored to counter this problem by inhibiting the resistance mechanism employed by the bacterium [1]. However, those strategies are still not effective enough and have so far achieved limited success due to limited knowledge about how the resistance mechanisms are triggered in bacteria upon antibiotic drug treatment [7]. Mycobacterium Tuberculosis has remained one of the killer infectious diseases that have widely spread with prominent drug resistance. Multidrug resistant Mycobacterium Tuberculosis has underscored the need for research into the mechanisms of drug resistance and the design of more effective anti-tuberculosis agents.

Systems biology approach is essential to gain novel insights into the pathways involved in the mechanism of drug resistance from biological networks. Due to the increasing availability of protein interaction networks, network-based analysis provides an opportunity to discover active (significant) networks under specific conditions. High-throughput microarray data technology

has led to genome-wide measurements of mRNA activity levels under different conditions and it is one of the data sources that can help us realize the active networks. Most of statistical methods such as fold change, t-test identify genes using only different expressed genes among different conditions with large set of the microarray data. These methods do not utilize the knowledge of protein interaction networks nor do they capture the coordination of multiple genes. Recent works estimated the weights of protein interactions based on differential gene expression values that scored edge or vertex in the sub-networks and applied a heuristic search method to extract the significant networks and infer regulatory and signaling modules [2,3,4,5]. They proposed a search of active sub-networks in terms of a minimum-weight path search or an unsupervised maximum score sub-network problem. Vertex-based scoring methods take all known interactions among proteins as the edges of the active sub-networks. They do not further select the active interaction relationships among protein while only a part of the interactions among a set of proteins may be active. This kind of methods are inconsistent with previous studies which found that not all protein interactions occur at a specific condition [6]. Edge-based scoring applied Pearson correlation coefficient for analyzing pair relationships which do not work in the small set of the microarray data and could be unsuitable to explore the true gene relationship because it is overly sensitive to the expression value. All of them applied greedy or heuristic search instead of exhausted search and may sacrifice the optimality of the identified active sub-networks.

Typically, the target of a drug inhibits the pathogen or arrests its growth but the resistance machinery is established via certain pathways. A recent idea for a systems-level analysis is called "co-targets" instead of being the ancillary or secondary targets that have a critical physiological function for the survival of the cell but help in modifying the properties of the drug to inhibit the resistance mechanism [7]. Thus, co-targets could be either essential or non-essential but it is necessary to have a strong influence in the network and to counter drug resistance. Raman and Chandra formulated this problem as a search for the shortest paths obtained from the bacteria after exposure to the drug and calculated betweenness attribute of genes in the protein interaction networks to identify the potential co-target [7]. However, this formulation has an obvious weakness because the shortest paths are the only routes of drug resistance and there are some "back-up" ways to make the robustness of

the mechanism in bacteria [8]. Ayati et al. did not identify significant drug resistance pathways from gene expression data to solve this problem. They used balanced bipartition problem with spectral bipartition to discover the co-targets which separate multiple essential pathways into disconnected pieces to effectively disrupt the survival of a bacterium even when it has multiple pathways to drug resistance [8]. However, they simply take all the interactions in the public database as the edges are in adequate and did not consider the weight of the interaction under antibiotic drug treatment.

With the availability of gene expression and protein interaction networks, it is feasible to address the issue of drug resistance from a systems perspective. Here, we presented an efficient heuristic search function for detecting the simple paths that differs from the above researches. Then, we applied random walk model to discover a set of co-targets which affect higher probability of the genes related to the mechanism of the drug resistance through main and back-up paths instead of only considering shortest paths. The paper is organized as follows: Section 2 describes the proposed methods. Section 3 explains the experiments and discusses the results. Section 4 makes the conclusions.

2 System architecture and workflow

The workflow of our methods consist of six steps: Step 1 integrated the public protein-protein interaction networks database and assigned weight values to the interactions based on the confidence value and gene expression from antibiotic drug treatment and control samples in step 2. Step 3 presented A* heuristic search method to identify the active sub-networks upon antibiotic drug treatment and then we also extract drug resistance pathways using known curated resistance proteins in step 4 [9]. Step 5 and 6 applied the method to modify the transition matrix in the random walk method to discover potential co-targets. The overall workflow of our method is shown in Figure 1.

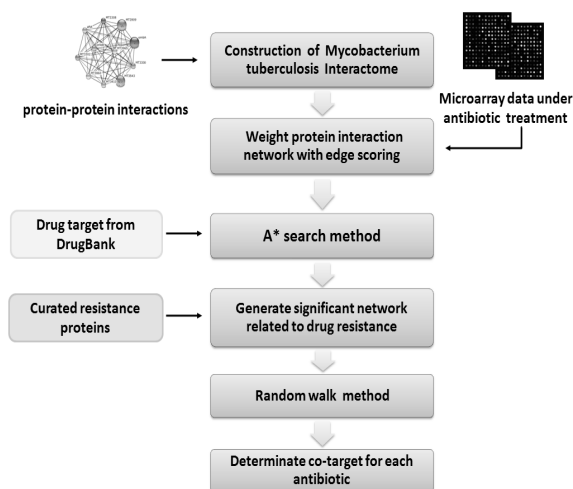


Figure 1 the overall workflow of our method

2.1 Network construction from microarray data and protein-protein interactions database

The microarray data implies gene expression information in the biological experiments. The microarray dataset consists of n genes and m experiments can be represented as an $n \times m$ matrix. It represents different gene expression levels in this matrix. Gene expressions either over-expressed or under-expressed can be revealed in terms of two colored channel in the microarray data representing the intensity of the antibiotic treatment and control samples. The gene expression ratios were calculated as the median value of the pixels minus background pixel median value for one color channel divided by the same for the other channel. We extracted the median value of the log base 2 of each gene in experimental dataset because the median value of the normalized ratio is much harder to be affected by noise than the mean value. We derive a genome-scale protein-protein interaction network from STRING database which includes interactions from published literature including experimentally studied interactions from genome analysis [10]. In STRING database, a continuous confidence score is assigned to each protein-protein interaction which is derived by benchmarking the performance of the predictions against a common reference set of trusted, true associations and also takes into account the frequency of the detection with various ways [10]. A higher confidence score is assigned while an interaction between two proteins is supported by several types of evidence.

We formulate an undirected protein interaction networks defined as $G(V,E)$ where the node set V represents protein which is the product of gene v ($v \in V$) and edge set E represents the protein interactions e ($e \in E$) in the network. Due to the network contain some false positives, we used the absolute value of the expression profile for each gene and the larger value denotes more significant differential expressed genes under drug treatment. We applied the weight to each edge which is defined in Equation (1) as the product of the absolute value of confidence score C_{ij} and the sum of the absolute value of gene expression values between two corresponding genes u and v in the edge.

$$w(e) = w(u, v) = A(u, v) = |C_{uv}| \times (|E_u| + |E_v|) \quad (1)$$

E_u and E_v are the average of the gene expression values of node u and v in the microarray. We use the adjacent matrices A of graph G to store the undirected networks as $A=(a_{uv})_{n \times n}$ where a_{uv} denotes the probability of interactions between nodes u and v .

2.2 A* algorithm as heuristic search

In order to study the part of the large scale of the protein interaction networks which is relevant to drug resistance, it is required to define the source nodes to understand the flow of drug actions. DrugBank database provide drug-related information and also determine the drug target of the antibiotic drug [11]. We assume that a source node not only refers to the drug target and

possible inhibitors associated with the function of the drug. It can be envisaged that upon inhibition of a protein and the drug-related functional mechanism often occur so as to minimize the effect of inhibition on the particular protein [12]. Therefore, we used the drug target and the genes associated with the drug-related function as source nodes for searching. In search for paths using a traditional tree search method, it may expand a large collection of new nodes while traversing new level of tree. In order to determine the range of path lengths in the network we would detect, we apply the heap-based Dijkstra's algorithm for each node to get the longest shortest path of all pairs of nodes in the network [13]. This information shows if any pair of nodes in the network can link to others at most the length and we thus use the length of the longest shortest path as the maximum length in the path searching. We assume that the active sub-networks extraction issue is a minimum score linear path searching problem with the fixed length. First, we normalized the weight $w(e)$ of the edge e calculated by Equation (1) to be the range [0,1]. Then, we transfer the larger weight of the edge to be a smaller score and the score of the edge e between two corresponding genes u and v is calculated as $score(e) = score(u,v) = -\log(w(u,v))$. The negative logarithm makes larger weight become smaller score and so on. First, we defined the score of a path as the sum of scores of edges in the path and the formula is defined in Equation (2):

$$score(p) = \sum_{e \in p} score(e) \quad (2)$$

where

$score(e)$ is the score of an edge e in the path p

To speed up the procedure in search of the minimum score linear path, it needs to prune the unexplored new nodes heuristically. We use the idea of A* search to design a pruning strategy and the heuristic function is to determine the weight of a pathway that reflects significance to some extent. In the preprocessing experiments, we determine the edge with minimum score as $score_{min}$ and an average score of edges as $score_{avg}$. Then, we calculate the scores of the simple paths with the same length l between different source and end proteins in the network. We ran the procedure 5000 times to determine the scores of all paths in the experiments formed a normal distribution and we defined the error rate based on the standard deviation $score_{std}$ to find the optimal pathway in estimating bound heuristic function of $h(x)$ for a node x . We employed A* search method can explore heuristically after searching a fix length d in the paths that calculates current weight of a path as function of $g(x)$. The overall heuristic function of $f(x)$ is defined in Equation (3) for finding a pathway with an optimal (minimum) score.

$$f(x) = g(x) + h(x) \\ = score(P_d) + score_{min} \times (l - d) \quad (3)$$

where

l means the length of a path,

d means the length from the source node that we have already traversed in the network,

$score(P_d)$ means the sum of the score up to the current node x with a length parameter d ,

$score_{min}$ means the minimum edge score in the network.

Because the lower $f(x)$ a node is estimated, the more likely is it to be searched. We set a bound score for a path p with length l that is defined as Equation (4) to control the quality of the path we could find:

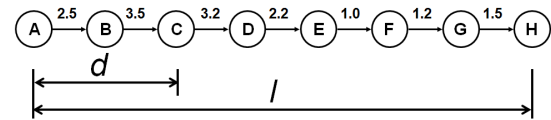
$$Bound(p) = (score_{avg} + \alpha \times score_{std}) \times l \quad (4)$$

α is a constant factor to control the bound

$score_{avg}$ means the average score calculated in the preprocessing experiments,

$score_{std}$ means the standard deviation calculated in the preprocessing experiments.

While we move to the next node through the edge in each search process, we compute heuristic function $f(x)$ and compare it with the initially-set bound score. If $f(x)$ exceeds the initially-set bound score, we do not expand the node further. For the nodes that are allowed to expand, their children nodes are expanded and their heuristic functions are computed and compared with the bound score again until the search reaches the end node. As the example in Figure 2, we consider finding a pathway with length $l=7$ from the initial node A to the end node H.



Initial variables:

$Score_{avg}=1.6$; $score_{min}=1.0$; $score_{STD}=0.5$; $\alpha=0.5$;

$Bound(p)=(1.6+0.5*0.5)*7=12.95$;

For node C:

$f(x)=6+1.0*(7-2)=11 < Bound(p)$

→ Continue to search node D

For node D:

$f(x)=9.2+1.0*(7-3)=13.2 > Bound(p)$

→ Stop searching

Figure 2 an example for A* searching method

First we explored a fix length $d=2$ from initial node A that lead us to node C, we start to estimate the score of a path with an additional length of 5 that yields a total weight 11 from current node C. The estimated score of the path is smaller than the bound score 12.95, therefore, we continue to traverse its children. The function of $f(x)$ of current node D is 13.2 and therefore we cannot search into its children. We applied heuristic method to prune the search space instead of exhaust searching for all the edges in the network.

The known drug resistance genes reported in the previous researches further help in classification of the paths [9] and we identified the function the potential drug resistance pathways where at least one of curated resistance proteins within paths. We extract the linear or tree-like path in the protein interaction network and we

assemble them to the active sub-networks N_{DR} with significant gene set G_{DR} .

2.3 Random walk to discover co-target

Random walk (RW) is a ranking algorithm [15]. It simulates a random walker starts on a set of seed nodes and moves to its immediate neighbors randomly at each step. Finally, all the nodes in the graph are ranked by the probability of the random walker reaching this node. The procedure of the RW model provides the basic idea to propagate the information from the drug target to the other genes in the network based to the gene expression.

2.3.1 Initial probability for primary drug treatment using RW

Based on the characteristic of RW, we applied this method to discover potential co-targets which have the maximum probability to affect the genes related to the drug resistance mechanisms. First, for every node v ($v \in V$), we defined $adj(v)$ which describes the set of nodes u with direct interaction with node v in the network G , and $ws(v)$ as the sum of the weight associated from node v to its neighbors u in adjacency matrix A , their formal definition is in Equation (5) and (6), respectively. The transition matrix M for RW is computed using the adjacency matrix A and $ws(v)$ and the transition probability from node v to node u is defined as Equation (7) where $w(v,u)$ is calculated by Equation (1)

$$adj(v) = \{u \mid (v, u) \in E\} \tag{5}$$

$$ws(v) = \sum_{u \in adj(v)} w(v, u) \tag{6}$$

$$M_{vu} = probability(v \rightarrow u) = w(v, u) / ws(v) \tag{7}$$

Let P_0 be the initial probability vector constructed in such way that equal probabilities assigned to all the source nodes with their probability sum equal to 1. Let P_s be a vector in which a node in the network holds the probability of finding itself in the random walker process up to the step s , the probability of P_{s+1} can be derived by

$$P_{s+1} = M^T P_s \tag{8}$$

We plunge the transition matrix M and initial probability vector P_0 into the iterative Equation (8). After certain steps, the probabilities will reach a steady state which is obtained by performing the iteration until the difference between P_s and P_{s+1} measured by L1 norm falls below a very small number such as 10^{-8} . We defined the vector $P_{reference}(d)$ representing the steady state probability vector for the treatment merely by drug target d and also represents the probability of the nodes in the network as the reference probability vector.

2.3.2 Discovering potential co-target

A combination of primary drug target and co-target should disrupt pathways and reduce the emergence of drug resistance thus allowing the main drug to kill the bacteria. Due to the calculation of the weight of the edge is done from the primary antibiotic treatment, we modify the transition matrix in order to determinate the possible probability of the interaction while setting candidate

co-target. We make the following constraints to specify the new transition matrix M' :

- (1) To inhibit proteins that are co-target, the probability of the interaction to this node in the transition matrix should be set to a small value ϵ .
- (2) The constraint of the transition matrix is that sum of the weight of the node should be equal to 1, so the rest of the weights must be set accordingly if at least one of the edges is set to ϵ .

In order to satisfy the above constraints, we have the following definition: Let $ct(v)$ be a set of proteins where the node belong to $adj(v)$ of node v and is also a co-target in Equation (9).

$$ct(v) = \{u \mid adj(v) \wedge u \text{ is a co-target}\} \tag{9}$$

For every node v in the network, if the nodes u in $adj(v)$ belongs to $ct(v)$, we want to reduce the probability of walking into co-target node with small value ϵ , else, we first count the number of the nodes in $ct(v)$ as $|ct(v)|$ and calculate the sum of the weights of those nodes in $adj(v)$ which are not in $ct(v)$ as $ws'(v)$ in Equation (10). Afterwards, we adjust the weight to each node which is not in $ct(v)$ based on their weight ratio of the remaining probability in Equation (11).

$$ws'(v) = \sum_{w \in adj(v) - ct(v)} w(v, u) \tag{10}$$

$$M'_{vu} = \begin{cases} e & u \in ct(v) \\ \frac{w(v, u)}{ws'(v)} (1 - |ct(v)|e) & u \notin ct(v) \end{cases} \tag{11}$$

Where

$|ct(v)|$ denotes the number of nodes in $ct(v)$

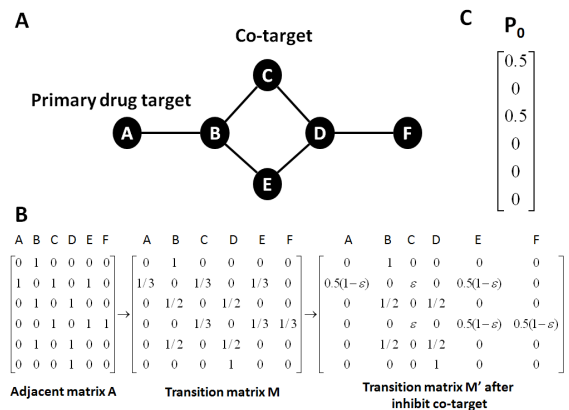


Figure 3 the example for transition matrix of co-target assignment

The small undirected network is represented in Figure 3(A) where node A is a primary drug target and all the weights of the edges are equal to one. Figure 3(B) is the adjacent matrix A and original transition matrix M calculated by Equation (7). While we choose the node C to be co-target, the modified transition matrix M' is calculated by Equation (9)-(11). Take node B as an example, first we get $adj(B) = \{A,C,E\}$ and $ct(B)=\{C\}$ from Equation (9) and then we set the probability of

M_{BC} and M_{DC} to be ε based on Equation (11). The probability of M_{BA} is calculated by

$$M'_{BA} = \text{probability}(B \rightarrow A) \\ = \left(\frac{\frac{1/3}{1/3 + 1/3}}{\frac{1/3}{1/3 + 1/3}} \right) (1 - (1)e) = \frac{1}{2}(1 - e)$$

In a similar manner are set the probabilities of M_{BE} , M_{DE} , and M_{DF} . The initial probability P_0 is formed such that equal probabilities are assigned to the nodes which are targeted by the drug and co-target with the sum equal to 1. In Figure 3(C), the initial probabilities for the pair of the primary drug target and co-target are set as 0.5 respectively. After certain steps, the probability will reach a steady state to the probability $P_{\text{cotarget}}(d, t)$ for the treatment by the primary antibiotic target d and its co-target t . Finally, we obtained an function $F(d, t)$ which is shown in the following Equation (11) for every primary drug target-co-target pair. The function $F(d, t)$ denotes the relative visitation frequency of drug resistance gene set G_{DR} between the co-target $P_{\text{cotarget}}(d, t)$ and reference probability $P_{\text{reference}}(d)$.

$$F(d, t) = \sum_{g \in G_{DR}} P_{\text{cotarget}}(d, t)_g / P_{\text{reference}}(d)_g \quad (11)$$

Where $P_{\text{cotarget}}(d, t)_g$ denotes the probability of the g^{th} gene which belongs to the function of drug resistance in the vector of the $P_{\text{cotarget}}(d, t)$

3 Computational experiments and results

We extracted protein interaction networks of Mycobacterium Tuberculosis H37rv from STRING database which contains 3,764 proteins with 179,920 undirected interactions among them. We extracted microarray experiments data which have been deposited in Gene Expression Omnibus at NCBI with accession number GSE1642 [16]. Isoniazid (INH) is a central component of drug regimens used worldwide to treat tuberculosis. H37Rv treated with 0.2mg/mL and 0.4mg/mL isoniazid (+1uL/mL EtOH) for 6h with MIC (0.02ug/mL) and control cells treated with equivalent amount of EtOH for 6h. It must be noted that it is possible that the high concentration may lead to abnormal expression but there may be a higher probability to develop drug resistance. Isoniazid is known to be inhibitors of mycolic acid biosynthesis. It can be envisaged that upon inhibition of a protein within drug treatment and metabolic adjustments often occur so as to minimize the effect of inhibition on the particular protein [7,12]. In order to incorporate the effect of such adjustments, we have considered the functional related genes as source rather than individual drug target and we use 21 proteins as source nodes for A* search to extract active sub-networks [4].

3.1 The drug response and resistance pathways of the antibiotic treatment

The variation of the gene expression in the microarray data upon exposure to anti-tubercular identify

lists of genes whose expression levels were either increased or decreased. There are 1,920 over-expressed genes, 1,806 down-expressed genes and the expression value of the 38 genes are equal to zero. Known 71 genes relevant to resistance mechanisms were classified into four types (a) efflux pumps which transport drugs out of the cell, (b) cytochromes and other target-modifying enzymes that cause potential chemical modification of drug molecules, (c) SOS-response and related genes leading to mutations or its regulatory region, (d) proteins involved in horizontal gene transfer (HGT) to import a target modifying protein from its environment. Table 1 shows the number of the over- and down- expressed genes belong to curated resistance proteins [9]. Our experiments observed seven up-expressed genes of antibiotic efflux pumps and ten in down-expression. There are five over-expressed and four under-expressed genes in SOS. Most over- and under- expressed genes have connection with cytochromes, 15 up-expression and 20 down-expression in cytochromes. We found that 32.3% (22/68) of the genes' absolute expression value are larger than the average of the absolute expression value of all genes in the microarray data. But we only found that expression values of two genes (iniA and efpA) are more than two standard deviations. Only dependent on the patterns of variation in terms of an increase or decrease in the expression levels of individual genes are hard to know the mechanism of the drug response and resistance.

Table 1 the number of the over- and down- expressed genes belong to curated resistance proteins

Drug resistance	Up	Down
Antibiotic efflux pumps	7	10
Hypothetical efflux pumps	2	2
Antibiotic degrading enzymes	1	0
Target-modifying enzymes	1	0
SOS and related genes	5	4
Genes implicated in horizontal gene transfer (HGT)	1	2
Cytochromes	15	20

Previous researches observed that paths to different resistance mechanisms for different drugs and it suggest that a given target may have a higher propensity for eliciting a specific mechanism of resistance [8]. Therefore, we applied the length of seven is the longest shortest path in bacteria network and detect the path with the length from three to seven as our experiment testing. We identified the potential drug resistance pathways under isoniazid treatment where at least one of curated resistance proteins within paths and assemble them to the active sub-networks. The part of the drug resistance network assembles by the paths while setting alpha value equal to three is shown in Figure 4. Nodes are labeled by their gene symbol as indicated. The thickness of an edge is proportional to the number of times that the active

sub-networks we extracted are traversed through this edge. The node with dashed line represents the gene is the known drug resistance genes.

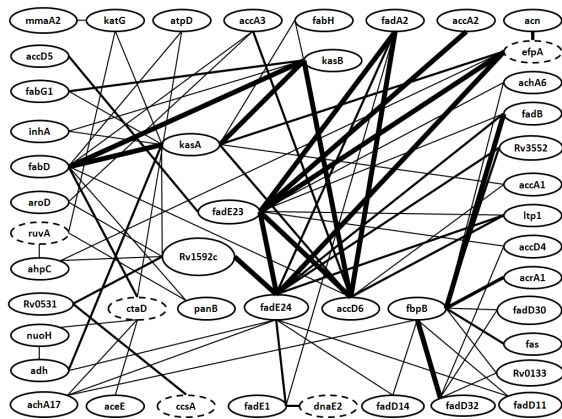


Figure 4 the part of the drug resistance networks

The global view of the Figure 4, we suggest that drug resistance related genes *efpA*, *ccsA*, *ctaD* and *dnaE2* strongly associated with *fadE* family which can contribute directly to the emergence of drug resistance. Genes *kasA*, *kasB* and *fabD* play important roles which have stronger relationship with *fadE* family in the active networks extracted by our method. Then, we show that the linear paths with small score in the network. Table 2 denotes the paths with small score which are belong to different resistance mechanisms and the value of $S_{avg}(P)$ is the score(p) divided by the number of node involved in the path. The top significant drug resistance paths is antibiotic efflux pumps with minimum score 1.05. It was interesting to observe that *efpA* is an important transporter known to confer resistance involved in the antibiotic efflux pumps paths in isoniazid [12]. Genes *fadE1/23/24*, *fadD*, *kasA*, *kasB*, and *accD6* encoding enzymes are involved in fatty acid oxidation and fatty acid biosynthetic pathway [17, 18, 19, 20]. Gene *accD6* is an acetyl-CoA carboxylase that is involved in the production of malonyl-CoA. The result has previously been shown that genes are over-expressed in *Mycobacterium Tuberculosis* in the presence of activated isoniazid in the wet-lab experiment [17]. The edges in the SOS response were common to paths from cell wall proteins and *ahpC* genes that encode type II fatty acid synthase enzymes involved in mycolic acid biosynthesis. In the cytochromes mechanism, *Rv1592c* and *Rv0531* are the genes with unknown functions and they are also transcriptionally induced by isoniazid [19]. Genes *fabG1* and *inhA* both encode mycolic acid biosynthetic enzymes and *fabG1-inhA* regulatory region have also been identified and associated with isoniazid resistance [17]. NADH dehydrogenase (*ndh*) has been associated with isoniazid resistance. The essential acetyl-CoA carboxylase is involved in fatty acid and mycolic acid biosynthesis in *Mycobacterium Tuberculosis* and those genes are also strongly associated with growth and cell wall function. Our findings suggest are consistency with the recent experimental results.

Table 2 top paths of the drug resistance mechanism in active sub-networks

Top paths in active sub-networks	$S_{avg}(P)$
Antibiotic efflux pumps	
<i>kasA--kasB--accD6--fadA2--fadE23--efpA--acn</i>	1.05
<i>fabD--kasB--accD6--fadA2--fadE23--efpA--acn</i>	1.08
<i>fabD--kasA--efpA--fadE23--echA6--fbpB--acrA1</i>	1.12
<i>fadD32--fbpB--fadD11--fadE24--efpA--fadE23--accA2</i>	1.14
SOS	
<i>fabD--kasB--accD6--fadE23--fadE24--fadE1--dnaE2</i>	1.43
<i>fabD--kasA--accD6--fadE23--fadE24--fadE1--dnaE2</i>	1.48
<i>inhA--kasB--kasA--fabD--panB--ruvA--ahpC</i>	1.64
Cytochromes	
<i>kasA--kasB--fabD--ctaD--echA17--fbpB--acrA1</i>	1.34
<i>kasB--fabD--kasA--ndh--nuoH--ctaD--aceE</i>	1.42
<i>fabG1--kasB--fabD--ctaD--echA17--fbpB--acrA1</i>	1.49
<i>kasB--accD6--fadA2--fadE24--Rv1592c--Rv0531--ccsA</i>	1.56
<i>accA3--accD6--fadE23--fadE24--Rv1592c--Rv0531--ccsA</i>	1.59
<i>fadD32--fbpB--fadD11--fadE24--Rv1592c--Rv0531--ccsA</i>	1.61

3.2 The potential co-target discovered by random walks

After we ran our random walk model for 868 genes in G_{DR} , we display top 5 co-targets in Table 3. The top 1 potential co-target, *Rv2721c* is associated with alanine and glycine rich membrane protein which has been suggested to be important for maintenance of the NAD pool [21]. Our method discovered *rv0483* (*lprQ*) which is previously shown to be cell wall-associated by proteomics and it could be a specific inhibitor to counter the drug resistance [22]. Lipoproteins such like *lprQ* carry out important functions efficiently at the membrane aqueous interface and its biosynthetic pathway is also essential for bacterial viability. Bacteria may be inherently resistant with particular type of cell wall structure with an outer membrane that establishes a permeability barrier against the antibiotic. Although *Rv0885*, *rv1109C* and *rv2137C* are all hypothetical proteins, they are all strongly functional interact with the lipoproteins, adrenodoxin oxidoreductase and cell wall processes which is deposited in STRING database. Although the biological validation for the predicated results from our method is difficult, it turns out that some of our predicted results had been reported in the public literature for validation.

Table 3 top 5 co-targets for countering drug resistance

Co-target	F(d,t)	Annotation
<i>rv2721c</i>	144.16	conserved alanine and glycine rich membrane protein

rv1109c	144.03	conserved hypothetical protein
rv0483	143.93	lipoprotein lprQ
rv0885	143.87	conserved hypothetical protein
rv2137C	143.86	conserved hypothetical protein

4 Conclusion

We develop a computational workflow for giving new insights to bacterial drug resistance which can be gained by a systems-level analysis of bacterial regulation networks. In our approach, we utilize information on STRING database and expression data to construct a weighted network and to decipher the active networks related to drug resistance using A* search method. We also identified the potential genes having higher probability using modified random walk model and suggested those genes that could be explored as co-targets. Knowledge of the active networks under specific condition will help us address more systematic and novel ways. The merit of this research would help biologists to understand the cellular mechanism more easily so that they could either based it to conduct further clinical diagnosis or verification. In the future, we could further integrate directed DNA-gene interaction and signal pathway to construct a more complete networks. The edge orientation of the undirected protein network based on the domain-domain interactions could be added to realize the signal flow in the network. The genome of the drug-resistant strain and non-drug-resistant strain should be compared to identify extra genes which are worth considering as significant components for co-targets and drug-resistance pathways.

References

- [1] Y. T. Tan, D. J. Tillett and I. A. McKay, "Molecular strategies for overcoming antibiotic resistance in bacteria," *Molecular medicine today*, vol. 6, no. 8, pp. 309-314, August 2000.
- [2] J. Scott, T. Ideker, R. M. Karp and R. Sharan, "Efficient algorithms for detecting signaling pathways in protein interaction networks," *Ninth Annual International Conference on Research in Computational Molecular Biology*, LNBI 3500, pp.1-13, 2005.
- [3] X. Zhao, R. Wang, L. Chen, and K. Aihara, "Automatic modeling of signal pathways from protein-protein interaction networks," *Proceedings Trim Size*, vol. 3, no. 42, September 29, 2007.
- [4] Z. Guo, et al., "Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network," *Bioinformatics*, vol. 23, no. 16, pp. 2121-2128, June 2007.
- [5] T. Ideker, O. Ozier, B. Schwikowski and A. Siegel, "Discovering regulatory and signaling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, pp. S233-S240, April 2002.
- [6] J. Han, et al., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, pp.88-93, July 2004.
- [7] K. Raman and N. Chandra, "Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance," *BMC Microbiology*, vol. 8, no. 234, pp. 1471-2180, July 2008.
- [8] M. Ayati, G. Taheri, S. Arab, L. Wong and C. Eslahchi, "Overcoming Drug Resistance by Co-Targeting," *IEEE International Conference on Bioinformatics & Biomedicine*, 2010
- [9] P. A. Smith and F. E. Romesberg, "Combating bacteria and drug resistance by inhibiting mechanisms of persistence and adaptation," *nature chemical biology*, vol. 3, no. 9, pp. 549-556, September 2007.
- [10] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel, "STRING: a database of predicted functional associations between proteins," vol. 31, no. 1, pp. 258-261. September 2002.
- [11] D.S. Wishart, et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.* 36(Database issue): D901-D906, 2008
- [12] L. Nguyen and C. J. Thompson, "Foundations of antibiotic resistance in bacterial physiology: the mycobacterial paradigm," *Trends in Microbiology*, vol.14, no.7, pp. 304-312, July 2006.
- [13] E. W. Dijkstra, "A note on two problems in connection with graphs," *Numerische Mathematik*, vol.1, pp.269-271, 1959.
- [14] K. Raman, P. Rajagopalan and N. Chandra, "Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-Tubercular Drugs," *PLoS Computational Biology*, vol. 1, no. 5, pp. e46, August 2005.
- [15] S. Köhler, S. Bauer, D. Horn and P. N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949-958, April 2008.
- [16] H.I. Boshoff, et al., "The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action," *The Journal of Biological Chemistry*, vol.17, no. 279, 2004
- [17] A. Banerjee, et al., "inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis," *Science*, vol. 263, pp.227-230, 1994.
- [18] M. Wilson, et al., "Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 12833-12838, 1999.
- [19] S. T. Cole, et al., "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence," *Nature*, vol. 393, pp. 537-544, 1998.
- [20] A. Lee, A. Teo, and SY Wong, "Novel Mutations in ndh in Isoniazid-Resistant Mycobacterium tuberculosis Isolates", *Antimicrob Agents Chemother.* 2001 July; vo. 45, no. 7, pp. 2157-2159.
- [21] B. Hutter and T. Dick, "Increased alanine dehydrogenase activity during dormancy in Mycobacterium smegmatis", *FEMS Microbiol Lett* vol. 167, pp.7-11, 1998
- [22] J.A. McDonough, et al. , "Identification of functional Tat signal sequences in Mycobacterium tuberculosis proteins", *J Bacteriol*, vol. 190, no. 19, pp.6428-38, 2008.

The rhesus macaque is three times as diverse but more closely equivalent in “damaging” coding variation as compared to the human

Qiaoping Yuan^{1*}, Zhifeng Zhou^{1*}, Stephen G. Lindell¹, J. Dee Higley², Betsy Ferguson³, Robert C. Thompson⁴, Juan F. Lopez⁵, Stephen J. Suomi⁶, Basel Baghal¹, Maggie Baker¹, Deborah C. Mash⁷, Christina S. Barr^{1**}, David Goldman^{1**}

¹Laboratory of Neurogenetics, National Institute on Alcohol Abuse and Alcoholism, NIH, Bethesda, MD 20892, USA; ²Laboratory of Clinical and Translational Studies, NIAAA, Bethesda, MD 20892, USA; ³Oregon National Primate Research Center, Oregon Health and Sciences University, 505 NW 185th Ave., Beaverton, OR 97006, USA; ⁴Department of Psychiatry, University of Michigan, Ann Arbor, MI 48104, USA; ⁵Mental Health Research Institute, University of Michigan Medical Center, 3064 NSL, 1103 East Huron Street, Ann Arbor, MI 48104, USA; ⁶Laboratory of Comparative Ethology, National Institute of Child Health and Human Development, NIH, Poolesville, MD 20837, USA; ⁷Department of Neurology, University of Miami School of Medicine, Miami, FL 33124, USA

Abstract - Using a parallel next-gen sequencing and analytic pipeline, we sequenced the whole mRNA transcriptome and trimethylated histone H3-lysine 4 marked DNA regions in hippocampus from 14 humans and 14 rhesus macaques. Using this equivalent methodology and sampling space, we identified 462,802 macaque SNPs, most novel and disproportionately located in functionally important genomic regions. At least one SNP was identified in each of more than 16,000 annotated macaque genes. Comparative analyses with these SNPs equivalently identified in the two species revealed that rhesus macaque has approximately three times higher SNP density and average nucleotide diversity as compared to the human. The effective population size of the rhesus macaque is estimated to be approximately 80,000 and several times that of the human. Across five different genomic regions (intergenic, 5 Kb upstream of transcription start site, introns, untranslated, coding), intergenic regions had the highest SNP density and average nucleotide diversity and coding sequences the lowest, in both human and macaque. Although there are more coding SNPs (cSNPs) per individual in macaque than in human, the ratio of d_N/d_S in macaque is significantly lower than that in human. Furthermore, the number of predicted “damaging” nonsynonymous cSNPs in macaque is more closely equivalent to that of the human.

Keywords: Macaque, Human, Sequencing variation, Single nucleotide diversity, SNP density, Comparative genomics

1 Introduction

Rhesus macaque (*Macaca mulatta*) monkeys and humans (*Homo sapiens*) are thought to have shared a common ancestor approximately 25 million years ago [1].

Due to their genetic, physiological and behavioral similarities with humans, and because of their hardiness, adaptability, and availability, the rhesus macaque has been widely used as a nonhuman primate model in biomedical research [2,3]. Humans presently are the most numerous and widespread of primates. Furthermore, hominid apes representing the ancestral lineage of humans were geographically widespread, their fossils having been found in both Africa and Asia. However, the human diaspora is relatively recent, with our African ancestry dating back only 80,000 to 150,000 yrs b.p [4]. Also, the number of humans worldwide numbered as low as one million as recently as 100,000 yrs ago [5], and due to limitations in dispersion and gene flow effective population sizes were probably much smaller still. Substantial evidence exists that the neutral genetic diversity of humans has been shaped, and in fact restricted, by an effective population size that until recently was less than 8,000 [6].

The geographic range of the rhesus macaque extends from Afghanistan to the East China Sea. The population presently numbers in the millions, and in its range and population size the rhesus macaque is only exceeded by the humans among primate species [7]. Fossil evidence indicates that the *Macaca* genus originated in North Africa, and dispersed to various sites in Asia at least three million years ago [8]. The rhesus macaque has adapted to a variety of natural environments, including savannah and forests, and various climatic zones. Rhesus macaques thrive in cities – where they live side by side with man. The diversity of environmental adaptations and large current and ancestral population sizes suggests that the genetic legacy of the rhesus macaque may include a higher quotient of both neutral and selectively significant genetic variation than humans. Consistent with a high degree of genetic variation, substantial morphological variation has been observed between rhesus macaques in the same populations and also between populations, with as many as 13 subspecies

* These authors contributed equally to this work.

** Correspondences: cbarr@mail.nih.gov

identified [9]. Within rhesus macaques there is some evidence for genetic distinctiveness at the molecular level, and Indian rhesus may be among the least diverse [10]. Several studies using protein polymorphisms have found higher levels of diversity in Rhesus macaques from China (where there are also more subspecies) than India, and there is some evidence for a genetic bottleneck in Indian Rhesus macaques [9]. However, substantial gene flow probably occurred later, which could refresh genetic variation. In a study of six rhesus macaque populations, including Indian, Burmese, and four Chinese populations, Indian macaques had one third to one sixth the mitochondrial DNA diversity as compared to four other populations. But the Indian macaques were approximately equal in diversity to one of the Western Chinese populations [9]. A recent study with more than 1,000 Single nucleotide polymorphisms (SNPs), which are more mutationally stable than other types of polymorphisms, revealed that Indian and Chinese rhesus macaques were nearly identical in genetic diversity [11]. Taken together, the evidence suggests that the rhesus macaque is likely to be a genetically diverse primate species but Indian macaques are if anything among the least heterogeneous populations. Genomic analysis of rhesus macaques of Indian origin would thus provide a conservative estimate of the variability of rhesus macaques.

A draft genome sequence of a single Rhesus macaque of Indian origin was completed in 2007 [3]. This draft sequence opened the opportunity to map the amount and type of macaque genomic variation. Furthermore, characterization of genetic variation in macaques would greatly improve the value of the rhesus macaque as an animal model for human biology. However, there has been no systematic genome-wide view of the genetic diversity within this species. At present, fewer than 8,000 SNPs from macaque have been recorded (dbSNP Build 131, http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi?view+summary=view+summary&build_id=131). In 2007, Malhi et al. reported about 23,000 candidate SNPs from pyrosequencing [12].

Compatible with its larger effective population size across evolutionary timeframes, the macaque appears to have higher sequence diversity than the human [3,13]. SNP density in macaques was estimated to range 1~7.8 SNPs/Kb [3,14]. However, the number of loci on which this conclusion is based is relatively small, and the loci were not selected in an unbiased fashion. Although >22 million human SNPs are recorded, the availability of <10,000 macaque SNPs prevents large scale sequence diversity comparison between human and macaque in different genomic regions. In this study, we used SNPs equivalently identified in 14 humans and 14 rhesus macaques by massively parallel sequencing with both H3K4me3 (trimethylated histone H3-lysine 4) ChIPseq (chromatin immunoprecipitation followed with massively parallel DNA sequencing) and RNAseq (whole transcriptome massively parallel shotgun sequencing) as sources of sequenced fragments. From more than 16,000 genes some half million macaque SNPs, most newly identified, were further analyzed and the extent of diversity was compared between

humans and macaques in different genomic regions to capture effects of neutral genetic drift and selection in these two primate species. By sequencing diversity in the tissue-specific transcriptomes and histone-marked regions of the two species, we were able, without the use of DNA capture technology (that did not exist for the macaque) or whole-genome sequencing, to compare diversity in equivalent, functionally relevant genomic regions and detect effects of selection and drift on sequence substitutions in protein-coding gene regions.

2 Methods

2.1 Samples and tissues

Postmortem brain tissue (hippocampus) of 14 unrelated human (*H. sapiens*) males, age 30-50 was obtained from the University of Miami Brain Endowment Bank (Miami, FL, USA). The ethnic background of the human sample was: 6 African Americans, 8 Caucasians/Hispanics. Postmortem hippocampus of 14 rhesus macaque (*M. mulatta*) males, most unrelated, age 3.5-7, was obtained from the National Institutes of Health Animal Center in Poolesville, Maryland. Among the macaques, eleven were of Indian origin, one was of Chinese origin and two were approximately 50% Chinese/50% Indian as indicated by forensic genotyping with a panel of 96 markers optimized for macaque origin identification (Primate Genetics Program, Oregon National Primate Research Center, Table S1). The macaques at the Poolesville colony are maintained in an outbred state, with frequent introduction of new breeding stock such that their genetic diversity is expected to be equivalent to natural populations.

2.2 Construction of double-stranded cDNA libraries

Total RNA was extracted from 100 mg of hippocampus collected postmortem. Briefly, tissue samples were submerged in guanidinium thiocyanate and phenol based RNA extraction solution STAT-60 (Invitrogen, Friendswood, TX) and homogenized using a glass-Teflon homogenizer. Following mixing with chloroform and centrifugation, the aqueous phase was collected and isopropanol was added. The samples were then loaded onto RNeasy spin columns (Qiagen, Valencia, CA) for purification. To eliminate residual genomic DNA contamination, RNA samples were incubated with DNase I (Qiagen) on column at room temperature for 15 min and washed several times before collection in elution buffer. To isolate mRNA, 35 µg of total RNA was heated at 65°C for 2 min, and then mixed with 0.5 mg of Dynabeads oligo (dT)₂₅ (Invitrogen) in binding buffer (20 mM Tris-HCl, pH 7.5, 1.0 M LiCl, 2 mM EDTA). After incubation at room temperature for 5 min and then washing several times, mRNA was eluted from the beads by heating at 80°C for 2 min. The purified mRNA was fragmented to the 150 – 500 base pair range by mixing with 10 x fragmentation buffer (Ambion, Austin, TX) and heating at 70°C for 3 min. The

samples were purified with RNeasy spin column. 200 ng of fragmented mRNA was reverse-transcribed to first strand cDNA by random priming, using 3 µg of random hexamer oligos and 200 units of Superscript II reverse transcriptase (Invitrogen). The reaction was carried out at 45°C for 1 hr in First Strand Buffer (Invitrogen) with 10 mM DTT and 0.5 mM dNTP. For second-strand cDNA synthesis, 400 units of *Escherichia Coli* DNA polymerase, 2 units of *E. Coli* RNase H, and 10 units of *E. Coli* DNA ligase was added, and the reaction was carried out at 16°C for 2 hr in Second Strand Buffer with 0.2 mM dNTP. 20 units of T4 DNA polymerase was also added at the end of incubation for endrepair. The synthesized double-stranded cDNA library was purified with QIAquick purification kit (Qiagen).

2.3 Chromatin immunoprecipitation (ChIP)

Postmortem brain tissue (100 mg) was cut into slices less than 1 mm in thickness, and fixed in 3 ml of 1% formaldehyde/PBS solution for 10 min at room temperature to cross-link chromatin DNA and proteins. The tissue samples were then homogenized using a glass-Teflon homogenizer. Following homogenization, chromatin was isolated using the Upstate Magna ChIP G kit (Millipore, Temecula, CA). Briefly, cells were lysed in Cell Lysis Buffer in the presence of protein inhibitor cocktail. Nuclei were isolated from lysed cells by centrifugation, and re-suspended in Nuclear Lysis Buffer. The chromatin DNA was then fragmented into the 150 – 500 base-pair range by sonication using a Branson Sonifer (Branson, Danbury, Connecticut). To immunoprecipitate specific genomic regions of chromatin DNA, isolated chromatin was incubated with antibodies (Abcam, Cambridge, MA) against H3K4me3 and magnetic protein G beads (Millipore) at 4°C for 2.5 hr. Following incubation, beads were washed with low salt, high salt, LiCl salt, and TE buffers, and reverse cross-linked by proteinase K digestion at 62°C for 2 hr. The enriched DNA was purified after reverse cross-linking by column purification.

2.4 Sequencing with Illumina Genome Analyzer

Sample preparation and sequencing on an Illumina Genome Analyzer (Illumina, San Diego, CA) were carried out according to Illumina protocols with some modifications. Briefly, double-stranded cDNA and ChIP-enriched genomic DNA were treated with T4 DNA polymerase and Klenow fragment for end repair. The 5' ends of DNA fragments were then phosphorylated by T4 polynucleotide kinase, and an adenosine base was added to the 3' end of the fragments by Klenow (3'-5' exo⁻). A universal adaptor was added to the both ends of the DNA fragments by A-T ligation. Following 18 cycles of PCR with Phusion DNA polymerase, the DNA library was purified on a 2% agarose gel, and fragments 170 – 350 bp in size were recovered. Approximately 10 ng of the prepared DNA was then used for cluster generation on a grafted Flow Cell, and sequenced on the Genome Analyzer for 36 cycles using the “Sequencing-by-synthesis” method.

2.5 SNP calling and sequence analyses

Sequences were called from image files with the Illumina Genome Analyzer Pipeline (GAPipeline) and aligned to the corresponding reference genome (UCSC rheMac2 for macaque and UCSC hg18 for human) using Extended Eland in the GAPipeline. The uniquely mapped reads were parsed with in-house Perl scripts to generate base coverage and SNP calls as described previously [15]. To reduce false positive and false negative SNP calling for low coverage sequence data, a two-step approach was used. Briefly, reads were first pooled from all samples in a species for SNP identification. At this step, no base in the uniquely mapped reads had a quality score < 8, only a single mismatch with quality score ≥ 15 was allowed in a single 36-base read, and a probable SNP had to have three independent reads representing the same alternative allele within the pooled samples. To reduce false SNP calls due to mis-mapping of cross-exon RNAseq reads, putative SNPs were filtered to remove instances in which the alternative allele was represented only by reads located one or two bases from either end of the RNAseq fragment. Candidate SNPs were then filtered at the individual sample level, where the frequency of the alternative allele in a single sample had to be the highest or second highest with a frequency ≥ 0.2. Genotypes were called for an individual sample only when sequencing coverage was ≥ 6x for the SNP site and when the allele with the lowest coverage was represented at least 3 times and heterozygotes with each allele covered by 30~70% of sequence reads. Gene structures for human were based on RefSeq Genes in UCSC hg18 and Ensembl Genes from UCSC rheMac2 were used for the macaque. PolyPhen-2 [16] was used to predict protein functional effects of nonsynonymous coding SNPs (nsSNPs). Fourteen novel macaque cSNPs were selected to be resequenced by Sanger sequencing using the BigDye Terminator Sequencing Mix (Applied Biosystems, Carlsbad, CA) and analyzed on the Applied Biosystems 3730 DNA Analyzer.

3 Results and Discussion

3.1 SNP density is three times higher in the rhesus macaque than the human

In this study diversity was determined in short sequence reads (36 bases) equivalently detected and analyzed in 14 humans and 14 rhesus macaques (Table 1) (The raw sequences generated in this study have been deposited in The Sequence Read Archive with the accession numbers of SRA028822, SRA027316, SRA029279 and SRA029275). It is important to point out that the analytical strategy of comparing diversity within the hippocampal transcriptome and in H3K4me3-marked DNA regions resulted in the analysis of equivalent regions in the macaque and in the human. There was a strong correlation between level of expression of genic associated sequences between the hippocampus of both species and in the regions strongly tagged by H3K4me3 (Fig. S1). From these equivalent

genomic regions with at least 3x sequencing coverage, a total of 462,802 high quality putative SNPs (most of which were novel) were detected in the macaque, and 230,028 (most of which were known) were detected in the human. At least one SNP was identified in each of 14,675 human annotated genes and 16,797 macaque annotated genes.

Table 1. Summary of sequence coverage and putative SNPs

		Human	Rhesus
Genome size in reference assembly (Mb)		3,080	2,864
Non-gap reference genome size (Mb)		2,858	2,647
Unique coding sequence size in reference (Mb)		32.5	31.8
Sample number		14	14
Average 36-base reads per sample		17.4 x 10 ⁶	14.4 x 10 ⁶
Total length (Mb) of uniquely mapped reads		8,770	7,266
Mb in genome with ($\geq 1x$ sequence coverage)		1,505	1,571
Mb in genome with ($\geq 3x$ sequence coverage)		426	435
SNPs in dbSNP_B131		23.7 x 10 ⁶	7,880
SNPs in this study		230,028	462,802
Also in dbSNP_B131		206,267 (89.7%)	34 (0.0%)
Transition	AG,GA,TC,CT	155,836 (67.7%)	312,064 (67.4%)
Transversion	AC,CA,TG,GT	37,046 (16.1%)	79,061 (17.1%)
Transversion	CG,GC	25,467 (11.1%)	46,820 (10.1%)
Transversion	AT,TA	11,679 (5.1%)	24,857 (5.4%)
Genes with SNPs		14,675	16,797
Genes with SNPs in exons		11,200	12,466
SNPs located in intergenic regions		107,461 (46.7%)	269,390 (58.2%)
SNPs locate in 5Kb upstream of TSS		10,036 (4.4%)	26,303 (5.7%)
SNPs located in UTR		18,432 (8.0%)	15,455 (3.3%)
SNPs located in intron		79,875 (34.7%)	130,443 (28.2%)
SNPs located in CDS		14,224 (6.2%)	21,211 (4.6%)
Synonymous		8,329 (58.6%)	13,798 (65.1%)
Non-synonymous		5,877 (41.3%)	7,367 (34.7%)
Damaging		1,741 (29.6%)	1,525 (20.7%)
Nonsense		18 (0.1%)	46 (0.2%)

Approximately 10–25% of the putative SNPs detected in intergenic regions were found to be covered with RNAseq reads (Table S2), suggesting that significant transcription activity occurred outside of defined genic regions in both species, consistent with those reported recently [17]. Among 230,028 putative human SNPs, 90% had been recorded previously in dbSNP. This rediscovery rate is slightly higher than the 77–89% rediscovery rate for SNPs in the 1000 Genomes Project Pilot 2 deep sequencing data [18]. Also bearing on the validity of the SNP detection pipeline, the transition to transversion ratio of human and macaque SNPs was non-random. Although the random transition to transversion ratio is 1:2, this ratio is approximately 2:1 in both human and macaque. Using the same SNP calling pipeline, 22 of 26 human nsSNPs were validated by Sanger

sequencing in a previous study [15]. Using Sanger sequencing, 13 of 14 novel macaque cSNPs identified in this study were also verified. Overall, the rhesus macaque had a SNP density approximately three times higher than humans (Fig.1A). Calculated across all genomic regions with at least 4x sequencing coverage in individual samples, the SNP densities for macaques and humans were 2.82 SNP/kb and 1.07 SNP/Kb, respectively (Table 2, Fig.1A).

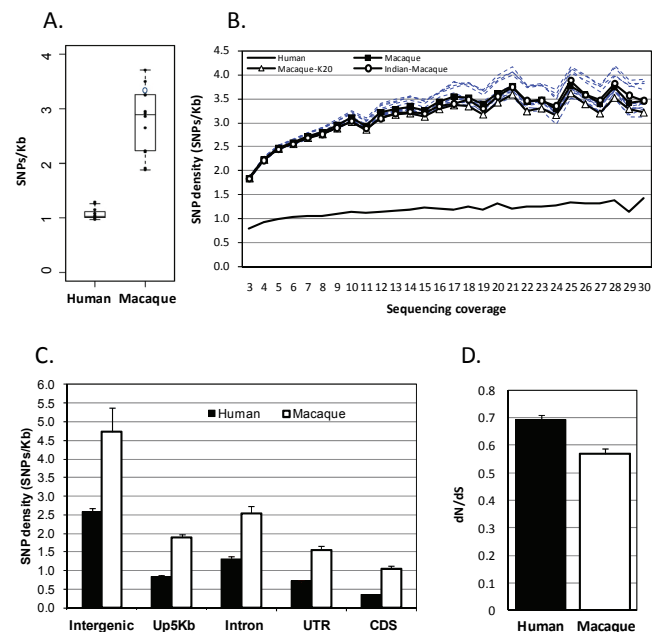


Fig. 1: Average SNP density (SNPs per 1Kb) in human and macaque. A). SNP density was calculated as the putative SNPs having different allele from reference genome divided by the unique sequenced bases in individual samples. Only bases having $\geq 4x$ sequence coverage were used for this calculation. Macaque sample K20, with Chinese origin, is labeled as an unfilled circle. B). Average SNP density in human and macaque was calculated for different sequencing coverage. Data from all macaque samples in solid line with filled square markers. Data with K20 omitted in solid line with unfilled triangle markers and others omitted one-by-one in dotted lines. Indian macaques only in solid line with unfilled circle markers. C). SNP density in 5 different genomic regions; D). The ratio of dN/dS for cSNPs. Error bar in C and D: standard error of mean.

Table 2. SNP density

Genome	Technology Used	SNPs/Kb
Venter	Sanger method	1.41*
Watson	454 Sequencing System (Roche)	1.46*
Chinese (YH)	Genome Analyzer (Illumina)	1.35*
African (NA18507)	Genome Analyzer (Illumina)	1.58*
African (NA18507)	SOLiD system (ABI)	1.69*
Korean (SJK)	Genome Analyzer (Illumina)	1.50*
Korean (AK1)	Genome Analyzer (Illumina)	1.51*
Proband (III-4)	SOLiD system (ABI)	1.50*
CEU,YRI	Genome Analyzer, SOLiD, 454	1.21-1.48**
Humans in this study	Genome Analyzer (Illumina)	1.07 (0.97-1.26)
Macaques in this study	Genome Analyzer (Illumina)	2.82 (1.88-3.71)

* SNP number was from Lupski et al. 2010 [19] and SNPs/Kb was calculated based on the total SNPs reported and 2.85 x 10⁹ of the sequenced human genome size and 80% of accessible genome [18].

** Based on the SNPs and accessible genome from high coverage Pilot Trios data of 1000 Genomes Project [18].

Because sequencing coverage for individual samples was low for most regions, putative SNPs were called by a conservative, two-step approach as described in methods. As a result, SNP density increased in both species as sequencing coverage increased (Fig. 1B), but it can be observed that the macaque had proportionately higher SNP density at all levels of sequencing coverage (Fig. 1B). One of the macaque samples was of Chinese origin and two were approximately equally admixed between Chinese macaque and Indian macaques as described in methods. However, in the comparison between macaque and human, this Chinese macaque (K20) and the two admixed macaques did not exert a larger effect on SNP density as compared to any of the Indian macaques. This was tested by omitting individual macaques one-by-one, and also by evaluating SNP density with all three of the animals with Chinese ancestry omitted (Fig. 1B). The result is consistent with what found in a recent study where no difference was found in genetic diversity between Chinese and Indian macaques using genotype analyses with more than 1,000 SNPs [11]. As mentioned, our human sample itself included individuals of different ethnic backgrounds. Therefore, the Chinese macaque and the two admixtures were included in all analyses unless specified otherwise.

At higher coverage, SNP density approached that found by higher coverage sequencing, being 1.5 SNPs/Kb for 30x coverage across human 14 samples. A range of $3.07 \sim 3.86 \times 10^6$ SNPs was found in individual human genomes [19] representing approximately 1.3~1.7 SNP/Kb. Also, a SNP density of 1.2 ~ 1.5 SNPs/Kb was found in the 1000 Genomes Project Pilot 2 data for two human family trios with >40 x sequencing coverage [18]. Here, SNP densities were estimated from 14 samples in both species and with highly similar sequencing coverage, representing a methodologically equivalent view of diversity. Since intergenic and intronic regions comprise the majority of the genome in both humans and macaques, the overall SNP densities reported here are most likely underestimates because a high proportion of our data derives from coding sequences (CDS) and untranslated regions (UTR) that have the lowest SNP densities, as will be discussed below and as shown in Fig. 1C.

SNP density was compared across five different categories of genomic regions: intergenic, 5 Kb upstream of TSS (transcription start site), introns, UTR (5'- and 3'-UTRs), and CDS as annotated in refGene (human) or ENSEMBL (macaque). In all five genomic regions, macaques had significant higher SNP densities than humans (Fig. 1C). Intergenic regions had the highest SNP density and coding regions the lowest SNP density in both species (Fig. 1C). In coding regions, 76% of the cSNPs would be expected to be nsSNPs if all base substitutions were equally likely [20]. But nsNP density was lower than synonymous cSNP density with a d_N/d_S ratio (the ratio of nonsynonymous versus synonymous substitutions, reflecting selection pressure acting on nonsynonymous sites relative to synonymous ones) in humans approximately 0.691 ± 0.017 and d_N/d_S ratio of 0.567 ± 0.022 in macaque (Fig. 1D). Although both adaptation and purifying selection

may have occurred at numerous genes for both species, purifying selection is most likely to be predominant across the whole genome in both species as their d_N/d_S ratio values were significantly less than 1. The selection pressure on nonsynonymous substitutions may have been stronger in the macaque than in the human since the d_N/d_S ratio in macaque is significantly (t-test, p-value <0.0001) lower than human. In an equivalent genomic search space, twice as many putative SNPs were identified in macaque as compared to the human (Table 1). However, macaques only had 1.2 times as many nsSNPs, reflecting that much of the increased diversity of the macaque, even in protein-coding regions of the genome, is likely to be selectively neutral. Furthermore, the nsSNPs of macaques were less likely to be “damaging” (including “possibly damaging” and “probably damaging”) as compared to the human (20% in macaque vs 30% in human), at least as predicted by PolyPhen-2 (Table 1). In line with this result, the higher d_N/d_S ratio in human may reflect a relative relaxation of purifying selection during hominoid evolution as a consequence of smaller effective population sizes or a high rate of adaptive substitution [21].

Using RNAseq and H3K4me3 ChIPseq data, a relatively high percentage of SNPs can be identified in gene coding and promoter regions, which represent functionally important domains of the genome. This could represent an advantage for certain types of gene-centric analyses. For instance, 6.2% of the human SNPs detected in this study (and 90% are previously known) were located in coding regions (cSNPs), whereas only 0.7% of the total SNPs identified in 1000 Genomes Project Pilot 2 data were cSNPs [18]. Here we sequenced only 0.426 Gb of unique human sequence at $\geq 3x$ coverage, but detected 14,224 cSNPs. This is a substantial number given that 24,192 cSNPs were detected in three Caucasian individuals with whole genome sequenced at high coverage, in the 1000 Genomes Project Pilot 2 (Fig. S2). The major limitation for SNP detection here was the proportion of genes that are not expressed in adult hippocampus or that are expressed at a low level in this tissue. The overlap of the cSNPs we detected with those reported in two individuals from the 1000 Genomes Project Pilot 2 data is consistent with the overlap that has been empirically observed between unrelated individuals (50~70% SNPs shared) on a pairwise basis (Fig. S3) [18]. Based on our sensitivity of detection of human SNPs, where 14,224 cSNPs were detected versus some 250,000 cSNPs that have been reported in NCBI (from a much larger population of subjects), we estimate that our region-focused sequencing of only 14 individuals enabled us to discover approximately 6% of the common cSNPs that are present in the rhesus macaque, although detection sensitivity was of course higher for the more abundant SNPs.

3.2 Rhesus macaques are three times as diverse as the human

The average nucleotide heterozygosity (diversity) for SNPs (θ_{SNP} , as defined by Levy et al. 2007[22]) in this study was measured as the ratio of heterozygous basepairs (both alleles with $\geq 3x$ coverage and $\geq 30\%$ of sequence reads) divided by all basepairs sequenced at this level, within each

individual. Macaque θ_{SNP} was 3 times higher than human θ_{SNP} (8.93×10^{-4} vs 3.06×10^{-4} , Fig. 2A).

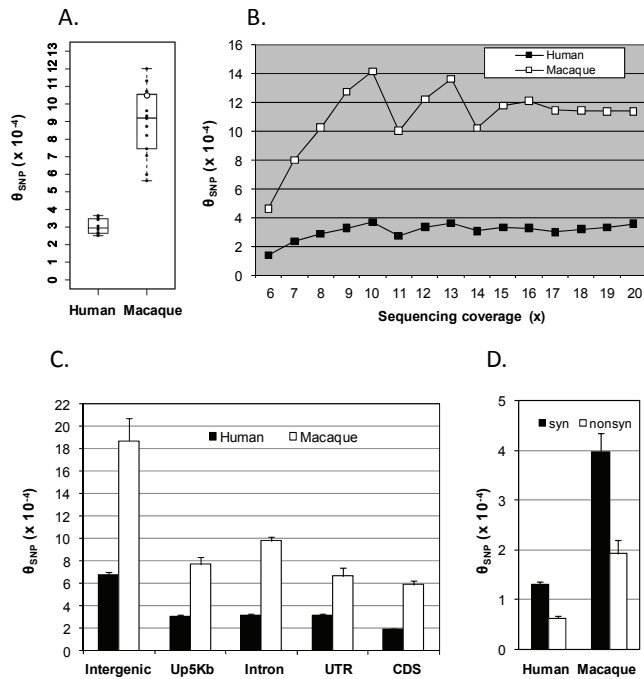


Fig. 2: Average nucleotide diversity (θ_{SNP}). A). θ_{SNP} in individual samples. Calculation was based on bases with $\geq 6x$ sequence coverage. Macaque sample K20, with Chinese origin, was labeled as an unfilled circle. B). Nucleotide diversity was average from all samples in each species at different sequencing coverage. C). Average nucleotide diversity in 5 different genomic regions; D). Average nucleotide diversity for synonymous cSNPs and nsSNPs. Error bar in C and D: standard error of mean.

Paralleling observations on SNP density, as sequencing coverage increases, more heterozygous basepairs are detected. With increasing sequencing coverage, θ_{SNP} increased, becoming asymptotic at about 10x coverage (Fig. 2B). At 20x sequencing coverage, θ_{SNP} was 11.4×10^{-4} in the macaque and 3.6×10^{-4} in the human (Fig. 2B). Similar to what we observed for SNP density, θ_{SNP} was highest in intergenic regions and lowest in coding regions in both species and macaque had significant higher θ_{SNP} than human in all five genomic regions (Fig. 2C). Our estimated θ_{SNP} in human from all regions (3.06×10^{-4}) is lower than values that can be calculated (See supplementary Table S4) from 1000 Genomes Project Pilot 2 data (7.2×10^{-4} to 9.3×10^{-4}) and is also slightly lower than values of 5.4×10^{-4} to 8.3×10^{-4} reported previously for humans [22-26]. Our overall lower human θ_{SNP} than those reported previously was expected due to our lighter sequencing coverage and higher genetic percentage of sequenced regions. However, within intergenic regions that comprise most of the genome our θ_{SNP} estimate of $\sim 6.78 \times 10^{-4}$ is actually very close to these previously reported estimates for humans based on high coverage whole genome sequencing. Therefore, the θ_{SNP} values we have computed for macaque and human appear to be robust, reflect parallel methodology and sampling and are informative for both genome-wide and regional increases in genetic diversity in the macaque compared to human.

Within coding regions it is possible to compare diversity that is more likely to be functionally significant with diversity that is more likely to be selectively neutral. In coding regions, both human and macaque had approximately 2 times more diversity for synonymous cSNPs as compared to nsSNPs (Fig. 2D), reflecting functional constraint and selection against changes in the protein sequence [25]. Concerning the possible functional significance of nsSNPs, Polyphen predicted that some 1,741 (29.6%) of the cSNPs we detected in the human and 1,525 (20.7%) of the cSNPs we detected in macaque were likely to be “damaging”. The macaque cSNPs we identified include a substantial resource of putatively functional sequence variants. Supporting the functional significance of many of these SNPs, individual humans and macaques were both half as likely to be homozygous for “damaging” nsSNPs than they were to be homozygous for synonymous cSNPs and nsSNPs scored as “benign” by Polyphen (Fig. 3).

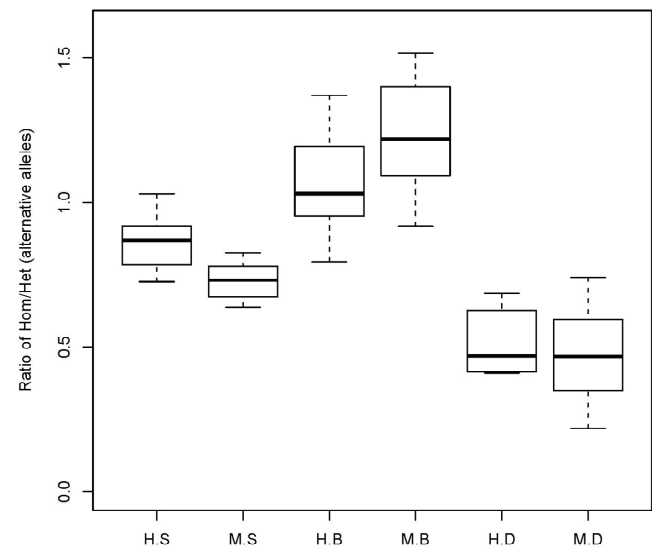


Fig. 3: The ratio of homozygous/heterozygous for the alternative alleles of cSNPs. H.S: human synonymous cSNPs; M.S: macaque synonymous cSNPs; H.B: human nsSNPs with “benign” prediction by PolyPhen; M.B: macaque nsSNPs with “benign” prediction by PolyPhen; H.D: human nsSNPs with “damaging” prediction by PolyPhen; M.D: macaque nsSNPs with “damaging” prediction by PolyPhen.

In line with the theory that most of the increased diversity of the rhesus macaque is selectively neutral in nature, the increase in macaque SNP density was not proportionately maintained from non-coding sequence to coding sequence, to nsSNPs and to putatively “damaging” nsSNPs. Instead, the macaque more closely resembled the human in its SNP density within these more functionally significant categories. Surprisingly, a different picture was observed using the diversity measure θ_{SNP} for human and macaque. By this standard, macaque was approximately three times as diverse as the human across all types of sequence categories. This could point to the maintenance of nsSNPs by balancing selection. This is an important mechanism of evolutionary adaptation in all genetically diverse species but may be operative at a larger percentage of loci in the macaque than in the human. Speculatively,

although the macaque does not have proportionately more nsSNPs, those that it does have are more likely to be maintained at higher frequency by balancing selection. However, although this would explain why nsSNP density does not increase proportionately with overall SNP density and with diversity, other validating data would be required to establish this point. One indirect test would be linkage disequilibrium analysis that could detect signals of selection (selective sweeps) at genes containing nsSNPs. In fact, one use of the SNPs we have discovered would be the creation of a marker panel enabling genome wide evaluation of LD. When that is done, the results may again be surprising.

At equilibrium, LD depends on the recombination rate and effective population size. Therefore, it might be anticipated that LD blocks in the rhesus macaque will be substantially smaller than the human. Thus a macaque SNP panel effective for genome-wide use might have to be larger than human 1M panels that are now the standard. However, it is also possible that cross-population admixture has already occurred in the rhesus macaque, at least in some samples of macaques, which could have led to the presence of much larger haplotype blocks than anticipated on the basis of population size. In this same vein, cross-population comparisons of genetic variation would be valuable. The macaques analyzed here are primarily of Indian origin, but as described earlier the species is widely dispersed. In particular there is a very large population of Chinese macaques with several Chinese subspecies proposed including a subspecies representing the island of Hainan, and several unique island-based colonies including Cayo Santiago, Puerto Rico, and Morgan Island, South Carolina. The similarity of diversity of the one Chinese and Chinese/Indian admixed macaques we studied does not address whether there are significant differences at the haplotype level, and based on the analysis of these several animals we have not developed a panel of markers informative for Chinese origin. That might also require the analysis of multiple Chinese populations. Because of their population sizes and breeding structures, macaque and human founder populations, both of which are available, offer an opportunity to observe the changing impact of population dynamics on genetic diversity of different types.

There is some evidence that the mutation rate may have slowed in the hominoid ape lineage, but based on the nucleotide diversity rates we have observed we can compare the effective population sizes of rhesus macaque and human. For this purpose, we used Watterson's (1975) [27] estimator $\theta = 4N\mu$ with average nucleotide diversity (θ_{SNP}) in intergenic regions (Fig. 2C) as θ because intergenic diversity is most likely to faithfully reflect neutral diversity at the whole genome level. Assuming an average mutation rate of 1×10^{-8} to 2.5×10^{-8} mutations per nucleotide site per diploid genome per generation for human [18,28-30] and an average mutation rate 5.9×10^{-9} mutations per nucleotide site per diploid genome per generation for macaque [14], the effective population size of humans is approximately 6,780-16,950 and the effective population size of the macaque is approximately 80,000. The most relevant comparison remains the diversity ratio between the human and macaque,

with the macaque emerging as having an effective population size several times larger.

As mentioned, our findings on the relative diversity of Chinese and Indian macaques were limited because we studied only one individual animal of Chinese origin and two that were admixed. Furthermore, the specific geographic origin of this one Chinese macaque, and the admixture component of the two other macaques, was unknown. That could be relevant, because the mitochondrial diversity of rhesus macaques from one Western Chinese population appeared to be equivalent to Indian macaques [9], which displayed lower mitochondrial diversity than several other macaque populations. However, it should be noted that the genetic diversity of nuclear DNA is less sensitive to the effects of population bottlenecks than is the diversity of the mitochondrial genome or the haploid Y chromosome. For example, a Finnish bottleneck that left a strong imprint on Y chromosome diversity led to no reduction in autosomal diversity [31]. Recently, Kanthaswamy et al revealed that Chinese and Indian macaques appeared to have near identical genetic diversity based on genotype analysis with more than 1,000 SNPs [11]. Regardless of whether there was a population bottleneck in the rhesus macaque population of India, the Indian macaques that we studied are several times as diverse as the human. Perhaps this is due to subsequent gene flow from other populations which would have restored nuclear DNA diversity of the species on the Indian subcontinent. Considering the geographic origin of the macaques we studied, it is clear that rhesus macaque is several times as diverse compared to the human, but with indications that selection has dampened the increase in functional diversity in this species.

4 Funding

This work was supported by the Intramural Research Program of the National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health.

5 References

- [1] Goodman, M., Porter, C.A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G. and Groves, C.P. (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol*, **9**, 585-598.
- [2] Barr, C.S. and Goldman, D. (2006) Non-human primate models of inheritance vulnerability to alcohol use disorders. *Addict Biol*, **11**, 374-385.
- [3] Rhesus Macaque Genome Sequencing and Analysis Consortium. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222-234.
- [4] Templeton, A. (2002) Out of Africa again and again. *Nature*, **416**, 45-51.
- [5] Thomlinson, R. (1975) *Demographic Problems: Controversy over population control*. 2nd ed. Dickenson Publishing Company, Ecino, CA.

- [6] Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E. and Visscher, P.M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res*, **17**, 520-526.
- [7] Zhang, R., Zhao, T., Quan, Q. and Southwick, C.H. (1991) Distribution of macaques (*Macaca*) in China. *Acta Theriologica Sinica* **11**, 171-185.
- [8] Abegg, C. and Thierry, B. (2002) Macaque evolution and dispersal in insular south-east Asia. *Biological Journal of the Linnean Society*, **75**, 555-576.
- [9] Smith, D.G. and McDonough, J. (2005) Mitochondrial DNA variation in Chinese and Indian Rhesus Macaques (*Macaca mulatta*). *American Journal of Primatology*, **65**, 1-25.
- [10] Ferguson, B., Street, S.L., Wright, H., Pearson, C., Jia, Y., Thompson, S.L., Allibone, P., Dubay, C.J., Spindel, E. and Norgren, R.B. (2007) Single nucleotide polymorphisms (SNPs) distinguish Indian-origin and Chinese-origin rhesus macaques (*Macaca mulatta*). *Bmc Genomics*, **8**, -.
- [11] Kanthaswamy, S., Satkoski, J., Kou, A., Malladi, V. and Smith, D.G. (2010) Detecting signatures of inter-regional and inter-specific hybridization among the Chinese rhesus macaque specific pathogen-free (SPF) population using single nucleotide polymorphic (SNP) markers. *Journal of Medical Primatology*, **39**, 252-265.
- [12] Malhi, R.S., Sickler, B., Lin, D., Satkoski, J., Tito, R.Y., George, D., Kanthaswamy, S. and Smith, D.G. (2007) MamuSNP: a resource for Rhesus Macaque (*Macaca mulatta*) genomics. *PLoS One*, **2**, e438.
- [13] Magness, C.L., Fellin, P.C., Thomas, M.J., Korth, M.J., Agy, M.B., Proll, S.C., Fitzgibbon, M., Scherer, C.A., Miner, D.G., Katze, M.G. *et al.* (2005) Analysis of the *Macaca mulatta* transcriptome and the sequence divergence between *Macaca* and human. *Genome Biology*, **6**, R60.
- [14] Hernandez, R.D., Hubisz, M.J., Wheeler, D.A., Smith, D.G., Ferguson, B., Rogers, J., Nazareth, L., Indap, A., Bourquin, T., McPherson, J. *et al.* (2007) Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science*, **316**, 240-243.
- [15] Bevilacqua, L., Doly, S., Kaprio, J., Yuan, Q., Tikkanen, R., Paunio, T., Zhou, Z., Wedenoja, J., Maroteaux, L., Diaz, S. *et al.* (2010) A population-specific HTR2B stop codon predisposes to severe impulsivity. *Nature*, **468**, 1061-1066.
- [16] Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods*, **7**, 248-249.
- [17] Xu, A.G., He, L., Li, Z., Xu, Y., Li, M., Fu, X., Yan, Z., Yuan, Y., Menzel, C., Li, N. *et al.* (2010) Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Comput Biol*, **6**, e1000843.
- [18] The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
- [19] Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A. *et al.* (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*, **362**, 1181-1191.
- [20] Wilke, C.O. (2004) Molecular clock in neutral protein evolution. *BMC Genet*, **5**, 25.
- [21] Eyre-Walker, A. and Keightley, P.D. (1999) High genomic deleterious mutation rates in hominids. *Nature*, **397**, 344-347.
- [22] Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol*, **5**, e254.
- [23] Bhangale, T.R., Rieder, M.J., Livingston, R.J. and Nickerson, D.A. (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet*, **14**, 59-69.
- [24] Livingston, R.J., von Niederhausern, A., Jegga, A.G., Crawford, D.C., Carlson, C.S., Rieder, M.J., Gowrisankar, S., Aronow, B.J., Weiss, R.B. and Nickerson, D.A. (2004) Pattern of sequence variation across 213 environmental response genes. *Genome Res*, **14**, 1821-1831.
- [25] Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet*, **22**, 239-247.
- [26] Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*, **22**, 231-238.
- [27] Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256-276.
- [28] Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297-304.
- [29] Kondrashov, A.S. (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human Mutation*, **21**, 12-27.
- [30] Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636-639.
- [31] Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D. and Long, J.C. (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet*, **62**, 1171-1179.

6 Supporting Materials

Table S1. Ancestry assay for rhesus macaque samples used in this study

SampleID	%Missing Data	Inferred cluster		90% Probability Interval	
		Chinese	India	Chinese	India
E13	0	0.006	0.994	(0.000,0.044)	(0.956,1.000)
E20	0	0.007	0.993	(0.000,0.046)	(0.954,1.000)
E78	0	0.005	0.995	(0.000,0.034)	(0.966,1.000)
E85	0	0.524	0.476	(0.392,0.653)	(0.347,0.608)
G14	0	0.013	0.987	(0.000,0.087)	(0.913,1.000)
G36	0	0.531	0.469	(0.395,0.664)	(0.336,0.605)
K09	0	0.004	0.996	(0.000,0.027)	(0.973,1.000)
K20	0	0.999	0.001	(0.991,1.000)	(0.000,0.009)
K29	0	0.003	0.997	(0.000,0.018)	(0.982,1.000)
K41	0	0.008	0.992	(0.000,0.052)	(0.948,1.000)
K44	0	0.006	0.994	(0.000,0.043)	(0.957,1.000)
K46	0	0.005	0.995	(0.000,0.034)	(0.966,1.000)
M18	0	0.019	0.981	(0.000,0.102)	(0.898,1.000)
M40	0	0.004	0.996	(0.000,0.027)	(0.973,1.000)

Table S2. Putative SNPs covered with sequence reads from ChIPseq and/or RNAseq

	Human SNPs having reads from			Macaque SNPs having reads from		
	ChIPseq	RNAseq	Both	ChIPseq	RNAseq	Both
Total	130914	35615	63499	386219	23109	53474
Intergenic	80482	11594	15385	240655	9457	19278
5Kbupstream	7404	131	2501	22057	403	3843
Intron	39458	8736	31681	113994	2267	14182
Exon	3570	15154	13932	9513	10982	16171
UTR	1966	9149	7317	4249	4815	6391
CDS	1604	6005	6615	5264	6167	9780
nsSNP	814	2513	2550	2453	1864	3050
synonymous	787	3485	4057	2781	4299	6718
nonsense	3	7	8	30	4	12

Table S3. SNPs from 1000 Genomes Project Pilot 2

	CEU.trio	YRI.trio
Total	3646764	4502439
Also in dbSNP	3239544(88.8%)	3446643(76.6%)
Located in intergenic	2131596(58.4%)	2599459(57.7%)
Located in 5Kb upstream	165384(4.5%)	210006(4.7%)
Located in intron	1336273(36.6%)	1674001(37.2%)
Located in UTR	31887(0.9%)	41315(0.9%)
Located in CDS	24192(0.7%)	32244(0.7%)
nsSNP	9696(40.1%)	12853(39.9%)
Synonymous	14506(60.0%)	19412(60.2%)

Table S4. The average nucleotide heterozygosity from 1000 Genomes Project Pilot 2

Population	SampleID	$\theta_{\text{SNP}} (\times 10^{-4})^*$
CEU	NA12891	7.16
CEU	NA12892	7.33
YRI	NA19239	9.26
YRI	NA19238	9.06

* Calculated as the heterozygous bases divided by the sequencing accessible genome size ($2.85 \times 10^9 \times 80\%$) using 1000 Genomes Project Pilot 2 data.

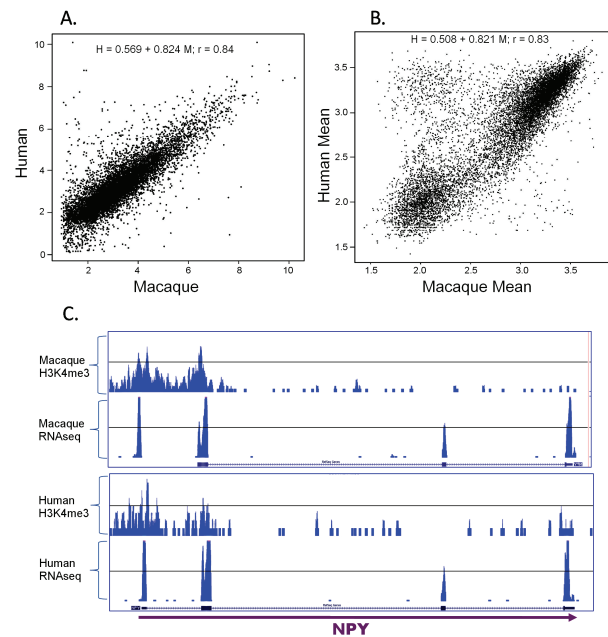


Fig. S1: A: RNAseq correlation between human vs macaque. Data points: mean of normalized gene expression level (log2). B: H3K4me3 ChIPseq correlation between human vs macaque. Data points mean of normalized area under curve (log10) of covered reads within 1Kb of TSS. C: Sequencing coverage (H3K4me3 ChIPseq and RNAseq) in NPY gene region.

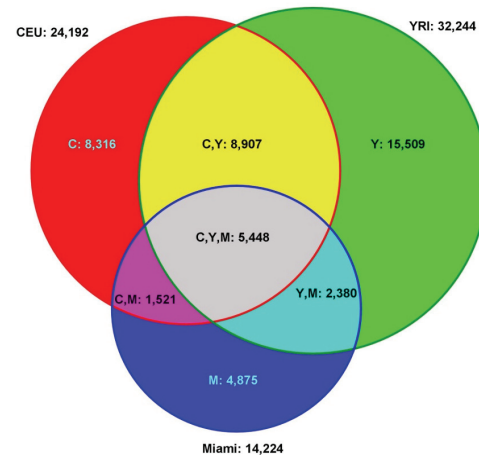


Fig. S2: Human cSNPs identified in 1000 Genomes Project Pilot 2 samples and this study. C or CEU: CEU trio from 1000 Genomes Project Pilot 2; Y or YRI: YRI trio from 1000 Genomes Project Pilot 2; M or Miami: 14 samples from Miami dataset in this study.

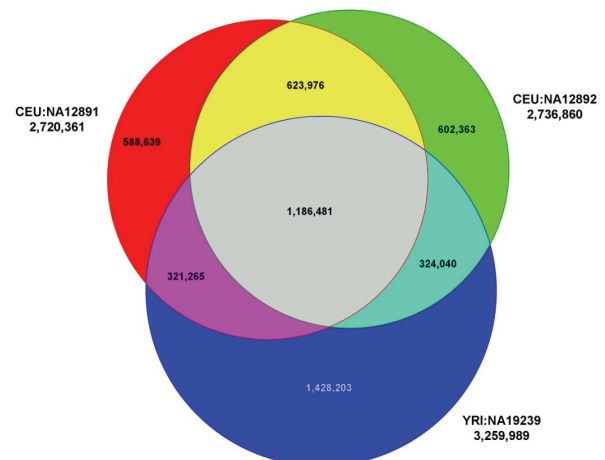


Fig. S3: SNPs shared between individuals in 1000 Genomes Project Pilot 2.

The Peak of a Pandemic? — A Phylogenetic Analysis of the H1N1 Influenza Virus from 2009 to Present

Anthony Deeter^{1,*}, Mark Dalman^{2,3,*}, Gayathri Nimishakavi¹, and Zhong-Hui Duan^{1,3,†}

¹Department of Computer Science, University of Akron, Akron, OH, USA 44325

²Department of Biology, University of Akron, Akron, OH, USA 44325

³Integrated Bioscience Program, University of Akron, Akron, OH, USA 44325

Abstract – *The swine origin influenza A (S-OIV) virus of 2009 reached pandemic proportions due to its novel sequence identity in human populations of North America and other localities. The S-OIV virus shows subtle change from 2009-2010 in humans, affirming that the HA and NA sequences have been unable to antigenically drift or shift enough to emerge as another pandemic. This study aimed to document the succession of S-OIV from 2009 to current in addition to investigating its relationship among other locations. Based on the phylogenetic analysis, the 2010 H1N1 is similar to other isolates circulating the previous year. Furthermore, the protein sequences with the highest non-synonymous to synonymous ratio were HA and NA thus indicating strong selective pressures for the antigen receptor binding sites to adapt even within human hosts.*

Keywords: 2009 influenza pandemic H1N1 influenza virus; antigenetic shift and drift; phylogenetic analysis; non-synonymous to synonymous ratio; neighbor joining method

1 Introduction

With the development of antibiotics within the past century, life expectancy has increased despite a prolonged window of susceptibility and transmissibility to viral and bacterial infections during humans' life span. Currently, with over a quarter of a million deaths and upwards of three million cases of influenza globally each year in humans, the emergence of a novel swine-origin influenza virus (S-OIV) in 2009 garnered much attention [1]. The common influenza (flu) is caused by the Orthomyxoviridae family of ssRNA viruses including Influenzavirus A, Influenzavirus B, Influenzavirus C, Isavirus and Thogotovirus. All but Isavirus are detected in vertebrates with Influenzavirus A seemingly the most virulent, diverse, and pathogenetic to humans. The present paper deals with the pandemic H1N1 flu virus, a novel subtype of Influenzavirus A.

Three months after its identification in Mexico in 2009, the S-OIV epidemic had reached alert phase 6, marking the first pandemic in almost forty years to reach that phase [1, 2]. The

S-OIV virus, despite its novel sequence and severity, is a triple reassortment from three different “donors” [3, 4]. Phylogenetic analyses conclude that of its eight nucleotide sequences, six of them (HA, PB2, PB1, PA, NP, NS) are highly similar to influenza viruses endemic to pigs in the late 1990's with the other two genes (NA and MP) from a bird lineage isolated in Europe [5-7]. None of the individual genes were previously found in Europe or North America, reaffirming conditions for a pandemic viral outbreak [3, 4, 8, and 9].

Therefore the early detection and continuous monitoring of novel strains in the environment are poised at the interface of molecular biology, viral biology, and, more recently, computer science. Furthermore, the inherent diversity, total number of sequences of Influenzavirus A and the lack of sampling resolution make phylogenetic analysis very complex. The present paper focuses on the antigenic shifting and drifting of the virus from 2009 to present and the post pandemic evolution of the 2009 H1N1 (S-OIV) Influenza virus.

2 Materials and Methods

In order to study the evolution of S-OIV since its emergence, phylogenetic trees were constructed using only unique, full length coding sequences, human host H1N1 nucleotide sequences from April 2009 to January 2011 [10]. The trees were constructed via the neighbor-joining method, with distances calculated using the Felsenstein F84 nucleotide method [11]. Non-synonymous (dN) to synonymous (dS) substitution ratio were then calculated using the Nei-Gojobori method [12]. The A/California/04/2009(H1N1) isolate was used as the first identified S-OIV strain (highlighted in green in figures) and 2010 strains were highlighted in red. A/Moscow/01/2009(H1N1), A/Boston/653/2009(H1N1), A/Korea/CJ01/2009(H1N1), A/Chile/4064/2009(H1N1), A/MexicoCity/WRAIR1752N/2010(H1N1), A/Newark/INS429/2010(H1N1), A/Vienna/INS179/2010(H1N1), and A/Ontario/3620/2010 were randomly selected in calculating dN/dS ratios. In constructing the phylogenetic trees, only unique isolates were used and a Perl script was written to

*both authors contributed equally; †duan@uakron.edu

randomly select a sample of 1000 sequences. The A/United Kingdom/1/1933(H1N1) is used as the outgroup. The A/California/04/2009(H1N1) and 2010 sequences were highlighted in each tree for reference and closely related branches were collapsed for tree readability.

3 Results and Discussion

Phylogenetic trees were constructed using neighbor-joining method to understand how the novel S-OIV (H1N1) influenza virus A strain has changed through its pandemic period (April 2009 - January 2011) in human hosts. The phylogenies for the unique protein coding sequences HA, NA, M1, M2 and PB2 are shown in Figures 1-5 (phylogenies

for other coding sequences are not shown here). These trees show the genotypic variation of encoded proteins in H1N1 influenza virus A. Our results indicate that the A/California/04/2009(H1N1) strain is genetically similar to 2010 isolated strains and is always present within every tree. Furthermore A/United Kingdom/1/1933 (H1N1) is the outgroup for all trees. Interestingly, 2010 isolates are genetically similar to the previous 2009 isolates once again reiterating that most epidemic H1N1 stem from circulating viral reservoirs. Surprisingly, a large polytomy occurred within the 2009 pandemic and 2010 isolates are much more diverse from one another, yet still similar to ones from 2009 than any other sequences.



Figure 1. Phylogenetic relationships among human H1N1 viruses (HA)

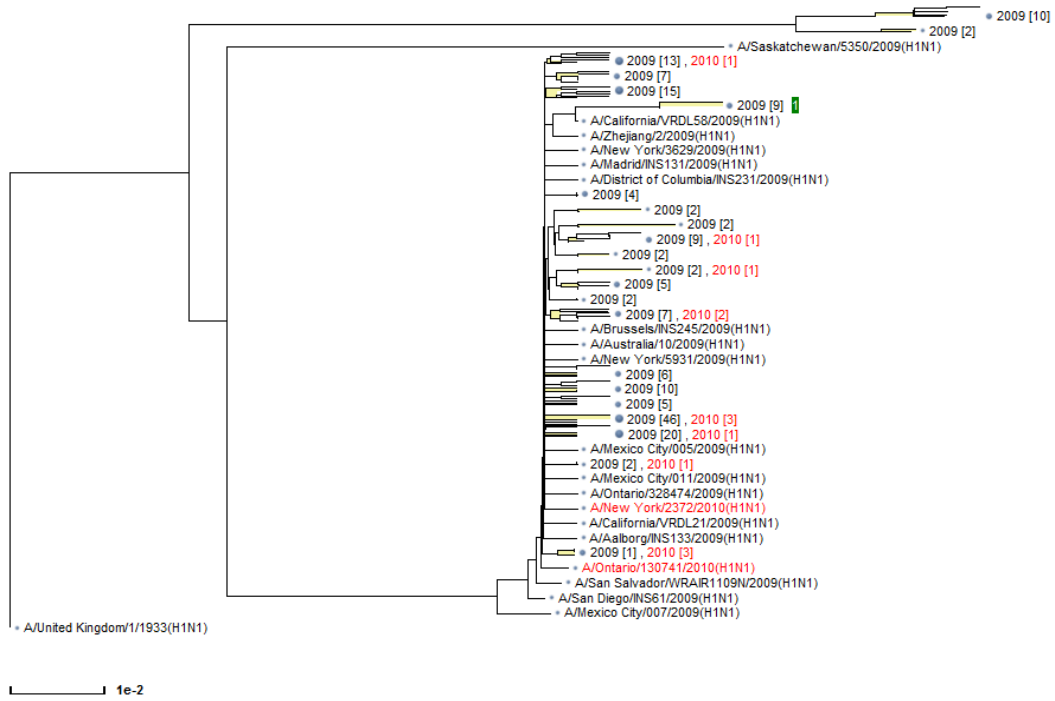


Figure 2. Phylogenetic relationship among human H1N1 viruses (M2)

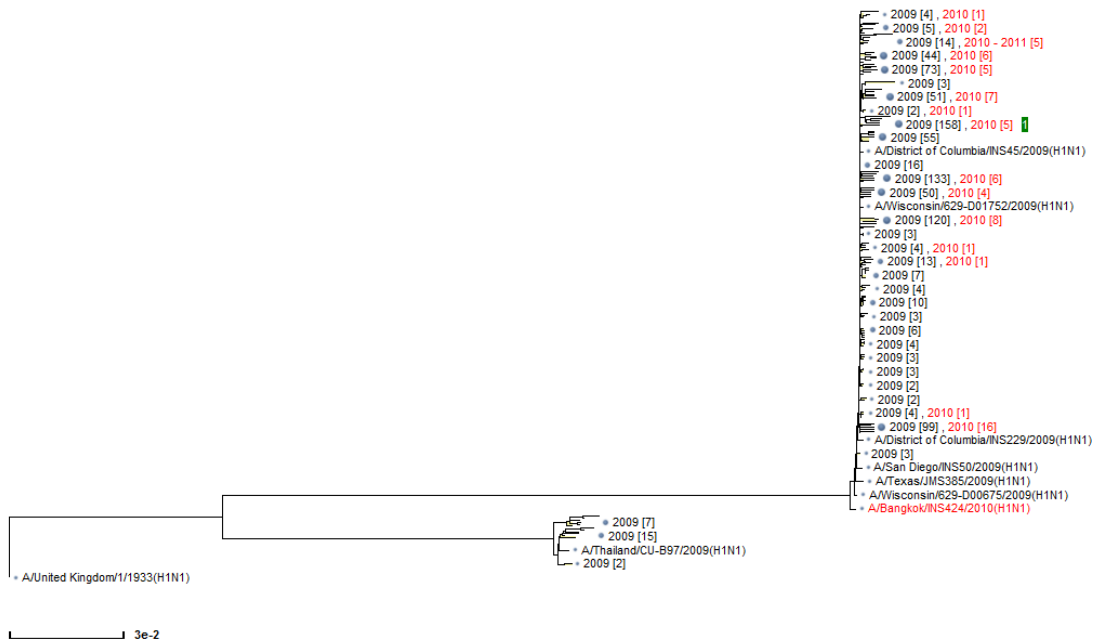


Figure 3. Phylogenetic relationship among human H1N1 viruses (NA)

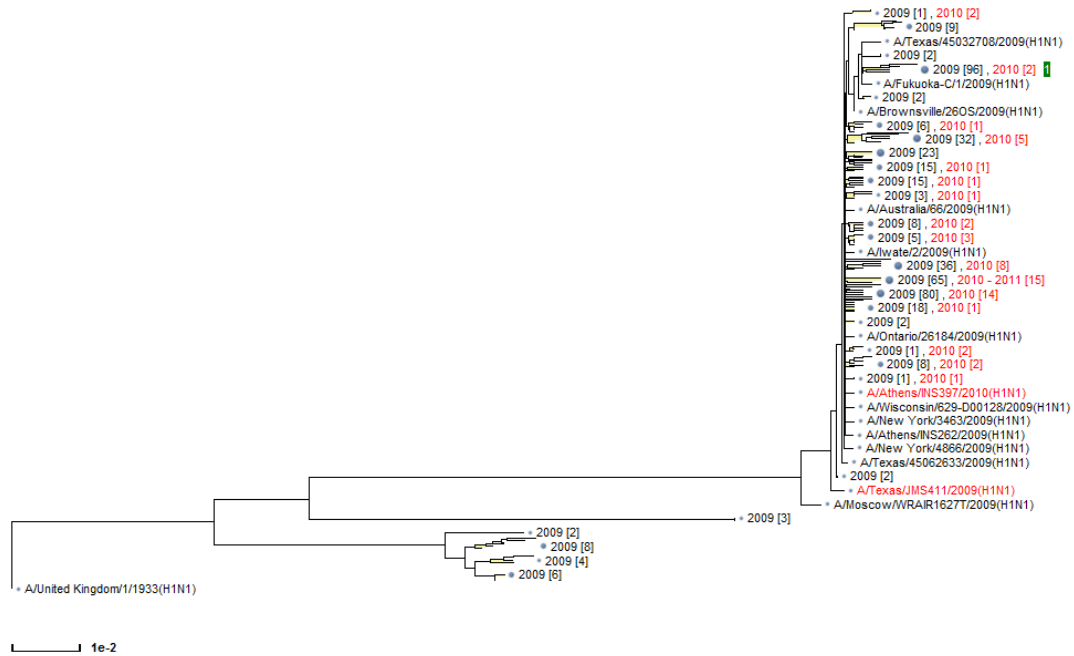


Figure 4. Phylogenetic relationship among human H1N1 viruses (M1)

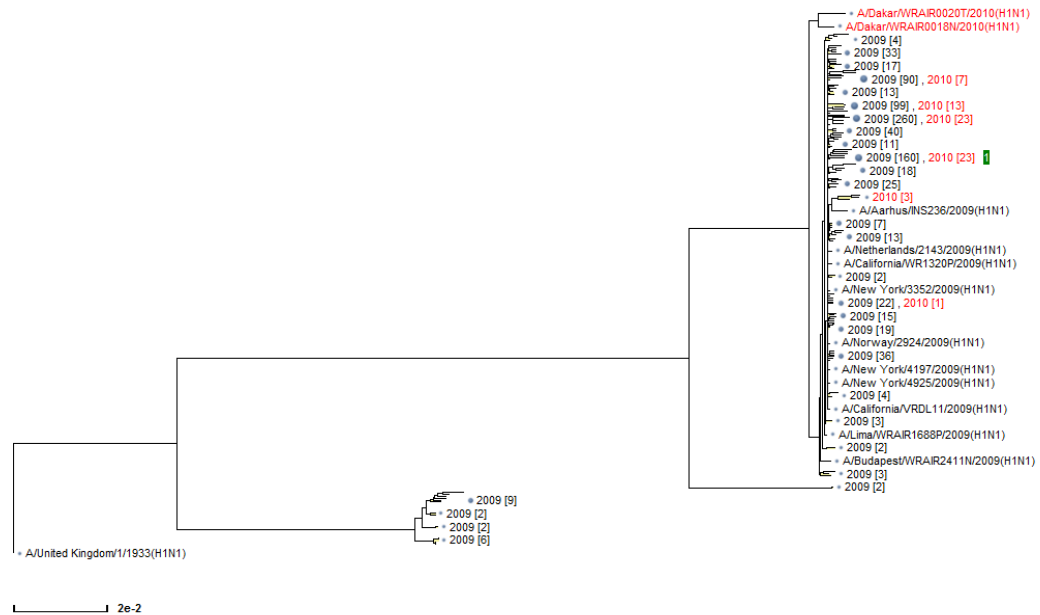


Figure 5. Phylogenetic relationship among human H1H1 viruses (PB2)

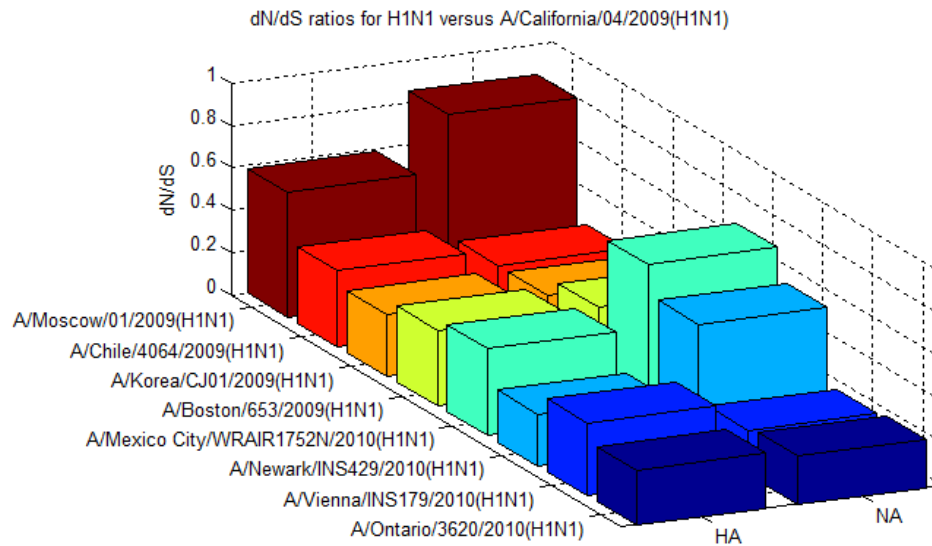


Figure 6. Non-synonymous to synonymous (dN/dS) ratios for selected human H1N1 isolates from 2009 and 2010.

To quantitatively identify the changes in primary sequence from 2009 to 2010, non-synonymous to synonymous (dN/dS) substitution ratio were calculated to identify whether changes in nucleotide sequence actually resulted in changes in primary sequence. Unusually high number of non-synonymous substitutions is widely accepted as a result of positive selection [13]. Within our case study, the non-synonymous substitutions between A/California/04/2009 (H1N1) strain and the 2010 sequences of HA and NA are considerably smaller than their respective synonymous ones. Ratios of less than one across all 2010 isolates suggest concrete evidence as to why no genetically dissimilar S-OIV isolates have arisen (Figure 6). Furthermore, in comparing randomly selected 2009 and 2010 sequences to A/California/04/2009(H1N1), there was less variation between A/California/04/2009(H1N1) and more recent 2010 sequences than amongst 2009 sequences (0.424 vs 0.371, respectively). This suggests there was more genetic variation among the initial outbreak than subsequently documented in 2010 and even more so, this signifies that the strains identified are likely small antigenic drifts from other viruses circulating in 2009.

Genetically dissimilar and novel isolates to a population are the cruxes of a pandemic. Additionally, the dN/dS ratios of HA and NA are considerably larger than those of other six proteins (not shown in paper); suggesting that as these ratios increase, the prevalence of new coded amino acids will increase. A new amino acid can be the difference between viral detection and infection [14-16]. However, the direction of selection is not well articulated within dN/dS ratios and begs the question of whether neutral theory is the evolutionary process underlying epidemic viral outbreaks

and the “perfect storm” reassortments in pigs and birds causing pandemic outbreaks [17].

By crossing the positive selection threshold, the possibility of novel strains of influenza virus A increase, requiring new vaccines. In other words, as the dN/dS ratio goes above one, the novelty of its structure begins to be advantageous to the virus’s transmission, ultimately increasing its fitness. Despite research indicating that selective pressures will increase non-synonymous substitutions, the lack of biochemical and evolutionary data is not in accordance. For all proteins, there are both essential and nonessential amino acids, those which are responsible for function and those that are not. HA and NA are two proteins on the viral coat which, by being genetically different through selective pressures from innate and adaptive immunity, can cause a pandemic. Yet proteins within the virion that are not as plastic show little variation from one host or year to the next (Figure 2). The later example is similar to most proteins in the human body in which there are areas capable of nonsynonomous substition and areas that have conserved sequences. Consequently, in viral biology producing a genetically different coat is advantageous as opposed to maintaining the status quo and being eradicated. An inability to infect (highly detected) or novelize (highly virulent) may both result in the eradication of the strains from the gene pool.

4 Conclusions

The swine origin influenzavirus A S-OIV pandemic of 2009 has a unique genetic composition as suggested by

almost a century of viral data. Our study reveals that despite being phylogenetically similar to 2010 influenza viruses in human, the dN/dS ratios indicate that the surface proteins HA and NA do antigenetically drift fastest amongst human hosts. Furthermore, the dN/dS ratios suggest that sequences during 2009 are significantly more dissimilar than recent 2010 isolates, suggesting that the 2009 S-OIV pandemic might have peaked during the summer of 2009.

Future studies should comparatively measure the substitution rates amongst host types and by locations to further elucidate whether avian and swine lineages are the most capable and dominating viral incubators or whether attention should be focused at a more macroscopic regional or continental understanding of viral transmission. Thus further research of immunoinformatics will increase the interdisciplinary understanding of viral transmission, vaccination, documentation, and retrieval.

5 References

- [1] WHO Pandemic (H1N1) 2009. <http://www.who.int/csr/disease/swineflu/en/>. Retrieved on 2/12/2011.
- [2] Christophe Fraser, Christl A. Donnelly, Simon Cauchemez, William P. Hanage, Maria D. Van Kerkhove, T. Déirdre Hollingsworth, Jamie Griffin, Rebecca F. Baggaley, Helen E. Jenkins, Emily J. Lyons, Thibaut Jombart, Wes R. Hinsley, Nicholas C. Grassly, Francois Balloux, Azra C. Ghani, Neil M. Ferguson, Andrew Rambaut, Oliver G. Pybus, Hugo Lopez-Gatell, Celia M. Alpuche-Aranda, Ietza Bojorquez Chapela, Ethel Palacios Zavala, Dulce Ma. Espejo Guevara, Francesco Checchi, Erika Garcia, Stephane Hugonnet, Cathy Roth, and The WHO Rapid Pandemic Assessment Collaboration. Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings; *Science*, 324:1557-1561, June 2009.
- [3] F. S. Dawood, S. Jain, L. Finelli, M. W. Shaw, S. Lindstrom, R. J. Garten, L. V. Gubareva, X. Xu, C. B. Bridges, T. M. Uyeki, and Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. Emergence of a Novel Swine-origin Influenza A (H1N1) Virus in Humans; *New England Journal of Medicine*, 360:2605-2615, June 2009.
- [4] Gavin J. D. Smith, Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J. Lycett, Michael Worobey, Oliver G. Pybus, Siu Kit Ma, Chung Lam Cheung, Jayna Raghwan, Samir Bhatt, J. S. Malik Peiris, Yi Guan, and Andrew Rambaut. Origins and Evolutionary Genomics of the 2009 Swine-origin H1N1 Influenza A Epidemic; *Nature*, 459:1122-1125, June 2009.
- [5] Mikhail Matrosovich, Alexander Tuzikov, Nikolai Bovin, Alexandra Gambaryan, Alexander Klimov, Maria R. Castrucci, Isabella Donatelli, and Yoshihiro Kawaoka. Early Alterations of the Receptor-Binding Properties of H1, H2, and H3 Avian Influenza Virus Hemagglutinins after Their Introduction into Mammals; *Journal of Virology*, 74: 8502-8512, September 2000.
- [6] Gabriele Neumann, Takeshi Noda, and Yoshihiro Kawaoka. Emergence and Pandemic Potential of Swine-Origin H1N1 Influenza Virus; *Nature*, 459(7249): 931-939, June 2009.
- [7] Adrian J Gibbs, John S Armstrong, and Jean C Downie. From Where Did The 2009 'Swine-Origin' Influenza A Virus (H1N1) Emerge? *Virology Journal*, 6(207): 1-11, November 2009.
- [8] Alan J. Hay, Victoria Gregory, Alan R. Douglas and Yi Pu Lin. The Evolution of Human Influenza Viruses; *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 356:1861-1870, December 2001.
- [9] Troy Day, Jean-Baptiste André, and Andrew Park. The Evolutionary Emergence of Pandemic Influenza; *Proceedings: Biological Sciences*, 273(1604): 2945-2953, December 2006.
- [10] NCBI Influenza Virus Sequence Database. <http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi?go=1>. Data retrieved: March, 15, 2011.
- [11] Joseph Felsenstein and Gary A. Churchill. A Hidden Markov Model Approach to Variation among Sites In Rate of Evolution. *Molecular Biology and Evolution*, 13:93-104, January 1996.
- [12] Masatoshi Nei and Takashi Gojoborit. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions; *Molecular Biology and Evolution*, 3:418-426, September 1986.
- [13] Sergei L. Kosakovsky Pond, Art F.Y. Poon, Andrew J. Leigh Brown, and Simon D.W. Frost. A Maximum Likelihood Method for Detecting Directional Evolution In Protein Sequences And Its Application To Influenza A Virus; *Molecular Biology Evolution*, 25(9):1809-1824 September 2008.

- [14] Christoph Scholtissek. Source for Influenza Pandemics; *European Journal of Epidemiology*, 10(4): 455-458, August 1994.
- [15] G. G. Brownlee and E. Fodor. The Predicted Antigenicity of the Haemagglutinin of the 1918 Spanish Influenza Pandemic Suggests an Avian Origin; *Philosophical Transactions: Biological Sciences*, 356(1416): 1871-1876, December 2001.
- [16] D. R. Perez, R. J. Webby, E. Hoffmann, R. G. Webster. Land-Based Birds As Potential Disseminators of Avian Mammalian Reassortant Influenza A Viruses. *Avian Diseases*, 47: 1114-1117, 2003.
- [17] Tomoko Ohta. The Nearly Neutral Theory of Molecular Evolution; *Annual Review of Ecology and Systematics*, 23: 263-286, 1992.

Protease Complement of the Thermophilic Bacterium *Coprothermobacter proteolyticus*

Hong Cai¹, Jianying Gu^{2,*}, Yufeng Wang^{1,*}

¹ Department of Biology, South Texas Center for Emerging Infectious Diseases
University of Texas at San Antonio, San Antonio, TX 78249, USA
hong.cai@utsa.edu, yufeng.wang@utsa.edu (*corresponding authors)

² Department of Biology, College of Staten Island
City University of New York, Staten Island, NY 10314, USA
jianying.gu@csi.cuny.edu

Abstract—Thermal bacteria that live in higher temperature have been considered as good candidates for bioremediation and processing of protein-rich wastewater. However, very little is known about the proteases, the enzymes that digest the protein wastes in these organisms. In this study, we present a comparative genomic analysis of the protease complement in a thermal bacterium *Coprothermobacter proteolyticus*. The proteases common to a group of thermophilic bacteria have been identified, providing a short list of important enzymes for experimental characterizations.

Keywords- genome, protease, *Coprothermobacter*, degradome, bioinformatics, gene family

1. INTRODUCTION

Bacteria, a member of the Domains of life, mediates the fundamental geochemical cycles that sustain life on earth. These microorganisms live in diverse habitats and environments. Although some bacteria are human pathogens, the majority of bacteria species are harmless and some have important applications in biotechnology. For example, bacteria that are capable of degrading organic compounds have been used in bioremediation and waste processing in industry.

The advent of high throughput genomic technology and the development of effective bioinformatics data mining approach have provided an unprecedented opportunity to investigate the adaption and evolution of bacteria. Previously uncharacterized organisms can now be explored at a genome level. This study is focused on a bioinformatics characterization of gene families in an understudied bacterium *Coprothermobacter proteolyticus* (strain DSM 5265). This bacterium is anaerobic. Its most important feature is the high growth temperature (about 63°C). It was first isolated from a thermophilic digester for fermenting water wastes and animal manure. Wastewater often contains proteins. *Coprothermobacter proteolyticus* was found to have strong protease activity to degrade proteins and peptides [1, 2]. Here we report a comprehensive survey of the protease complement (or degradome) in the genome of *C. proteolyticus*, which may be a good candidate for facilitating waste water processing under high temperature.

2. METHODS

A total of over 34,000 sequences of characterized and predicted proteases were obtained from the Merops database (<http://www.merops.ac.uk>) [3]. These sequences were searched

against the *C. proteolyticus* predicted protein sequences using BLASTP with default settings and an E-value cutoff of less than 10⁻⁵ for defining protease homologs. Partial sequences (less than 80% of fulllength) and redundant sequences were excluded. The domain/motif organization of predicted *C. proteolyticus* proteases was revealed by an InterPro search. For each putative protease, the known protease sequence or domain with the highest similarity was used as a reference for annotation; the catalytic type and protease family were predicted in accordance with the classification in Merops, and the enzyme was named in accordance with SWISS-PROT enzyme nomenclature (<http://www.expasy.ch/cgi-bin/lists?peptidas.txt>) and the literature.

3. RESULTS

One of the most prominent physiological features of the anaerobic thermophilic *Coprothermobacter proteolyticus*, formerly *Thermobacteroides proteolyticus*, is its well-documented proteolytic activity [1, 2]. Although proteolytic activity is common in the anaerobic bacteria that are mesophilic, it is observed in only a few thermophiles [4-7]. *C. proteolyticus* has attracted the attention of researchers interested in its potential applications in high temperature environments, including the treatment of protein-rich wastewater, for example. Despite this interest, however, not a single protease in *C. proteolyticus* has been systematically characterized at the biochemical and molecular level to date.

Our comparative genomic analysis revealed that its proteolytic repertoire (degradome) consists of a total of 59 protease homologs, which account for approximately 1.9% of the proteome (Table 1). The fraction of proteases in the *C. proteolyticus* genome is close to the average observed in the 1,569 organisms with completed genomes (2.6%). Using the Merops protease nomenclature, which is based on intrinsic evolutionary and structural relationships [3], the *C. proteolyticus* proteases were divided into four known and one unknown catalytic classes that encompass 38 families. These families include: Two aspartic protease families and five cysteine protease families, each represented by a single member; 24 metalloproteases belonging to 17 families, 23 serine proteases belonging to 12 families, and two families (five proteases) with unknown catalytic types. Clearly, gene duplication occurred at a very small scale during the evolution of *C. proteolyticus* proteases, which accounts for the large number of singletons.

A glance at the *C. proteolyticus* degradome reveals some significant features. The entire catalytic class of proteasome-specific threonine proteases is missing, which is consistent with the observation that the proteasome is absent. *C. proteolyticus* has an abundant catalog of metalloproteases (40.7%) and serine proteases (40.0%), compared to aspartic (3.4%) and cysteine proteases (8.5%). The most abundant protease family, serine protease subtilisin (S8), has 6 members. Interestingly, many subtilisins that have been characterized are thermostable [8-10].

The lineage specific expansion of subtilisins in *C. proteolyticus* is likely to be adaptive: at least two subtilisins (COPRO5265_1473 and COPRO5265_1474) are the products of one tandem gene duplication event. Specifically, two subtilisins (COPRO5265_1474 and COPRO5265_1431) are extracellular Vpr peptidases. Vpr was previously only found in a number species from the *Bacillales* [11]; the homologs found in *C. proteolyticus* expand the range of Vpr to the *Clostridiales*.

Table 1. Protease complements in *Coprothermobacter proteolyticus* and other model organisms.

Organism	Catalytic Class					Total	Percentage of the Proteome ^a
	Aspartic	Cysteine	Metallo	Serine	Threonine		
<i>Coprothermobacter proteolyticus</i>	2 (3.4%) ^b	5 (8.5%)	24 (40.7%)	23 (40.0%)	0 (0%)	59 ^c	1.9
<i>Neurospora crassa</i>	19 (8.1%)	41 (17.4%)	81 (34.5%)	75 (31.9%)	19 (8.1%)	235	2.4
<i>Saccharomyces cerevisiae</i>	19 (11.1%)	41 (24.0%)	57 (33.3%)	38 (22.2%)	16 (9.4%)	171	2.4
<i>Caenorhabditis elegans</i>	27 (5.6%)	125 (25.9%)	190 (39.4%)	115 (23.9%)	25 (5.2%)	482	2.4
<i>Drosophila melanogaster</i>	46 (6.2%)	86 (11.5%)	207 (27.7%)	373 (49.9%)	35 (4.7%)	747	5.4
<i>Homo sapiens</i>	320 (29.3%)	190 (17.4%)	252 (23.0%)	291 (26.6%)	41 (3.7%)	1,094	4.5
<i>Arabidopsis thaliana</i>	233 (27.6%)	162 (19.2%)	112 (13.3%)	306 (36.2%)	31(3.7%)	849	3.1

^a. The percentage of the whole genome that encodes putative proteases.

^b. Percentage of individual catalytic class in the protease complement is included in parentheses.

^c. The total proteases in *Coprothermobacter proteolyticus* includes 5 protease homologs with unknown classifications.

C. proteolyticus possesses a core degradome structure that may be common in the thermophilic bacteria, as shown by comparison with *Moorella thermoacetica* and *Thermoanaerobacter tengcongensis*, which are the most closely related sequenced species in the family *Thermoanaerobacteriaceae* to have a detailed analysis of its proteases published in Merops [12]. Nineteen protease families are present in all the three organisms. For example, at least three proteases may be actively involved in the secretion system: signal peptidase I (S26) typically processes newly-synthesized secreted proteins by removing the hydrophobic signal peptides when the precursors are translocating the membrane; the bacteria-specific signal peptidase II (A8) is membrane bound and it plays an important role in the production of cell wall by removing the signal peptide from the murein prolipoprotein; type IV prepilin peptidase (A24) processes prepeptides by removing leader peptides. Fifteen protease families found in *C. proteolyticus* are also present in either *Moorella thermoacetica* or *Thermoanaerobacter tengcongensis*, but not both. Four protease families are uniquely present in *C. proteolyticus*. They are papain (C1), dipeptidase A (C69), RTX toxin (M6), and carboxypeptidase Taq (M32). Among them, Taq (M32), by its presence in a

variety of thermophiles and hyperthermophiles [13], has a demonstrated ability to tolerate high temperatures. While the RTX toxin was implicated in several bacterial pathogens to be a virulence factor as host immune inhibitor, its role in the non-pathogenic *C. proteolyticus* remains unclear [14].

2. CONCLUSIONS

We performed a comparative genomic study of the proteases in thermophilic bacterium *Coprothermobacter proteolyticus*. These enzymes play important roles in digesting and breaking down proteins and peptides into smaller fragments. Functional characterization of these enzymes in this bacterium may provide a better understanding of the mechanisms of physiological adaptation to hot temperature and a better assessment of its potential application to wastewater processing.

ACKNOWLEDGMENT

This work is supported by NIH grant AI067543 to YW, and the PSC-CUNY Research Award PSCREG-39-497 to JG.

REFERENCES

- [1] Ollivier BM, Mah RA, Ferguson TJ, Boone R, Garcia JL, Robinson R (1985) Emendation of the genus *Thermobacteroides*: *Thermobacteroides proteolyticus* sp. nov., a proteolytic acetogen from a methanogenic enrichment. *Int J Syst Bacteriol.* 35: 425-428.
- [2] Rainey FA, Stackebrandt E (1993) Transfer of the type species of the genus *Thermobacteroides* to the genus *Thermoanaerobacter* as *Thermoanaerobacter acetothylacus* (Ben-Bassat and Zeikus 1981) comb. nov., description of *Coprothermobacter* gen. nov., and reclassification of *Thermobacteroides proteolyticus* as *Coprothermobacter proteolyticus* (Ollivier et al. 1985) comb. nov. *Int. J. Syst Bacteriol.* 43: 857-859
- [3] Rawlings ND, Barrett AJ, Bateman A. (2009) MEROPS: the peptidase database. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D227-33.
- [4] Toda Y, Saiki T, Uozumi T, Beppu T (1988) Isolation and characterization of a protease-producing, thermophilic anaerobic bacterium, *Thermobacteroides leptospartum* sp. nov. *Agric Biol Chem* 52: 1339-1344.
- [5] Engle M, Li Y, Rainey F, DeBlois S, Mai V, Reichert A, Mayer F, Messner P, Wiegel J (1996) *Thermobrachium celere* gen. nov., sp. nov., a rapidly growing thermophilic, alkalitolerant, and proteolytic obligate anaerobe. *Int J Syst Bacteriol* 46: 1025-1033.
- [6] Tarlera S., Mux L, Soubes M, Stams AJM (1997) *Caloramator proteoclasticus* sp. nov., a new moderately thermophilic anaerobic proteolytic bacterium. *Int J Syst Bacteriol* 47: 651-656.
- [7] Etchebehere C, Pavan ME, Zorzopulos J, Soubes M, Muxi L. (1998) *Coprothermobacter platensis* sp. nov., a new anaerobic proteolytic thermophilic bacterium isolated from an anaerobic mesophilic sludge. *Int J Syst Bacteriol.* 48:1297-1304.
- [8] Siezen RJ, Leunissen JA (1997) Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Sci.* 6:501-523.
- [9] Toogood HS, Smith CA, Baker EN, Daniel RM. (2000) Purification and characterization of Ak.1 protease, a thermostable subtilisin with a disulphide bond in the substrate-binding cleft. *Biochem J.* 350:321-328.
- [10] Godde C, Sahn K, Brouns SJ, Kluskens LD, van der Oost J, de Vos WM, Antranikian G. (2005) Cloning and expression of islandisin, a new thermostable subtilisin from *Fervidobacterium islandicum*, in *Escherichia coli*. *Appl Environ Microbiol.* 71:3951-3958.
- [11] Corvey C, Stein T, Dusterhus S, Karas M, Entian KD (2003) Activation of subtilin precursors by *Bacillus subtilis* extracellular serine proteases subtilisin (AprE), WprA, and Vpr. *Biochem Biophys Res Commun.* 304: 48-54.
- [12] G. Bao Q, Tian Y, Li W, Xu Z, Xuan Z, Hu S, Dong W, Yang J, Chen Y, Xue Y, Xu Y, Lai X, Huang L, Dong X, Ma Y, Ling L, Tan H, Chen R, Wang J, Yu J, Yang H. (2002) A complete sequence of the *T. tengcongensis* genome. *Genome Res.* 12:689-700.
- [13] Motoshima H, Kaminogawa S (2004) Carboxypeptidase Taq. In *Handbook of Proteolytic Enzymes*, 2nd edition (Barrett,A.J., Rawlings,N.D. & Woessner,J.F. eds), p.776-778, Elsevier, London.
- [14] Dalhammar G, Steiner H (1984) Characterization of inhibitor A, a protease from *Bacillus thuringiensis* which degrades attacins and cecropins, two classes of antibacterial proteins in insects. *Eur J Biochem* 139: 247-252.

Computational analysis on *Cuminum cyminum* compounds against aldose reductase as anti-diabetic agents

Naresh Babu Muppalaneni¹, Allam Appa Rao²

¹ Associate Professor, Department of Computer Science & Engineering, Avanathi's Research & Technological Academy

² Vice-Chancellor, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, INDIA

Abstract

*Various proteins play important roles in diabetes and a number of plants have been tested for their efficacy in modulating diabetes. Of all the proteins, we selected aldose reductase enzyme to analyze few plant compounds computationally for their efficacy towards protein inhibition. A total of 85 compounds from different parts of a plant, *Cuminum cyminum* were studied. Analysis was conducted using Molegro Virtual Docker software as docking program and the molecules drawn in ISIS Draw software are energy minimized using cosmic - optimize 3D module of Tsar (Tools for structure activity relationships) software. Before docking plant compounds, software validation was performed and found that the co-crystallized ligand was within 2.0 Å. Further, docking and re-scoring of top ten compounds with GOLD, AutoDock vina, eHiTS, PatchDock and MEdock followed by rank-sum technique revealed high binding affinity of compound Apigetrin.*

Keywords—Computer Science, Computer Application, Computer Aided Drug Design, type 2 Diabetes, Docking, GOLD, Molegro, aldose reductase

1. Introduction

Human body gets energy by making glucose from foods like bread, rice, potatoes etc., To use this glucose human body needs insulin. Insulin is hormone that helps the body control the level of glucose in the body. Type 2 diabetes is disease in which pancreas does not produce enough insulin or body may not utilize insulin produced. Diabetes mellitus is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both. The chronic hyperglycemia of diabetes is associated with long-term damage, dysfunction, and failure of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels. [1].

Furthermore, the researchers suggested that high intakes of plant foods and low intakes of meat products may help high blood pressure treatment and proper insulin levels and hence these benefits can be

linked to the presence of specific compounds in plants. Various plants have been tested for their efficacy in modulating diabetes, however, when literature was searched for computer-aided docking studies on compounds from plants vs proteins that mediate diabetes, very few reports were found to contain the required information. Also, many virtual screening studies have been reported in literature stating the importance of dataset, algorithms and scoring functions, whereas none of the works contain screening compounds from plants. This provided us the rationale to screen plant based compounds using Molegro Virtual Docker software. Hence, in this paper we report screening various compounds from *Cuminum cyminum* against Aldose reductase extracted from Protein Data Bank (PDB).

2. Materials And Methods

2.1 Virtual Screening

Virtual screening utilizes docking and scoring of each compound from a dataset and the technique employed is based on the prediction of binding modes and binding affinities of each compound in the dataset by means of docking to an X-ray crystallographic structure [2]. Some recent studies [3] have focused on certain crucial factors such as the size and diversity of the ligand dataset, wide range of targets and the evaluation of docking programs. Taking these aspects into consideration, diverse compounds from seven plants and three protein targets are evaluated.

However, in general, it is important to visualize the docked poses of high-scoring compounds because many ligands are docked in different orientations and may often miss interactions that are known to be important for the target receptor. This sort of study becomes more difficult as the size of the dataset increases. Therefore, an alternative approach is to eliminate unpromising compounds before docking by restricting the dataset to drug-like compounds; by filtering the dataset based on appropriate property and sub-structural features and by performing diversity analysis [4].

2.2 Data Set

Chemical compound names from each plant were obtained from Dukes Ethnobotany (<http://www.ars-grin.gov/duke/>) and the respective structures are searched in various literature databases. This resulted in 85 compounds, selected based on the property and sub-structural features, from *Cuminum cyminum* were drawn using ISISDraw software (www.mdli.com). The 2D structures are converted into 3D structures by using corina 3D analysis tool in Tsar (Tools for structure activity relationships) software (www.accelrys.com). The geometries of these compounds were optimized using cosmic optimize 3D module and the charges were added. All molecules were written as mol2 files.

2.3 Receptor X-ray structure

The X-ray crystal structure of Aldose reductase, 1AH3, in complex with inhibitor was recovered from Protein Data Bank. We used the molecular docking program Molegro Virtual Docker (MVD) for virtual ligand screening based on docking, and a consensus scoring and ranking was employed to generate classes using MolDock score of Molegro software respectively.

2.4 Molegro Docking

Water molecules were discarded from the pdb file, added hydrogens and missing side chains were reconstructed. Automated docking studies were then performed using the genetic algorithm to explore the full range of ligand conformational flexibility and the rotational flexibility of selected receptor hydrogens. The docking poses are ranked based on a scoring function, MolDock score. The scoring scheme was derived from PLP [Piecewise Linear Potential] scoring functions originally proposed by Gehlhaar et al [5] and later extended by Yang et al [6]. In the present work, the binding site was defined as a spherical region which encompasses all protein atoms within 15.0 Å of each crystallographic ligand atom. Default settings were used for all calculations.

Before screening plant compounds, the docking protocol was validated. 1AH3 with bound ligand was docked individually into its corresponding binding pocket to obtain the docked pose and the RMSD of all atoms between these two conformations was 0.87 Å (Table 1) indicating that the parameters for docking simulation are good in reproducing the X-ray crystal structure.

Table1: Table showing the RMSD values of 1AH3 in three runs.

SINo	PDB ID	Run1	Run2	Run3
1	1AH3	0.8736	0.8721	0.670

2.5 Consensus Scoring and Ranking

Generally, docking programs have the ability to predict the experimental orientations of protein-ligand

complexes. The ability to predict the ideal binding mode of a ligand and to differentiate correct poses from incorrect ones is based on reliable scoring functions. However, it has been reported that various combinations of scoring functions would reduce errors when compared to single scoring scheme which improves the probability of identifying true hits [7]. Thus, it has been demonstrated that consensus scoring is generally more effective than single scoring for molecular docking [8,9] and represented an effective way in getting improved hit rates in various virtual database screening studies [10]

In our study, we tested three different scoring functions such as GOLD score of GOLD docking routine, dock score implemented in eHiTS (electronic High Throughput Screening) and MolDock score of Molegro software respectively. Docking program GOLD was used to dock compounds to generate an ensemble of docked conformations and each scoring function is applied to generate classes based on the obtained dock scores followed by ranking the best conformations. During ranking, signs of some scoring functions are changed to make certain that a lower score always indicates a higher affinity.

3. Results

Dock runs of 85 compounds on protein 1AH3 using MVD resulted in few best compounds that were evaluated based on the binding compatibility [docked energy (kcal/mol)] with the receptor. The software generated 5 conformers for each docked molecule and in each case, binding energies greater than the co-crystallized ligand were only selected.

Dock scores of co-crystallized ligand of 1AH3 run in triplicates are within -105.52 to -107.01 kcal/mol, respectively, and hence any molecule from the dataset that result in scores higher than the range are considered more appropriate. Therefore, in the first step, virtual screening with docking and scoring resulted in few best hits [Table-2]. In the second step, consensus scoring was applied to generate different scores for these compounds. Likewise, re-scoring docking poses with independent functions is another valuable approach which gained prominence in recent studies. Therefore, re-scoring of best docked poses based on their interaction energies with respective protein active site residues was done using MolDock score scoring function.

Table 2: Table showing the dock scores of best compounds from *Cuminum cyminum*

S.No.	Compound	Affinity (kcal/mol)
1	Riboflavin	-133.388
2	Apigenin-5-o-glucoside	-131.705
3	Apigetrin	-130.833
4	Apiin	-127.982
5	Benzyl Cinnamate	-117.458

6	Luteolin	-116.643
7	Stigmasterol	-116.379
8	Cosmosin	-115.701
9	Luteolin-7-o-glucoside	-112.54
10	Cynaroside	-111.372

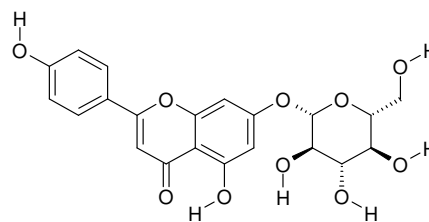


Figure 1: 2-dimensional structure of Apigetrin

4. Discussion

In our study, we tested seven different scoring functions such as GOLD, Molegro, AutoDock vina (Windows platform), e-HiTS (Linux platform) and PathDock, MEdock (docking servers). Re-scoring was carried out using all the above scoring functions and each molecule was optimized using optimization routine. Post-scoring results are evaluated for RMSD (Root Mean Square Deviation) and found to be within 2Å°. In all the above cases, ranking was done individually by clustering best scored compounds into equally split four classes using Tsar software, of which compounds in Class4 represents the highest class or top rank. Classes were generated for all scoring functions and instead of taking an average, rank-sum technique [8] was employed to retrieve best compounds. The ranks obtained from each of the individual scoring functions were added to give a rank-sum [Table-3]. The advantage of a sum over an average is that the contribution from each individual score can more easily be split out for illustrative purposes in the former instance. Finally, from top rank-sum classes, Riboflavin, Apigenin-5-o-glucoside and Apigetrin compound conformers are considered as potential ligands against Aldose reductase. The docking scores of the above best compounds in the seven different softwares, generated classes using Tsar software and the sum of the classes for each ligand are shown in Table 3 and Table 4.(Appendix)

From our analysis, it is evident that plant based compound Apigetrin exhibited anti-diabetic activity as it obtained best rank among other compounds and the the major interacting residues are reported in Table-5 and the 2-D image of apigetrin in Figure-1.

Table 5: Number of H-bond interactions and the corresponding interacting residues of apigetrin with active site amino acid residues of aldose reductase.

Compound	MolDock Score	No. of Interactions	Interacting residues
Apigetrin	-133.388	4	OG - Ser302
			NE1 - Trp20
			NE2 - Gln49
			O - Tyr48

5. Conclusion

Screening methods are routinely and extensively used to reduce cost and time of drug discovery. It has been clearly demonstrated that the approach utilized in this study is successful in finding novel anti-diabetic inhibitors from plants. Compound Apigetrin, in particular, from *Cuminum cyminum* showed high binding affinity against Aldose reductase, 1AH3. The docked pose of the compound exactly fits into the active site region and the ligand formed more number of H-bond interactions than the co-crystallized ligand. Therefore, this study states the importance of small molecules from various plant sources and their use to enhance protein-ligand interaction studies, in silico.

6. References

- [1] Expert committee on diagnosis and Classification of Diabetes Mellitus, Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, DIABETES CARE, VOLUME 26, SUPPLEMENT 1, JANUARY 2003
- [2] Jalaie M, V Shanmugasundaram 2006 Virtual screening: are we there yet? Mini Rev Med Chem 6:1159-1167
- [3] Warren GL, CW Andrews, AM Capelli, B Clarke, J LaLonde, MH Lambert, M Lindvall, N Nevins, SF Semus, S Senger, G Tedesco, ID Wall, JM Woolven, CE Peishoff, MS Head 2006 A critical assessment of docking programs and scoring functions. J Med Chem 49:5912-5931
- [4] Waszkowycz B 2008 Towards improving compound selection in structure-based virtual screening. Drug Discov Today 13:219-226
- [5] D. K. Gehlhaar, G. Verkhivker, P. A. Rejto, D. B. Fogel, L. J. Fogel, S. T. Freer, Proceedings of the Fourth Annual Conference on Evolutionary Programming. 1995, 615-627
- [6] J. M. Yang, C. C. Chen, Proteins. 2004, 55, 288-304
- [7] Kitchen DB, H Decornez, JR Furr, J Bajorath 2004 Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3:935-949
- [8] Clark RD, A Strizhev, JM Leonard, JF Blake, JB Matthew 2002 Consensus scoring for ligand/protein interactions. J Mol Graph Model 20:281-295

- [9] Wang R, Y Lu, S Wang 2003 Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46:2287-2303;
 [10] Charifson PS, JJ Corkery, MA Murcko, WP

Walters 1999 Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 42:5100-5109

APPENDIX

Table 3: Scores of the top 10 *Cuminum cyminum* compounds obtained from different docking softwares. All values are in kcal/mol

S.No.	Cuminum compounds	Molegro (kcal/mol)	Ehits (kcal/mol)	Vina (kcal/mol)	Gold (kcal/mol)	MEDock (kcal/mol)	Patchdock (kcal/mol)
1	Riboflavin	-133.388	-7.4034	-7.9	29.15	-9.24	4670
2	Apigenin-5-o-glucoside	-131.705	-4.7679	-8.4	58.05	-7.51	4792
3	Apigetrin	-130.833	-6.5608	-9.3	56.89	-11.61	5114
4	Apiin	-127.982	-5.471	-8.7	50.47	-8.14	5748
5	Benzyl Cinnamate	-117.458	-4.7007	-7.8	55.91	-9.21	4476
6	Luteolin	-116.643	-5.5682	-8.3	49.89	-12.6	3864
7	Stigmasterol	-116.379	-2.4912	-9.2	21.18	-6.13	5436
8	Cosmosin	-115.701	-6.1684	-9.1	46.83	-11.76	5090
9	Luteolin-7-o-glucoside	-112.54	-4.8228	-8.7	51.73	-10.77	4946
10	Cynaroside	-111.372	-5.9918	-8.8	52.7	-11.95	4932

Table 4: Classes generated using Tsar software.

S.No.	Compound	Molegro	Ehits	Vina	Gold	MEDock	Patchdock	Sum
1	Riboflavin	4	4	1	1	2	2	14
2	Apigenin-5-o-glucoside	4	2	2	4	1	2	15
3	Apigetrin	4	4	4	4	4	3	23
4	Apiin	4	3	3	4	2	4	20
5	Benzyl Cinnamate	2	2	1	4	2	2	13
6	Luteolin	1	3	2	4	4	1	15
7	Stigmasterol	1	1	4	1	1	4	12
8	Cosmosin	1	3	4	3	4	3	18
9	Luteolin-7-o-glucoside	1	2	3	4	3	3	16
10	Cynaroside	1	3	3	4	4	3	18

ETIOLOGY OF THE DISEASES CAUSED BY BACTERIUM ESCHERICHIA COLI ACCORDING TO AN ELECTROMAGNETIC MODE

JUAN ESTEBAN CORREA LÓPEZ

Physic engineer ,EAFIT University, Medellin, Antioquia, Colombia

Abstract: *There are evidences according to which the colonies of Escherichia coli bacterium form parabolic cylindrical structures. In such circumstances many symptoms are generated which are produced by a parasitic capacitance. This last is generated by the bacteria and it was calculated using a mathematical model using computer algebra. The mathematical model was built using Laplace equation, Whittaker functions, Hermite functions and the corresponding boundary condition. The resulting mathematical model was implemented using a maple algorithm. This algorithm can be extended to other kinds of bacteria whose colonies are characterized by different classes of specific geometries. Our results suggest that the antibodies are not able to find the bacteria because the induced parasitic capacitance alters the electromagnetic signals that the brain sends to the immune system doing that antibodies lost the signals that are indentifying those colonies.*

Keywords: *parabolic cylinder, parasitic capacitance, computer algebra, E-Coli, Laplace equation, Whittaker functions, Hermite functions.*

1 Introduction

The bacterium Escherichia coli or better known as E-coli through the history has been the bacteria most studied around the world because this is the principal cause of the gastrointestinal diseases in humans. The colonies of these bacteria form in some circumstances parabolic cylindrical structures creating a parasitic capacitive [1] effect due to electric potential that each bacterium contain and for hence due to the electric potential that all structure contains, as will be see later. This parasitic capacitance changes the electrostatic potential of the intestine, causing the gastro intestinal symptoms and altering the control of the electromagnetic signals that are regulating the immune system.

In this work will obtain a mathematical expression that defines the etiology of the diseases caused by bacterium Escherichia coli in terms of a parasitic capacitance and with this aim we will derive the electric potential and de electric

field using especial functions such as Hermit function and Whittaker function. All computation will be made using Maple.

2. Problem

After decades of study in biology have resulted evidence of the structure that form the Escherichia Coli colonies. Some colonies have particular forms that create symptoms on the humans and the animals for this reason is justified study them with computational math.

In figure 1 and figure 2 are shown particular forms of Escherichia coli colonies that will be studied in this work with the objet to obtain physical explanations of their effects on humans and animals. The images in figures 1 and 2 were obtained by scanning electron microscopy.

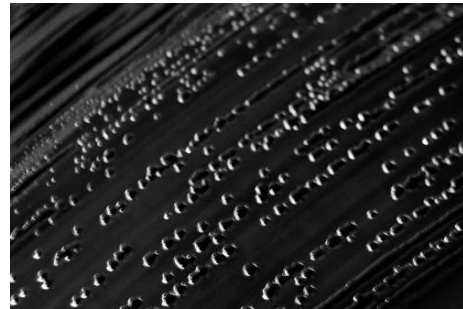


Figure 1, Photography of Escherichia-Coli

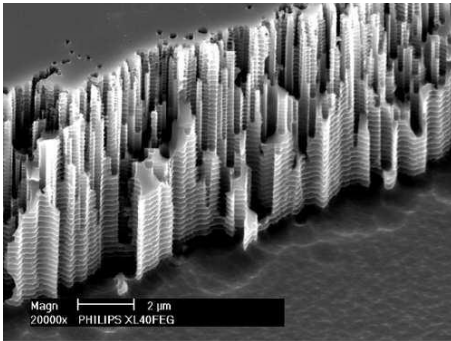


Figure 2, Photography of Escherichia-Coli with SEM (Scanning Electron Microscope).

Idealizing a little bit, we can represent a colony of the bacteria E-coli by a set of parabolic cylinders that create an effect of parasitic capacitance, as is illustrated in figure 3.

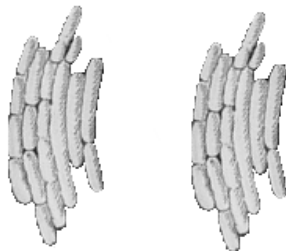


Figure 3, parabolic cylinder formed for E-coli bacteria.

As each bacterium contains a small amount of electric charge then the resulting parabolic cylinder formed by the colony will have an electric potential and this potential will be called $V(\alpha, \beta, z)$ as shown in Figure 4.

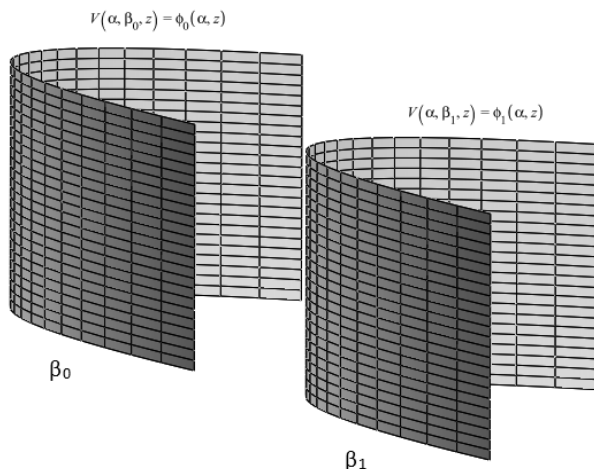


Figure 3, idealized model of two parabolic cylinders made by E-coli bacteria

The equation that will use to determine the electrical potential between the parabolic cylinders is the Laplace equation, which in cartesian coordinates is described as:

$$\frac{\partial^2}{\partial x^2} V(x, y, z) + \frac{\partial^2}{\partial y^2} V(x, y, z) + \frac{\partial^2}{\partial z^2} V(x, y, z) = 0$$

For practical purposes the Laplace equation will be worked in parabolic cylindrical coordinates defined by [2]:

$$x = \frac{1}{2} \alpha^2 - \frac{1}{2} \beta^2, \quad y = \alpha\beta, \quad z = z$$

Where $\alpha \in [0, \infty)$, $\beta \in [0, \infty)$, and $z \in (-\infty, \infty)$.

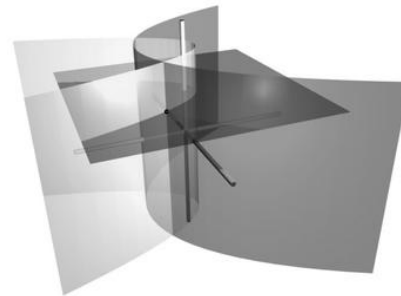


Figure 4. Coordinate surfaces of parabolic cylindrical coordinates.

Rewriting the Laplace equation in parabolic cylindrical coordinates we obtain:

$$\frac{1}{\alpha^2 + \beta^2} \left(\frac{\partial^2}{\partial \alpha^2} V(\alpha, \beta, z) + \frac{\partial^2}{\partial \beta^2} V(\alpha, \beta, z) + (\alpha^2 + \beta^2) \left(\frac{\partial^2}{\partial z^2} V(\alpha, \beta, z) \right) \right) = 0$$

To find the solution to the problem is required to establish boundary conditions. In our case we use Dirichlet conditions which consist in specify the solution $V(\alpha, \beta, z)$ on the border of the application domain of the Laplace equation. In our case the borders that delimit such domain are two parabolic cylindrical surfaces determined as $\beta = \beta_0$ y $\beta = \beta_1$ giving that we only take into account the potential generated between these two surfaces.

$$V(\alpha, \beta_0, z) = \phi_0(\alpha, z)$$

$$V(\alpha, \beta_1, z) = \phi_1(\alpha, z)$$

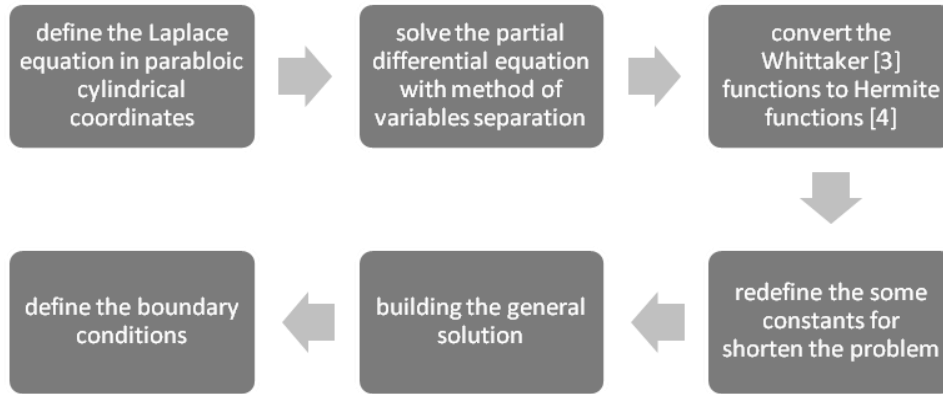
3 Method

To solve the problem we used computer algebra, specifically Maple and with its help packages including "VectorCalculus" and "PDETools".

for reasons of space was not possible to show all the algorithm is illustrated for this reason the procedure took place in the following flow chart but if you want you can download it copy the following URL in your web browser

<http://dl.dropbox.com/u/7791924/ETIOLOGY%20OF%20THE%20DISEASES%20CAUSED%20BY%20BACTERIUM%20ESCHERICHIA%20COLI>

<http://cid-ad28443fd7d93b36.office.live.com/self.aspx/P%3fbablico/ETIOLOGY%20OF%20THE%20DISEASES%20CAUSED%20BY%20BACTERIUM%20ESCHERICHIA%20COLI%20ACCORDING%20TO%20AN%20ELECTROMAGNETIC%20MODE.mw>



4 Results

$$\phi_0(\alpha, z) = \int_{-\infty}^{\infty} \phi_0(\alpha, \lambda) \cos(\lambda z) d\lambda \qquad \phi_1(\alpha, z) = \int_{-\infty}^{\infty} \phi_1(\alpha, \lambda) \cos(\lambda z) d\lambda$$

$$\begin{aligned}
 V(\alpha, \beta, z) = & \int_{-\infty}^{\infty} e^{-\frac{1}{2} \lambda \alpha^2} \cos(\lambda z) \text{HermiteH}(n, \sqrt{\lambda} \alpha) \left(\sqrt{\beta_0} \sqrt{\lambda} \sqrt{\beta_1} \left[-\text{HermiteH}(-n-1, \right. \right. \\
 & \left. \left. \sqrt{\lambda} \beta_0 \right) e^{-\frac{1}{2} \lambda \beta_0^2} \left(\int_{-\infty}^{\infty} \frac{\text{HermiteH}(n, x) e^{-\frac{1}{2} x^2} \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) + \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \right. \\
 & \left. \left. \text{HermiteH}(-n-1, \sqrt{\lambda} \beta_1) e^{-\frac{1}{2} \lambda \beta_1^2} \right) \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta^2\right) \right) / \\
 & \left(\sqrt{\beta} n! 2^n \sqrt{\pi} \left(-\text{HermiteH}(-n-1, \sqrt{\lambda} \beta_0) e^{-\frac{1}{2} \lambda \beta_0^2} \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) \right. \right. \\
 & \left. \left. + \text{HermiteH}(-n-1, \sqrt{\lambda} \beta_1) e^{-\frac{1}{2} \lambda \beta_1^2} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) \right) \right) + \left(\text{HermiteH}(-n-1, \right. \\
 & \left. \sqrt{\lambda} \beta) e^{-\frac{1}{2} \lambda \beta^2} \sqrt{\lambda} \left(- \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + \left(\int_{-\infty}^{\infty} \frac{\text{HermiteH}(n, x) e^{-\frac{1}{2} x^2} \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) \right) \right) / \left(n! 2^n \sqrt{\pi} \left(\right. \right. \\
 & \left. \left. -\text{HermiteH}(-n-1, \sqrt{\lambda} \beta_0) e^{-\frac{1}{2} \lambda \beta_0^2} \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + \text{HermiteH}(-n-1, \sqrt{\lambda} \beta_1) \right. \right. \\
 & \left. \left. e^{-\frac{1}{2} \lambda \beta_1^2} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) \right) \right) \right) d\lambda
 \end{aligned}$$

4.1 Electric field

$$\begin{aligned}
 E(\alpha, \beta, z) = & -\frac{1}{\sqrt{\alpha^2 + \beta^2}} \left[\sum_{n=0}^{\infty} \left(\left(\sqrt{\beta_0} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta^2\right) \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_0) e^{-\frac{1}{2}\lambda\beta_0^2} \right. \right. \right. \\
 & \left. \left. \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) - \sqrt{\beta_0} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta^2\right) \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \right. \\
 & \left. \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_1) e^{-\frac{1}{2}\lambda\beta_1^2} + \text{HermiteH}(-n-1, \sqrt{\lambda}\beta) e^{-\frac{1}{2}\lambda\beta^2} \sqrt{\beta} \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \right. \\
 & \left. \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta_1^2\right) - \text{HermiteH}(-n-1, \sqrt{\lambda}\beta) e^{-\frac{1}{2}\lambda\beta^2} \sqrt{\beta} \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \right. \\
 & \left. \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta_0^2\right) \right) \left(\lambda \alpha \text{HermiteH}(n, \sqrt{\lambda}\alpha) - 2\sqrt{\lambda} n \text{HermiteH}(n-1, \sqrt{\lambda}\alpha) \right) e^{-\frac{1}{2}\lambda\alpha^2} \cos(\lambda z) \sqrt{\lambda} 2^{-n} \Bigg/ \\
 & \left(\sqrt{\beta} \left(\text{HermiteH}(-n-1, \sqrt{\lambda}\beta_0) e^{-\frac{1}{2}\lambda\beta_0^2} \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta_1^2\right) - \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_1) e^{-\frac{1}{2}\lambda\beta_1^2} \sqrt{\beta_1} \right. \right. \\
 & \left. \left. \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta_0^2\right) \sqrt{\pi} n! \right) \right) \Bigg] \frac{1}{\sqrt{\alpha^2 + \beta^2}} \left[\sum_{n=0}^{\infty} \left(e^{-\frac{1}{2}\lambda\alpha^2} \cos(\lambda z) \text{HermiteH}(n, \right. \right. \\
 & \left. \left. \sqrt{\lambda}\alpha \right) \left(\sqrt{\beta_0} \lambda^{3/2} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta^2\right) \beta^2 \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_0) e^{-\frac{1}{2}\lambda\beta_0^2} \right. \right. \\
 & \left. \left. \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) + \sqrt{\beta_0} \sqrt{\lambda} \sqrt{\beta_1} \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_0) e^{-\frac{1}{2}\lambda\beta_0^2} \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \right. \\
 & \left. \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta^2\right) n + \sqrt{\beta_0} \sqrt{\lambda} \sqrt{\beta_1} \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_0) e^{-\frac{1}{2}\lambda\beta_0^2} \right. \\
 & \left. \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \text{WhittakerM}\left(-\frac{1}{2}n + \frac{3}{4}, \frac{1}{4}, \lambda\beta^2\right) - \sqrt{\beta_0} \sqrt{\lambda} \sqrt{\beta_1} \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_0) e^{-\frac{1}{2}\lambda\beta_0^2} \left. \right. \\
 & \left. \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \text{WhittakerM}\left(-\frac{1}{2}n + \frac{3}{4}, \frac{1}{4}, \lambda\beta^2\right) n - \sqrt{\beta_0} \lambda^{3/2} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta^2\right) \beta^2 \left. \right. \\
 & \left. \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_1) e^{-\frac{1}{2}\lambda\beta_1^2} - \sqrt{\beta_0} \sqrt{\lambda} \sqrt{\beta_1} \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \\
 & \left. \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_1) e^{-\frac{1}{2}\lambda\beta_1^2} \text{WhittakerM}\left(-\frac{1}{2}n - \frac{1}{4}, \frac{1}{4}, \lambda\beta^2\right) n - \sqrt{\beta_0} \sqrt{\lambda} \sqrt{\beta_1} \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \right. \\
 & \left. \text{HermiteH}(-n-1, \sqrt{\lambda}\beta_1) e^{-\frac{1}{2}\lambda\beta_1^2} \text{WhittakerM}\left(-\frac{1}{2}n + \frac{3}{4}, \frac{1}{4}, \lambda\beta^2\right) + \sqrt{\beta_0} \sqrt{\lambda} \sqrt{\beta_1} \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \right.
 \end{aligned}$$

$$\begin{aligned}
 & dx \left[\text{HermiteH}(-n-1, \sqrt{\lambda} \beta_1) e^{-\frac{1}{2} \lambda \beta_1^2} \text{WhittakerM}\left(-\frac{1}{2} n + \frac{3}{4}, \frac{1}{4}, \lambda \beta^2\right) n - 2 \lambda \text{HermiteH}(-n-2, \sqrt{\lambda} \beta) e^{-\frac{1}{2} \lambda \beta^2} \beta^{3/2} \right. \\
 & \left. \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + 2 \lambda \text{HermiteH}(-n-2, \sqrt{\lambda} \beta) e^{-\frac{1}{2} \lambda \beta^2} \beta^{3/2} \right. \\
 & \left. \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) - 2 \lambda \text{HermiteH}(-n-2, \sqrt{\lambda} \beta) e^{-\frac{1}{2} \lambda \beta^2} \beta^{3/2} n \right. \\
 & \left. \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + 2 \lambda \text{HermiteH}(-n-2, \sqrt{\lambda} \beta) e^{-\frac{1}{2} \lambda \beta^2} \beta^{3/2} n \right. \\
 & \left. \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) - \text{HermiteH}(-n-1, \sqrt{\lambda} \beta) \lambda^{3/2} \beta^{5/2} e^{-\frac{1}{2} \lambda \beta^2} \right. \\
 & \left. \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + \text{HermiteH}(-n-1, \sqrt{\lambda} \beta) \lambda^{3/2} \beta^{5/2} e^{-\frac{1}{2} \lambda \beta^2} \right. \\
 & \left. \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) 2^{-n} \right) / \left(\beta^{3/2} \left(-\text{HermiteH}(-n-1, \right. \right. \\
 & \left. \left. \sqrt{\lambda} \beta_0) e^{-\frac{1}{2} \lambda \beta_0^2} \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + \text{HermiteH}(-n-1, \sqrt{\lambda} \beta_1) e^{-\frac{1}{2} \lambda \beta_1^2} \sqrt{\beta_1} \text{WhittakerM}\left(\right. \right. \\
 & \left. \left. -\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) \sqrt{\pi} n! \right) \right) d\lambda \Big|_{\mathbb{R}} - \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \left(\left(\left(-\sqrt{\beta_0} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta^2\right) \text{HermiteH}(-n-1, \sqrt{\lambda} \beta_0) \right. \right. \right. \\
 & \left. \left. e^{-\frac{1}{2} \lambda \beta_0^2} \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) + \sqrt{\beta_0} \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta^2\right) \left(\right. \right. \right. \\
 & \left. \left. \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \text{HermiteH}(-n-1, \sqrt{\lambda} \beta_1) e^{-\frac{1}{2} \lambda \beta_1^2} - \text{HermiteH}(-n-1, \sqrt{\lambda} \beta) e^{-\frac{1}{2} \lambda \beta^2} \sqrt{\beta} \right. \\
 & \left. \left(\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_0(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + \text{HermiteH}(-n-1, \sqrt{\lambda} \beta) e^{-\frac{1}{2} \lambda \beta^2} \sqrt{\beta} \left(\right. \right. \\
 & \left. \left. \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2} x^2} \text{HermiteH}(n, x) \phi_1(x, \lambda)}{\sqrt{\lambda}} dx \right) \sqrt{\beta_1} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) e^{-\frac{1}{2} \lambda \alpha^2} \sin(\lambda z) \lambda^{3/2} \text{HermiteH}(n, \sqrt{\lambda} \alpha) 2^{-n} \right) / \\
 & \left(\sqrt{\beta} \left(-\text{HermiteH}(-n-1, \sqrt{\lambda} \beta_0) e^{-\frac{1}{2} \lambda \beta_0^2} \sqrt{\beta_0} \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_1^2\right) + \text{HermiteH}(-n-1, \sqrt{\lambda} \beta_1) e^{-\frac{1}{2} \lambda \beta_1^2} \sqrt{\beta_1} \right. \right. \\
 & \left. \left. \text{WhittakerM}\left(-\frac{1}{2} n - \frac{1}{4}, \frac{1}{4}, \lambda \beta_0^2\right) \sqrt{\pi} n! \right) \right) d\lambda \Big|_{\mathbb{R}}
 \end{aligned}$$

Now to find the capacitance we know that:

$$\begin{aligned} \sigma \Big|_{\beta = \beta_0} &= \epsilon_0 \left(E_{\beta} \Big|_{\beta = \beta_0} \right) \\ dQ \Big|_{\beta = \beta_0} &= \sigma (\alpha^2 + \beta^2) d\alpha dz \\ Q \Big|_{\beta = \beta_0} &= \int_0^{\infty} \int_{-\infty}^{\infty} \epsilon_0 \left(E_{\beta} \Big|_{\beta = \beta_0} \right) (\alpha^2 + \beta_0^2) d\alpha dz \\ C &= \frac{Q}{V_1 - V_0} \end{aligned}$$

5 Conclusions

In this work was made a study about the Escherichia Coli bacterium for giving a explanation the etymology of this bacterium in electromagnetic terms could be interpreted as follows:

E-Coli bacteria when they enter a human or animal body stays in the intestinal wall forming parabolic cylindrical structures as evidenced in the photographs above. By the fact that those bacteria are alive. These bacteria have a certain amount of electric charge, hence the electric field and electric potential. As in electrical circuits that only the proximity of the components produce a "parasitic capacitance" the Escherichia coli produces the same capacitance produces interference in the order of the electromagnetic signals of the host body such as immune system and gastrointestinal system making antibodies can not easily locate the position of the bacteria in the body and altering the electromagnetic signals of the amino acids that control the gastrointestinal biochemistry.

Software has given an acceleration of the developments that are at the present around the world. In this case Maple allowed developing an algorithm that gives an explanation for the etiology of the E-coli bacterium using electromagnetic concepts which alters the physiology of the beings who suffer from this infection.

Particularly the study was done on E-coli colonies with parabolic cylindrical form but this algorithm can be extended to different bacteria and forms that make their colonies only defining a specific geometry and a coordinate system that helps simplify the problem as much as possible.

6 References

- [1] http://en.wikipedia.org/wiki/Parabolic_cylindrical_coordinates
- [2] http://en.wikipedia.org/wiki/Parasitic_capacitance
- [3] <http://mathworld.wolfram.com/WhittakerFunction.html>
- [4] <http://mathworld.wolfram.com/HermitePolynomial.html>

SESSION

MODELLING, SIMULATION AND OPTIMIZATION OF BIOLOGICAL SYSTEMS

Chair(s)

TBA

Multi-Scale modelling of the Bile Acid and Xenobiotic System

Noel Kennedy¹, Paul Thompson¹, Huiru Zheng², Werner Dubitzky¹

The 2011 International Conference on Bioinformatics & Computational Biology

¹ School of Biomedical Sciences, University of Ulster, Coleraine, Northern Ireland, UK

² School of Computing & Mathematics, University of Ulster, Belfast, Northern Ireland, UK

Abstract: Systems biology has developed considerably in the past decade combining the different disciplines of mathematical modelling, computational simulation and biological experimentation facilitating the quantitative analysis of biological systems. This is often severely hampered by the lack of time-resolved data which ultimately leads to problems in validating any models created. To address the inherent complexity in biological systems, a recent trend in systems biology is exploring multi-scale modelling and simulation methodologies. We consider the Bile Acid and Xenobiotic System (BAXS) as a typical example of a multi-scale system. In the absence of dynamic data from biological experimentation the models we have developed are based on artificial data which enables us to explore multi-scale modelling and validation techniques and the integration of individual models. The outcome of this study will direct further research into multi-scale modelling methodology and ultimately will produce a novel framework for validation in the absence of dynamic data.

Keywords: Systems biology, multi-scale modelling, simulation, xenobiotics, bile acids.

1. Introduction

The main focus for this research is addressing the inherent complexity in biological systems by exploring multi-scale modelling and simulation methodologies. To facilitate this investigation we model the *bile acid and xenobiotic system (BAXS)*, a typical example of a multi-scale biological system adopting a multi-scale modelling and simulation approach. The BAXS describes a genetic network that facilitates two distinct but intimately overlapping physiological processes; The enterohepatic circulation and maintenance of bile acid concentrations (Figure 1) and the detoxification and removal from the body of harmful xenobiotic (e.g. drugs, pesticides), and endobiotic compounds (e.g. steroid hormones)^[1]. The system involves the coordination of several levels of gene activity, including control of mRNA and protein expression and regulation of metabolising enzyme and transporter protein function in tissues such as liver, intestine/colon and kidney. Bile acids are necessary for the emulsification and absorption of dietary fats and are therefore valuable compounds, however as their build-up can cause harm, their concentrations need to be appropriately regulated and recycled. Similarly there is a requirement for a system that can 'sense' the accumulation of xenobiotic and endobiotic

compounds and facilitate their detoxification and removal from the body. The BAXS accomplishes this and maintains enterohepatic circulation (the circulation of biliary acids from the liver, depicted Figure 1) through a complex network of sensors in the form of *nuclear receptors* that function as ligand-activated transcription factors.

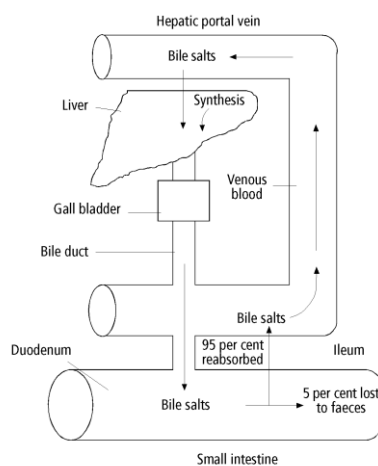


Figure 1. Schematic illustration of enterohepatic circulation.

They serve to detect fluctuations in concentration of many compounds and initiate a physiological response by regulating the BAXS. Transcriptional regulation by nuclear receptors involves both activating and repressive effects upon specific 'sets' of genes. There is considerable overlap exhibited between nuclear receptors in the genes they target and also the ligands that bind to and activate them. It is these factors that contribute to the phenomenon of drug-drug interactions, e.g. between St. John's Wort and Cyclosporine^[2] or St. John's Wort and Oral contraceptive^[3]. Positive feed-forward and negative feed-back loops can also occur, e.g. within the cholesterol metabolic pathway^[4, 5]. Multi-scale modelling of the BAXS will benefit biologists interested in exploring such phenomena. Multi-scale systems biology modelling efforts aim to explore such multi-scale systems quantitatively by means of simulations that integrate several (usually independently developed) single-scale models into a coherent multi-scale model^[6]. Our aim is to capture and model separate BAXS processes individually and combine them using a multi-scale modelling approach. For example, in the BAXS the initial stimuli leading to a physiological response would be the binding of a ligand by a

nuclear receptor. The process following the ligand-receptor binding event involves the bound nuclear receptor binding to response elements in the target genes and the cascading effects of increased gene expression that would ensue. Subsequent processes include conjugation and transporter functions^[7]. Each single process can be modelled separately regardless of the different scales the may operate in. They can be referred to as separate modalities of biology thus the approach taken is 'multi-modal'. The 'modularity' or multi-biology approach better reflects the way biologists would do experiments, investigating one constituent process at a time, each yielding a separate data set. Single-scale / single-biology models can be built from these experiments and then these individual models can be integrated into a multi-scale/multi-biology model. Each single scale model can then be reverse engineered separately and then integrated with a suitable coupling approach. Alternatively all single scale models can be reverse engineered in a single reverse engineering process however this approach must include the coupling within the reverse engineering phase. Through such experimentation the aim is to address the problems associated with multi-scale modelling and validation, specifically the coupling of processes operating on different scales.

Developing dynamic models of biological process and systems requires dynamic (time-resolved) quantitative data. Such time-series data provides measurements being recorded at certain, pre-defined intervals over a period of time. For many biological systems or processes of interest, sufficient dynamic data required for modelling may not be available^[8, 9, 10]. For example, many experimental protocols in biology require the killing of their specimen. This approach precludes the collection of individual-based time series data. Systems biology is still a developing field and current biological experimentation is rapidly changing to produce quantitative data facilitating the development (including validation) of dynamic models. Currently however, for many biological systems of interest, there is insufficient data to develop and validate dynamic models.

2. BAXS processes

Nuclear receptors are a class of proteins found within the interior of cells that are responsible for sensing the presence of steroid and thyroid hormones and certain other molecules. In response, these receptors work in concert with other proteins to regulate the expression of specific genes, thereby controlling the development, homeostasis, and metabolism of the organism. Nuclear receptors have the ability to directly bind to DNA and regulate the expression of adjacent genes. Hence, these receptors are classified as *transcription factors*¹. The regulation of gene expression by nuclear receptors occurs only when a *ligand* — a molecule that affects the receptor's behavior (i.e., activate or deactivate it) — is present. More specifically, ligand binding to a nuclear

receptor results in a conformational change of the receptor molecule complex, which in turn activates the receptor resulting in up-regulation of gene expression. A unique property of nuclear receptors that differentiates them from other types of receptors is their ability to directly interact with and control the expression of genomic DNA. As a consequence, nuclear receptors play a key role in both embryonic development and adult homeostasis.

Our BAXS modelling efforts are directed first at the effects of ritonavir on the metabolism of hyperforin in the liver and the overlap of this process with FXR mediated primary and secondary bile acid metabolism. We refer to this as the *Liver scenario* which is depicted in the diagram of Figure 2. Its main constituent elements and processes are described below.

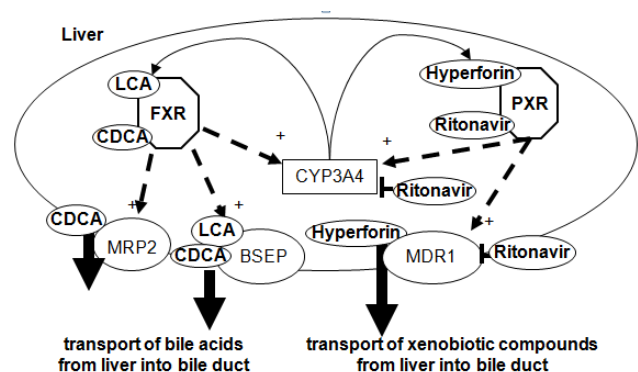


Figure 2. Metabolism of hyperforin and bile acid in liver. PXR-mediated metabolism of hyperforin in the liver inhibited by ritonavir, FXR mediated bile acid metabolism and the transport process.

Pregnane X receptor (PXR) is a nuclear receptor highly expressed in the liver encoded by the *NR1I2* (nuclear receptor subfamily 1, group I, member 2) gene. Its primary function is to sense the presence of foreign toxic substances and in response up-regulate the expression of proteins involved in the detoxification and clearance of these substances from the body^[11].

Farnesoid X receptor (FXR), a nuclear receptor encoded by the *NR1H4* (nuclear receptor subfamily 1, group H, member 4) gene is also known as the bile acid receptor. It is highly expressed in the liver and its primary function is to sense the presence of bile acids and protect the body from elevated bile acid concentrations^[12].

Hyperforin is a herbal antidepressant found in St. John's wort and is an activating ligand for PXR^[13]. Activated PXR up-regulates transcription of *CYP3A4* (measured in hours) producing enzymes which metabolise Hyperforin (measured in seconds to minutes)^[14]. PXR also targets the gene encoding MDR1^[15], a transporter protein which transports hyperforin from the cell (measured in seconds to minutes).

Ritonavir is a protease inhibitor, often prescribed to HIV patients as part of antiretroviral therapy^[16]. HIV protease is an enzyme which cuts the raw material for HIV into specific pieces needed to build a new virus. Protease inhibitors block the protease enzyme preventing it from working, thus

¹ Transcription factors activate or repress the transcription of a gene by controlling the time and rate of transcription of a gene's DNA into RNA.

incomplete, defective copies of HIV are formed which cannot infect cells. Ritonavir is also an activating ligand for PXR [17], however without receptor binding it can repress metabolism and transporter activity induced from transcription of CYP3A4 and MDR1 through competitive inhibition (measured in seconds and minutes). This could lead to a possible build-up of hyperforin in the liver.

The bile acid receptor (BAR), also known as farnesoid X receptor (FXR) is activated by primary and secondary bile acids, lithocholic acid (LCA) and chenodeoxycholic acid (CDCA). It up-regulates transcription of CYP3A4, MRP2 and BSEP, the latter two encoding transporter proteins which transport bile acids into the bile duct. The overlap of both processes occurs at the CYP3A4 gene and several scenarios can be explored. A patient taking hyperforin will have increased expression of CYP3A4 which may lead to a deficiency in bile acid concentration as this gene produces enzymes which metabolise bile acids. Similarly a patient with high bile acid concentrations may reduce the efficacy of hyperforin (if taken) as transcription of CYP3A4 is increased. If ritonavir is added to this example then bile acids and hyperforin could accumulate to toxic levels in the liver. A second scenario which will be considered in future work looks at the effects of ritonavir on the metabolism of hyperforin in the intestine and the overlap of this process with VDR-mediated vitamin D metabolism.

3. Multi-scale modelling

Starting with early studies beginning in 1990s [18] multi-scale modelling and simulation has now turned into a focal point of attention across many scientific and engineering disciplines. An increasing number of scientific papers are published, workshops are organized and some specialized journals exist. Communities (ranging from physics and biology to medicine, finance, and engineering) are confronted with the problem of understanding multi-scale systems that are central to their field of study. For instance, the Virtual Physiological Human project [19], funded by the EC, is a good example of a community concerned with multi-scale modelling and simulation of human physiology. The COAST project developed a multi-scale modelling methodology [20] whose basic building blocks comprise single-scale models and their mutual multi-scale couplings. Many, if not all, multi-scale models can be expressed in this general multi-scale modelling framework. In the COAST framework, a multi-scale model can be represented as a directed graph on a *scale separation map* (SSM), which is a plot that has the relevant range of scales on its axes (usually space and time, but other quantities are possible). The single-scale models are positioned on the SSM according to their characteristic scales, and the coupling templates are represented as directed edges (Figure 3). While many approaches to systems biology involve single-scale models, there is a growing body of work that aims at modelling of life phenomena across several *scales*. Multi-scale systems biology is concerned with experiments and hypotheses that involve different scales of biological organization from intracellular molecular interactions to cellular behaviour and the behaviour of cell populations (Figure 3). Multi-scale

systems biology modelling efforts aim to explore such multi-scale systems quantitatively by means of simulations that incorporate several different simulation techniques because of the different *temporal scales* and *spatial scales* involved [6, 21, 22].

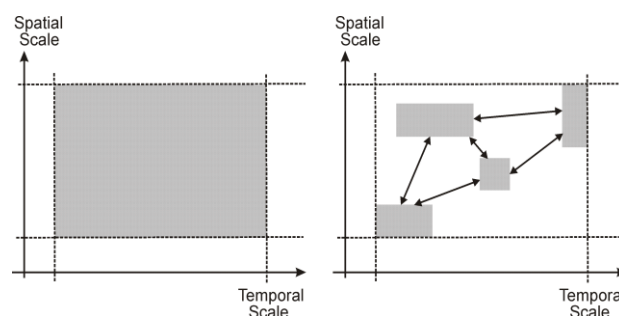
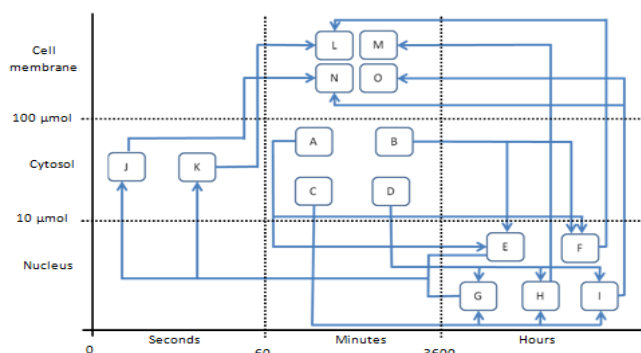


Figure 3. The scale separation map.
Decomposition of a multi-scale system: Left, a multi-scale model spanning many temporal and spatial scales. Right, the resulting decomposed model, consisting of four coupled single scale models.

Qualitative diagrammatic multi-scale models are very common in biomedical research. Ultimately all biological properties on the level of tissues or organs are based on molecular interactions occurring within or on the surface of cells. Biologists frequently describe the hypothetical role a specific molecular mechanism may play in a tissue-level disease by means of a diagram with an arrow connecting molecular entities to a higher scale entities associated with the disease. However, if one wants to subject the proposed causal relationships to a stringent quantitative exploration one needs to transform the knowledge embodied in the arrow-based diagram into a formal description suitable as input for computer simulations. The SSM depicted in Figure 4 represents the Liver BAXS scenario as described above.



Legend:

Ligand/Receptor binding:

- A PXR binds hyperforin
- B PXR binds ritonavir
- C FXR binds LCA
- D FXR binds CDCA
- Receptor activates gene:**
- E PXR activates CYP3A4
- F PXR activates MDR1
- G FXR activates CYP3A4
- H FXR activates MRP2
- I FXR activates BSEP

Enzyme activity on substrate (inhibited by ritonavir)

- J CYP3A4 metabolises LCA
- K CYP3A4 metabolises Hyperforin
- Transport of substrate from cell (inhibited by ritonavir)**
- L MDR1 transports metabolised hyperforin to exosol
- M MRP2 transports CDCA to exosol
- N BSEP transports metabolised LCA to exosol
- O BSEP transports metabolised CDCA to exosol

Figure 4. SSM representing the Liver scenario.

Each individual process in this scenario has been identified in terms of the spatial and temporal scales within which they occur. The first group of processes (labelled A to D in the diagram) operate within the cytosol and involve the binding of ligand to nuclear receptor which can be measured on a time scale of minutes. The next group of processes (labelled E to I) take place in the nucleus and result in an increased rate of gene expression. These processes operate on the scale of hours. Processes J and K take place in the cytosol, involve the metabolism of the ligand through increased enzyme activity and include the inhibitory effects of another substrate on the metabolic process through competitive inhibition. These processes are measured on a scale of micro-seconds to seconds.

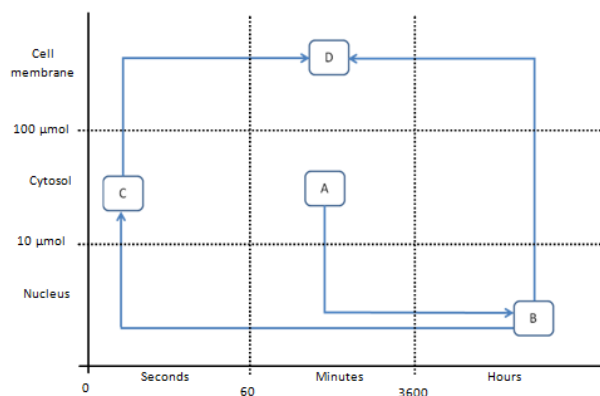
The processes labelled L to O are localized in and at the cell membrane and involve the transport of metabolized substrates across the membrane out of the cell by transporter proteins. These processes also include competitive inhibition of another substrate. These processes occur over a time scale of minutes. To simplify the modelling approach, the processes are grouped together such that process A represents the binding of ligand and nuclear receptor, process B represents gene expression, process C represents enzyme activity on a substrate, including competitive inhibition, and process D represents activity of transporter proteins as shown in Figure 6. Additionally, the initial models created represent the pathway resulting from PXR activation only. This will be further developed to include the FXR pathway once the modelling techniques have been established.

The ligand receptor binding process is governed by mass action kinetic laws^[23] which determine the rate at which the overall reaction occurs. The reaction equations below describe how this process occurs and how the kinetic laws are applied.



Eq. 1 shows that ligand (L) plus nuclear receptor (R) bind to create the ligand/nuclear receptor complex (LR). The rate at which this occurs is determined by the kinetic constant k_{on} which is the association rate for the ligand binding to the nuclear receptor. This reaction is reversible therefore Eq. 2 shows the dissociation of the ligand/receptor complex into its constituent compounds and the rate is determined by the kinetic constant k_{off} which is the dissociation rate of the bound nuclear receptor complex. The combination of both reactions determines the overall rate of complex formation.

The transactivation process resulting in increased gene expression is triggered by the activated PXR complex resulting from process A (either bound to hyperforin or ritonavir) translocating to the cell nucleus and binding to DNA. Among the target genes are *CYP3A4* which produces the enzyme cytochrome p450, and *MDR1* which produces the transporter protein p-glycoprotein, an ATP binding cassette transporter (ABC-transporter).



Legend:

- | | | | |
|----------|-------------------------------------|----------|-----------------------------------|
| A | Ligand/Receptor binding. | C | Enzyme activity on substrate. |
| B | Receptor activates gene expression. | D | Transport of substrate from cell. |

Figure 5. Simplified SSM representing the Liver scenario. Grouping all similar process types together for modelling purposes.

The transcription process follows kinetic laws determined by the Hill function for transcriptional activation^[24, 25]. Eq. 3 shows the equation determining the overall rate of mRNA production

$$\frac{k_1 [A]^n}{k_m^n + [A]^n} \quad \text{Eq. 3}$$

where A denotes the activator (the concentration of the PXR compound), k_1 the maximal transcription rate of the gene, k_m the activation co-efficient and n the Hill coefficient.

As mRNA is produced it translocates to the cytosol and is translated into protein at the ribosome. Eq. 3 shows the equation determining the overall rate mRNA is translated into protein. The rate of this reaction follows the kinetic laws of mass action

$$k_2 [\text{mRNA}] \quad \text{Eq. 4}$$

where k_2 is the translation rate which represents the number of protein molecules produced per mRNA molecule per unit of time.

The ligand receptor binding model was implemented in COPASI^[26], a software tool for simulation and analysis of biochemical networks and their dynamics. The final model forms a mathematical representation of the biological process under study upon which dynamic simulations can be run. Table 1 details the initial concentrations used in the model for ligand-receptor binding. Table 2 shows the reactions between species in the model and the parameter values used. Due to the absence of data from biological experimentation the values used in the model were estimated through a process of trial and error.

Table 1. Initial concentrations for ligand receptor binding model.

Species	Initial concentration (μmol/l)
Hyperforin	600
Ritonavir	500
PXR	10

Table 2. Ligand receptor binding model reactions.

Reaction	Equation	Rate
Ass. of Hyp with PXR	$PXR + Hyp \rightarrow PXR:Hyp$	$8e-06 \text{ l}/(\mu\text{mol}\cdot\text{s})$
Diss. of Hyp and PXR	$PXR:Hyp \rightarrow PXR + Hyp$	$6.5e-07 \text{ 1/s}$
Ass. of Rit with PXR	$PXR + Rit \rightarrow PXR:Rit$	$9e-06 \text{ l}/(\mu\text{mol}\cdot\text{s})$
Diss. of Rit and PXR	$PXR:Rit \rightarrow PXR + Rit$	$7.5e-07 \text{ 1/s}$

Ass. = association; Diss. = dissociation;
Hyp = hyperforin; Rit = ritonavir

A simulation was run in COPASI with the duration set to 600 seconds (10 minutes) and interval size at 2 seconds resulting in a dataset with 300 time-steps. Figure 6 shows the resulting graph of plotting the simulated data.

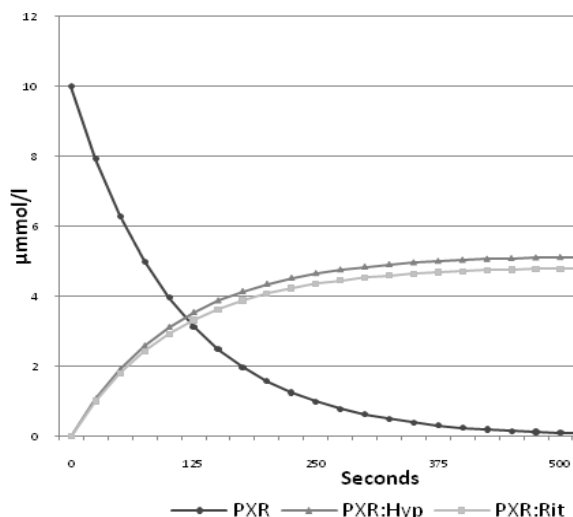


Figure 6. Ligand receptor binding model:

Species concentration (vertical axis) over time (horizontal axis).

A second model was created in COPASI to simulate the reactions involved in process B, which result in activation of gene expression.

Table 3 shows the initial concentrations used in the gene expression model and Table 4 details the reactions rates and parameter values used. The duration for the simulation was set to 100 000 seconds (27.7 hours) with 2500 time steps of 40 seconds each. Figure 7 shows the result of plotting the simulated data. Again, the initial values, rates and parameters have been estimated through a process of trial and error due to the lack of experimental data.

Table 3. Initial concentrations for gene expression model.

Species	Initial concentration (μmol/l)
PXR:Hyp	5.14
PXR:Rit	4.82

Table 4. Gene expression model reactions.

Reaction	Equation	Rates / Parameters
Diss. of PXR:Hyp complex	$PXR:Hyp \rightarrow PXR + Hyp$	0.00085 1/s
Diss. of PXR:Rit complex	$PXR:Rit \rightarrow PXR + Rit$	0.00095 1/s
Transc. of CYP3A4 by PXR:Hyp	$\rightarrow CYP3A4(m); PXR:Hyp$	$k1 = 0.003,$ $n = 1, km = 0.5$
Transc. of CYP3A4 by PXR:Rit	$\rightarrow CYP3A4(m); PXR:Rit$	$k1 = 0.006,$ $n = 1, km = 0.5$
Transc. of MDR1 by PXR:Hyp	$\rightarrow MDR1(m); PXR:Hyp$	$k1 = 0.005,$ $n = 1, km = 0.5$
Transc. of MDR1 by PXR:Rit	$\rightarrow MDR1(m); PXR:Rit$	$k1 = 0.007,$ $n = 1, km = 0.5$
Transl. of CYP3A4 mRNA	$CYP3A4(m) \rightarrow CYP3A4$	$k2 = 2.4e-05 \text{ 1/s},$ $d2 = 1e-05 \text{ 1/s}$
Transl. of MDR1 mRNA	$MDR1(m) \rightarrow MDR1$	$k2 = 2.7e-05 \text{ 1/s},$ $d2 = 1e-05 \text{ 1/s}$

Diss. = dissociation; Transc. = transcription;
Transl. = translation; Hyp = hyperforin; Rit = ritonavir

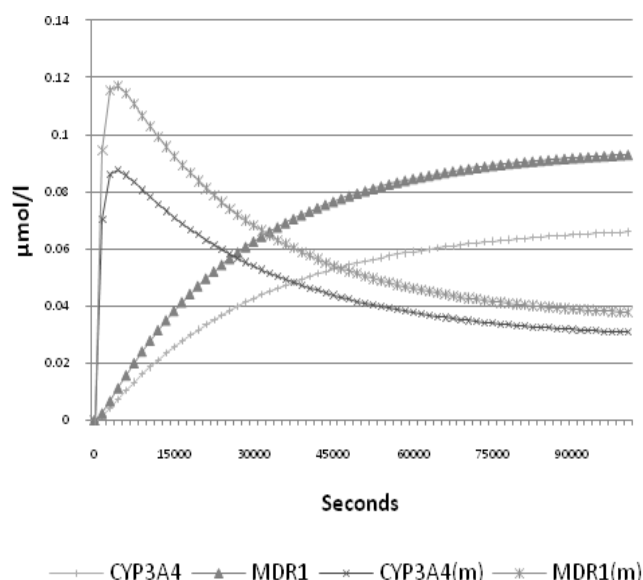


Figure 7. Gene expression model:

Species concentration (vertical axis) over time (horizontal axis).

4. Results

The ligand-receptor binding model indicates a steady increase in bound PXR correlated to a steady decrease in available (unbound) PXR. The initial concentrations of ritonavir and hyperforin decrease steadily (not shown) relative to the accumulation of bound PXR. The entire process is modelled over 600 seconds and reaches a steady state after approximately 500 seconds where the rate of formation of bound PXR begins to level out. The data indicates that after 600 seconds the concentration of PXR bound to ritonavir is 4.82 μmol/l and the concentration of PXR bound to hyperforin is 5.14 μmol/l. To initiate the transcription process only a minimum concentration of activated PXR is required. Process B can therefore start

before process A has finished therefore the processes are not necessarily sequential in nature.

An exchange of data from process A to B is required during the simulation of process A at predefined intervals. The gene expression model indicates a sharp increase in mRNA production peaking at approximately 5000 seconds (approximately 1.5 hours) after which there is a gradual decline.

The translation of mRNA into protein is indicated as a gradual increase in MDR1 and CYP3A4 concentrations which approach steady state at approximately 100 000 seconds (27.7 hours).

The process of enzyme activity on a substrate (process C) is yet to be modelled, however it is dependent on the concentration of the enzymes produced in the gene expression process (process B). As with the integration of processes A and B the relationship between processes B and C is not necessarily sequential. A minimum concentration of enzyme is required to initiate the metabolic process, the rate of which increases as enzyme concentration increases. Each model has been determined as the trigger for the subsequent process, however the processes are not sequential, therefore the integration or 'coupling' of models needs to be studied in more detail. This forms one of the major research areas for this project.

5. Model integration

To investigate how separate individual processes operating on different scales interact with each other a stock and flow diagram was created in Stella² for the processes under study (Figure 8). The stock and flow diagram treats the components of the model as stocks, e.g. 'Le' is a stock of ligand outside the cell. The flows represent the rate of change of the stock, either localization or change of state, e.g. 'Le' flows into the cell at a defined rate and accumulates as 'L' which represents the stock of ligand in the cell. The flow from the ligand stock (L) combines with the flow from receptor stock (R) to accumulate as bound ligand receptor stock (LR). This stock has a positive effect on the flows resulting in enzyme production (E1 and E2) represented by the arcs connecting the stock to the flows. Enzyme 1 stock (E1) has a positive effect on the flow of ligand (L) to its metabolised form (L\OH) and enzyme 2 (E2) has a positive effect on the flow of metabolised ligand (L\OH) out of the cell. Finally the stock of inhibitor (I) has a negative effect on the flow of ligand to metabolised ligand and the flow of metabolised ligand out of the cell. By studying the model in terms of stocks and flows it is easy to visualise the interactions in the model as an exchange of stocks. In terms of *coupling* multi-scale models the exchange of data must therefore represent a concentration of a component or components in the individual processes. For example the integration of processes A and B, ligand binding and gene expression, is an exchange of data representing the

concentration of activated PXR, the interaction of processes B and C, gene expression and enzyme activity, is an exchange of data representing enzyme concentrations. As the processes are not necessarily sequential, exchange of data has to occur at predefined time steps within the model operating on the smaller scale.

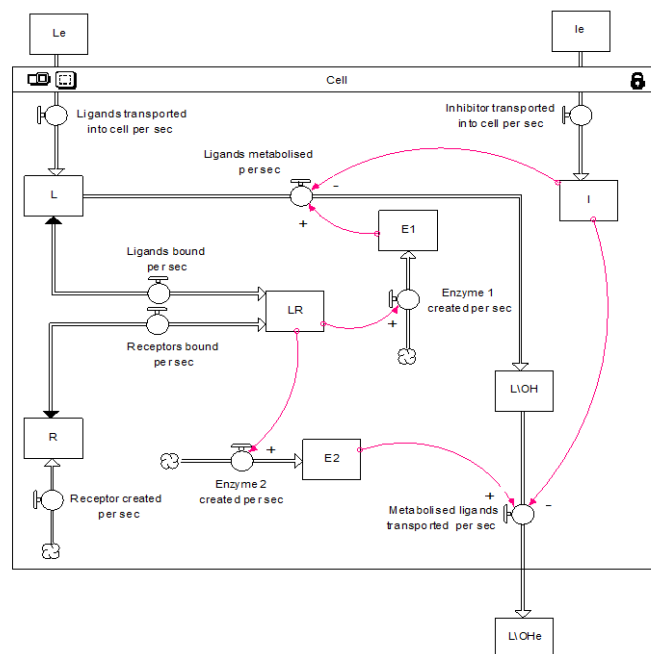


Figure 8. Stock and flow diagram.

Representing the metabolism of hyperforin inhibited by ritonavir in the liver.

E.g. the duration of the ligand binding model is 600 seconds with interval sizes of 2 seconds whereas the duration of the gene expression model is 100 000 seconds with time intervals of 40 seconds. This would mean that for every 20 time steps of process A an exchange of data can occur with process B. This forms the basis for developing methodology for coupling multi-scale processes and allows us to explore this problem further. The development of a 'data generator' using Java has begun which will be able to open two separate instances of Copasi and run two simulations together. It will also be able to interrupt the simulations at specific time intervals and facilitate the exchange of data in either direction as required.

6. Discussion

This study leads us to suggest the most suitable approach in multi-scale modelling and simulation is to deconstruct the entire system into individual processes and model each separately. The *coupling* of models can then be explored in more detail. We suggest the integration or *coupling* of separate models involves an exchange of data representing a stock or concentration of a component within the individual models. The development of a data generator in Java allows this integration of models to be further explored and developed to include other modelling methodologies. This research project has also raised several issues which require further investigation and prompt further research in the fields

² STELLA is a general-purpose modelling and simulation tool of isee systems: www.iseesystems.com.

of biology and systems biology. The models created in Copasi use artificial data to quantify the kinetic rates of reactions within the processes under study. This project would benefit greatly if biological experimentation in this area could provide real data upon which to validate the models. Further models will be developed to capture the additional processes detailed in the SSM and the 'data generator' will be implemented to explore the coupling of these separate processes. Ultimately the 'data generator' will be developed to explore the integration of different spatial scales in biology, including the integration of models using different methodologies e.g. cellular automata, agent based modelling. Multi-scale modelling and simulation is more complex than single-scale modelling and simulation. On the biology side it involves different temporal and spatial scales as well as different types of biological process and entities. On the mathematical side, different methods may be used to model the different sub-models of a multi-scale model. Furthermore, specific methods may be used to couple the different sub-models. On the computational side many intricate issues arise.

A new EC-funded project with University of Ulster participation aims to develop computational strategies, software and services for distributed multi-scale simulations across disciplines, exploiting existing and evolving European e-infrastructure^[27]. Our preliminary literature research on evaluation and validation of multi-scale modelling and simulation in biology shows that there is a lack of suitable detailed work in this area. This and the lack of suitable dynamic data for modelling of the BAXS has prompted us to pursue the development of a testing environment which would allow us (1) To generate unlimited dynamic data related to the BAXS, (2) develop and study multi-scale modelling and simulation approaches for the BAXS, and (3) study, apply and develop validation techniques for multi-scale modelling and simulation in systems biology. The basic idea of this testing environment is based on the Turing-like test for biology^[28].

7. References

- [1] Kliewer, S.A. & Willson, T.M. 2002, "Regulation of xenobiotic and bile acid metabolism by the nuclear pregnane X receptor", *Journal of Lipid Research*, vol. 43, no. 3, pp. 359-364.
- [2] Barone, G., Gurley, B., Ketel, B., Lightfoot, M. & Abul-Ezz, S. 2000, "Drug interaction between St. John's wort and cyclosporine", *The Annals of Pharmacotherapy*, vol. 34, no. 9, pp. 1013-1016.
- [3] Hall, S.D., Wang, Z., Huang, S., Hamman, M.A., Vasavada, N., Adigun, A.Q., Hilligoss, J.K., Miller, M. & Gorski, J.C. 2003, "The interaction between St John's wort and an oral contraceptive[ast]", *Clinical Pharmacology and Therapeutics*, vol. 74, no. 6, pp. 525-535.
- [4] Eloranta, J.J. & Kullak-Ublick, G.A. 2005, "Coordinate transcriptional regulation of bile acid homeostasis and drug metabolism", *Archives of Biochemistry and Biophysics*, vol. 433, no. 2, pp. 397-412.
- [5] Goodwin, B., Redinbo, M.R. & Kliewer, S.A. 2002, "Regulation of CYP3A gene transcription by the pregnane X receptor", *Annual Review of Pharmacology and Toxicology*, vol. 42, no. 1, pp. 1-23.
- [6] Dada, J.O. & Mendes, P. 2011, "Multi-scale modelling and simulation in systems biology", *Integr.Biol.*, vol. 3, no. 2, pp. 86-96.
- [7] Stieger, B. & Meier, P.J. 1998, "Bile acid and xenobiotic transporters in liver", *Current Opinion in Cell Biology*, vol. 10, no. 4, pp. 462-467.
- [8] Mendoza, L. & Xenarios, I. 2006, "A method for the generation of standardized qualitative dynamical systems of regulatory networks", *Theor Biol Med Model*, vol. 3, pp. 13.
- [9] Kitano, H. 2002, "Computational systems biology", *Nature*, vol. 420, no. 6912, pp. 206-10.
- [10] Endy, D. & Brent, R. 2001, "Modelling cellular behaviour", *Nature*, vol. 409, no. 6818, pp. 391-395.
- [11] Kliewer, S.A. 2003, "The Nuclear Pregnane X Receptor Regulates Xenobiotic Detoxification", *The Journal of Nutrition*, vol. 133, no. 7, pp. 2444S-2447S.
- [12] Chawla, A., Repa, J.J., Evans, R.M. & Mangelsdorf, D.J. 2001, "Nuclear Receptors and Lipid Physiology: Opening the X-Files", *Science*, vol. 294, no. 5548, pp. 1866-1870.
- [13] Chang, T. 2009, "Activation of Pregnane X Receptor (PXR) and Constitutive Androstane Receptor (CAR) by Herbal Medicines", *The AAPS Journal*, vol. 11, no. 3, pp. 590-601.
- [14] Lemaire, G., Mnif, W., Pascucci, J., Pillon, A., Rabenoelina, F., Fenet, H., Gomez, E., Casellas, C., Nicolas, J., Cavailles, V., Duchesne, M. & Balaguer, P. 2006, "Identification of New Human Pregnane X Receptor Ligands among Pesticides Using a Stable Reporter Cell System", *Toxicological Sciences*, vol. 91, no. 2, pp. 501-509.
- [15] Lin, Y.S., Yasuda, K., Assem, M., Cline, C., Barber, J., Li, C., Kholodovych, V., Ai, N., Chen, J.D., Welsh, W.J., Ekins, S. & Schuetz, E.G. 2009, "The Major Human Pregnane X Receptor (PXR) Splice Variant, PXR.2, Exhibits Significantly Diminished Ligand-Activated Transcriptional Regulation", *Drug Metabolism and Disposition*, vol. 37, no. 6, pp. 1295-1304.
- [16] Koudriakova, T., Iatsimirskaia, E., Utkin, I., Gangl, E., Vouros, P., et al., 1998, "Metabolism of the Human Immunodeficiency Virus Protease Inhibitors Indinavir and Ritonavir by Human Intestinal Microsomes and Expressed Cytochrome P4503A4/3A5: Mechanism-Based Inactivation of Cytochrome P4503A by Ritonavir", *Drug Metabolism and Disposition*, vol. 26, no. 6, pp. 552-561.
- [17] Willson, T.M. & Kliewer, S.A. 2002, "Pxr, car and drug metabolism", *Nat Rev Drug Discov*, vol. 1, no. 4, pp. 259-266.
- [18] Broughton, J.Q., Abraham, F.F., Bernstein, N. & Kaxiras, E. 1999, "Concurrent coupling of length scales: Methodology and application", *Phys.Rev.B*, vol. 60, no. 4, pp. 2391-2403.
- [19] VPHnoe 2010, , *Virtual Physiological Human network of excellence*. Available: <http://www.vph-noe.eu/> [2011, 3/1/2011] .
- [20] Hoekstra, A.G., Lorenz, E., Falcone, J. & et al. 2007, "Towards a complex automata framework for multi-scale modeling", *International Journal for Multiscale Computational Engineering*, vol. 5, pp. 491-502.
- [21] Meier-Schellersheim, M., Fraser, I.D.C. & Klauschen, F. 2009, "Multiscale modeling for biologists", *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 1, no. 1, pp. 4-14.
- [22] Schnell, S., Grima, R. & Maini, P.K. 2007, "Multiscale modeling in biology", *American Scientist*, vol. 95, no. 1, pp. 134-142.
- [23] Lees, P., Cunningham, F.M. & Elliott, J. 2004, "Principles of pharmacodynamics and their applications in veterinary pharmacology", *Journal of Veterinary Pharmacology and Therapeutics*, vol. 27, no. 6, pp. 397-414
- [24] Kim, H.D. & O'Shea, E.K. 2008, "A quantitative model of transcription factor-activated gene expression", *Nat Struct Mol Biol*, vol. 15, no. 11, pp. 1192-1198
- [25] Imperial College 2008, *Synthetic Biology*. Available: http://openwetware.org/wiki/Imperial_College/Courses [2011, 2/23/2011]
- [26] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. & Kummer, U. 2006, "COPASI - a Complex Pathway Simulator", *Bioinformatics*, vol. 22, no. 24, pp. 3067-3074.
- [27] EU MAPPER Project 2010, , *MAPPER Project*. Available: <http://www.mapper-project.eu/> [2011, 3/3/2011] .
- [28] Harel, D. 2005, "A Turing-like test for biological modeling", *Nat Biotech*, vol. 23, no. 4, pp. 495-496.

Computerized Platform for Optimal Organ Allocations in Kidney Exchanges

Yanhua Chen, Jack D. Kalbfleisch, Yijiang Li, Peter X.-K. Song and Yan Zhou

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

E-mail: {chenyanh, jdkalbfli, yijiang, pxsong, zhouyan}@umich.edu

Abstract—*Kidney transplantation has emerged as the treatment for the most serious forms of kidney disease, but the supply of kidneys from deceased donors cannot meet the fast-growing demand. Recently, Kidney Paired Donation (KPD) program, a modality which enables willing but incompatible live donor-candidate pairs to swap donors, offers a promising solution for closing the gap between kidney supply and demand. Most of current KPD programs focus mainly on organ allocations strategies achieving the maximum number of transplants or matches. However, patients' quality of life after transplants can be more important for kidney candidates. In this paper, we propose a novel algorithmic platform to optimize cross-matches with the maximum benefits for donor-candidate pairs. Utilizing the power of integer programming, our platform implements a recently proposed method that takes probabilistic-based utility as the objective function, so that the overall expected utility, instead of the number of matches, is maximized. Moreover, involving altruistic donors in the allocations lead to a significant improvement in successful transplants. Empirically, we demonstrate the computerized platform for optimal organ allocations in kidney exchanges through extensive simulation experiments.*

Keywords: Kidney exchange; Optimal matching; Integer programming; Computerized platform

1. Introduction

Kidney transplantation has emerged as the treatment for the most serious forms of kidney disease. However, there is a considerable shortage of donor kidneys in the U.S.: more than 80,000 patients are on the waiting list for transplants by the end of 2010 [9]. In the real world clinical application, deceased donation and living donation are the two resources of organs for kidney transplantation, and living-donor transplant has a higher chance of success. Unfortunately, about one-third of patients with willing live donors will be excluded from kidney transplantation because of ABO blood type mismatch or HLA incompatibility [8]. ABO blood type mismatch infers to: type O people are universal donors for any candidates; people who have type AB blood can donate to only the same blood type patients; and a type A or B donor can donate to the same type or a type AB candidate. HLA incompatibility occurs when a recipient

candidate is sensitized to some of the Human Leukocyte Antigens (HLA) of his/her willing donor. Therefore, KPD program is established as a promising clinical solution to overcome the shortage of donors. The essential idea of such program is to exchange living kidney donors between two willing but incompatible donor-candidate pairs. The fundamental question in the KPD program is how to make an optimal decision of kidney exchanges that benefit patients the best.

An Integer Programming (IP) approach is widely used to tackle the optimization problem of selecting the optimal matches among incompatible donor-candidate pairs. Unfortunately, most of all current methods focus on determining the optimal two-way and/or three-way cycle exchanges through the means of graphic representation. Such constraint on the length of cycles to be less than 3 is imposed due to logistic consideration [1]. In this setting, many articles have considered to maximize the total number of transplants; see for examples, [11], [12], [14], [13], [1], [3]. In the real kidney exchanges, it is not only necessary to consider how to increase the number of transplants, but also needs to improve the quality of life for recipients after their transplants so that the transplants can make them live better. Therefore, we consider an expected-utility-based algorithm proposed by [6], which takes account of the medical-outcome-based utility (e.g., the life years gained from real transplants (LYFT) [16]) as well as the probability of successful actual transplants. In addition, most of the KPD exchanges only consider the paired donor-candidates to swap donors between them. Recently, these swaps also include chains triggered by altruistic donors (ADs) because chains offer more advantages [10], [4], [2]. On the one hand, it relaxes the reciprocity requirement of KPD, so pairs can find a donor from other pairs or ADs, rather than matching both the donor and candidate of another pair. More importantly, the simultaneity requirement of KPD is relaxed, even if one donor of chain reneges, the candidate has some opportunity to get transplants. Therefore, we integrate ADs into the expected-utility-based algorithm to improve the kidney exchanges. The idea is to define a virtual recipient for an AD and carry out the similar optimization using the algorithm of paired exchanges. A complete review of KPD program is presented in [15].

In summary, we implement an innovative method that

takes account of utility and uncertainty into the optimization of graph matching and further integrates ADs into the traditional KPD program. Through simulation experiments, we demonstrate the superiority of the expected-utility-based approach in comparison to the existing allocation strategies. Thus, our algorithmic platform brings more benefits for a greater number of kidney patients. In addition, we develop a general KPD graphic user interface (GUI) software that allows to model, visualize, and monitor the real world kidney exchanges. The remainder of the paper is organized as follows. We first present the mathematical formulation, optimization algorithm and theoretical work of kidney exchange problem in details in Section 2. In Section 3, we provide thorough computerized platform, experimental results and GUI software. Finally, we give a conclusion and discuss some future work in Section 4.

2. Optimization Algorithm

2.1 Graph-based Formulation

A kidney exchange problem can be represented as a directed graph $G = (V, E)$. Let $|V|$ be the number of vertices (nodes) and $|E|$ be the number of edges in a KPD graph, where $|\cdot|$ denotes cardinality. Figure 1 shows an example. Each vertex in graph G represents an incompatible donor-candidate pair (e.g., vertex 1) or an altruistic donor (e.g., vertex 7). Each edge from vertex i to vertex j indicates that the donor kidney in vertex i is compatible with the candidate in vertex j (e.g., $7 \rightarrow 1$). In this directed graph, each edge is assigned a weight representing *edge utility* e_{ij} of the kidney transplant from the donor in vertex i to the candidate in vertex j (e.g., $e_{71} = 9$). In addition, an *edge probability* p_{ij} is considered for each edge to reflect the chance of an actual successful kidney transplant from i to j (e.g., $p_{71} = 0.6$). All the directed edges are established for compatibility of ABO blood types and HLA sensitization.

The goal of optimization for kidney exchange program is to find a collection of mutually disjoint cycles or chains that attain the maximum overall expected utility of graph G . This task of optimizing matches on graph G can be realized by the following setup of an integer programming [6]:

$$\begin{aligned} & \max \sum_{c \in C} y_c u_c, \\ & \text{s.t. } y_c \in \{0, 1\}, \forall c \in C, \\ & \sum_{c \in C(i)} y_c \leq 1, 1 \leq i \leq |V|. \end{aligned} \quad (1)$$

where C is the exchange set of all cycles or chains with length 2 and/or 3 in graph G . $C(i)$ is the exchange set of cycles or chains in C that contain vertex i and y_c is a vector of indicators representing if cycle or chain exchange set c is to be executed for transplant ($y_c = 1$) or not ($y_c = 0$). Notice that u_c is the expected utility of cycle or chain exchange set c , which has been fully discussed in [6]. According to [6], where $u_c = \sum U_c P_c$.

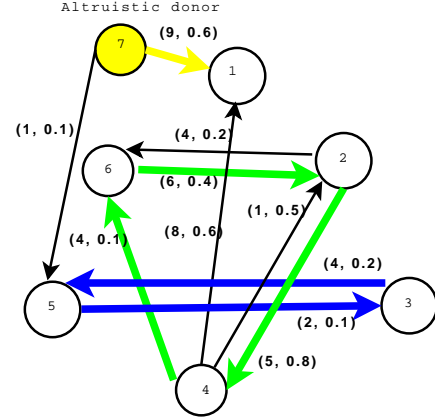


Fig. 1: A toy kidney exchange program including an altruistic donor and six incompatible pairs. It contains 3 two-way cycles ($\{2,4\}$, $\{2,6\}$, $\{3,5\}$), 1 three-way cycle ($\{6,2,4\}$) and 3 chains beginning with an altruistic donor ($\{7,1\}$, $\{7,5\}$, $\{7,5,3\}$). The optimal matches selected by IP are: $\{7,1\}$, $\{6,2,4\}$, and $\{3,5\}$, which represent the optimal exchanges $7 \rightarrow 1$, $6 \rightarrow 2 \rightarrow 4 \rightarrow 6$ and $3 \rightarrow 5 \rightarrow 3$.

U_c is the maximum utility of the possible exchange sets in c , while $P_c = \prod_{\substack{i,j \in c \\ e_{ij} \in E_s}} p_{ij} \prod_{\substack{i,j \in c \\ e_{ij} \in (1-E_s)}} (1 - p_{ij})$ for the corresponding exchange sets c , where E_s indicates a set of edges e_{ij} leading to actual transplants. Therefore, the calculation of expected utility is based on all possible configurations in exchange set corresponding to each edge either resulting in an actual successful transplant or not in the real lab match run. And for each such configuration, we aim to choose the available cycle that yields the highest expected utility. In addition, the expected utility of a chain initiated by an AD can be computed in a similar way except creating a dummy cycle from the ending vertex of chain. For example, add a dummy edge from vertex 1 to vertex 7 with edge utility $e_{17} = 0$ and edge probability $p_{17} = 1$, which results in a 2-way cycle $\{7,1\}$. In Figure 1, using the above formula, we compute the expected utilities of cycles as $u_{\{2,4\}} = 2.4$, $u_{\{2,6\}} = 0.8$, $u_{\{6,2,4\}} = 3.35$, $u_{\{3,5\}} = 0.12$. Also, the expected utilities of chains are calculated as $u_{\{7,1\}} = 5.4$, $u_{\{7,5\}} = 0.1$, $u_{\{7,5,3\}} = 0.156$. Then, plugging the expected utilities u_c into Equation (1), we use IP to find the optimal solution of the virtual matches: $7 \rightarrow 1$, $6 \rightarrow 2 \rightarrow 4 \rightarrow 6$ and $3 \rightarrow 5 \rightarrow 3$. Finally, not all the optimal virtual matches lead to actual operations. For instance, some higher order cycles (e.g. three-way cycles) are less likely to be chosen because such cycles tend to be more difficult to successful carry out [1]. If lab match run suggests one transplant fails (e.g., edge e_{62} is broken), then the entire three-way exchange $6 \rightarrow 2 \rightarrow 4 \rightarrow 6$ is labeled as a failure in the existing methods. However, [6] suggests a method with fall-back option; that is, we can choose the kidney exchange between 2 and 4 as a sub-cycle. As a result, the transplants now include $7 \rightarrow 1$, $2 \rightarrow 4 \rightarrow 2$, and $3 \rightarrow 5 \rightarrow 3$.

2.2 Algorithm

The computerized platform for kidney exchanges is based on a graphic optimization algorithm, described in detail as the following steps:

- 1) Given incompatible donor-candidate pairs and ADs at time $t = 0$, build a directed graph $G = (V, E)$ with each vertex representing a donor-candidate pair or an AD and each edge from vertex i to j denoting compatibility, so that there is a possibility match between the donor in vertex i to the candidate in vertex j .
- 2) Assign edge utility e_{ij} and edge probability p_{ij} to each match pair of donor i and candidate j . e_{ij} is derived from medical-outcome-based utility or some existing KPD scoring systems [11], and p_{ij} is derived from a statistical model for probability of successful transplants.
- 3) Find chains beginning at vertices of ADs with length size equal to 2 and/or 3.
- 4) Add dummy edges from the end vertices of chains to ADs, on which assign the edge utility $e_{ij} = 0$ and the edge probability $p_{ij} = 1$.
- 5) Find all cycles with length size 2 and/or 3 in graph G using the depth-first search algorithm.
- 6) Compute the expected utility u_c according to the configuration of each cycle exchange set.
- 7) Solve Equation (1) to get indicators y_c representing the optimal virtual donor-candidate matches.
- 8) Determine the final optimal kidney transplants according to Bernoulli trails with a certain success probability in the real lab match run. If such a Bernoulli trial is realized, the transplant will lead to an successful operation; otherwise, it fails.
- 9) Compute the number of completed transplants and associated utility of optimal kidney transplants.
- 10) Remove the vertices of donor-candidate pairs and ADs that finish successful transplants from graph G , and those end vertices of chains are "bridge donors" [10] as new ADs.
- 11) At time $t = t + 1$, form the new incompatible donor-candidate pairs and ADs based on pair arrival rate λ according to a Poisson process, then go to step 1).

2.3 Theoretical Analysis

In this section, we show that the decision version of our algorithm for kidney exchange program is NP-complete given in Equation (1).

Theorem 1: Given a graph $G = (V, E)$ and an integer n ($n \geq 3$), the problem of deciding if G admits a perfect cycle/chain cover containing cycles/chains of length at most n is NP-complete.

Our proof of Theorem 1 follows that in [1]. First, it is easy to demonstrate this problem is in NP. Second, we can prove

that it is NP-hard through a reduction from a 3D-Matching problem. Due to the space limitation, we omit detail of the proof.

3. Experiments

3.1 Computerized Platform and Evaluation Measurement

We tested the algorithm on a computerized platform by mimicking a general kidney exchange simulation system proposed in [6], which appropriately reflects the real world clinical application. In this computerized platform, we hope to evaluate different kidney allocation strategies. The flowchart for the computerized platform is illustrated in Figure 2. First, we generated data of candidates and donors separately. Candidates are sampled at random from the University of Michigan kidney paired donation database, which currently has 187 incompatible donor-candidate pairs. This database provides us the important information of ABO blood type and HLA useful to characterized each sampled candidate. Donors, on the other hand, are generated by the population distributions of ABO and HLA. In particular, the distribution of ABO blood types for the US population is: $O(44\%)$, $A(42\%)$, $B(10\%)$, and $AB(4\%)$, according to Stanford Blood Center (2010) ¹, and the distribution of HLA is derived from HLA haplotypes frequencies of the US population [7]. Through random sampling, we can appoint ADs directly from the set of drawn donors or construct an incompatible donor-candidate pair if either their ABO blood types mismatch or HLA incompatibility. In this way, simulated donors and candidates reflect real-world of data. Second, KPD parameters needed for data generation, including an initial pair number n and percentage of ADs, are specified for the first match run. Third, a directed graph $G = (V, E)$ involving edge utilities and edge probabilities is created by using characteristics of candidates and donors. In this paper, for illustration, we assign values of edge utilities and edge probabilities according to uniform random distributions on interval denoted by $[\min, \max] = [a, b]$, such as $U[10, 20]$ and $P[0.1, 0.5]$, respectively. Fourth, for a given KPD graph, we find all cycles and chains with length size equal to 2 and/or 3 by the depth-first search algorithm. Furthermore, using IP optimization algorithm discussed in Section 2, we search for the optimal solution regarding the maximum potential matches (transplants) under each allocation strategy applying Gurobi optimization software [5]. Fifth, the ready transplant matches are finalized as actual successful transplants in the real lab match run according to Bernoulli trails with a certain success probability. At the end, the actual successful transplants are output from the platform. Moreover, in an evolving KPD program, successful donor-candidate matches will leave the database and some

¹http://bloodcenter.stanford.edu/about_blood/blood_types.html

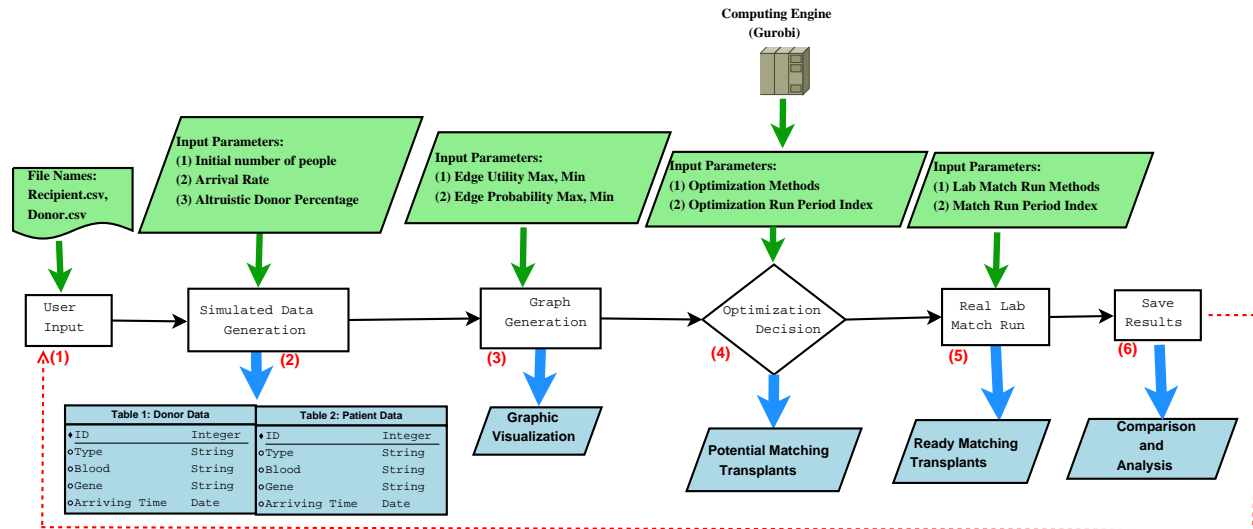


Fig. 2: A flowchart of computerized platform for kidney exchanges.

new pairs will enter into the pool according to a Poisson process with an arrival rate λ . Thus, a new match run will be performed at another time (see the dot line in Figure 2). In order to make a better comparison, we fixed the number of match runs as $k = 12$, mimicking the reality that there is one match run each month within a year. In the following simulation experiments, we evaluated the kidney exchange results based on two criteria: the accumulated number of transplants and accumulated utility. The higher the number of transplants or the utility is, the higher mutual benefits for the kidney transplant patients. For each allocation strategy, we conducted 100 test runs, and reported the averaged accumulated number of transplants and averaged accumulated utility.

3.2 Results

We began by creating a KPD pool of by specifying three input parameters: the initial number of pairs $n = 200$, the arrival rate of pairs $\lambda = 10$ or $\lambda = 20$, and the percentage of ADs 5%. Then we generated a directed graph by assigning edge utility and edge probability as $U[10, 10]$ and $P[0.1, 0.5]$, respectively. First, we aimed to compare two allocation strategies in terms of accumulated number of transplants, in the settings where the KPD only involved donor-candidate pairs (namely no ADs). The two strategies to be compared are (1) *Cycle-Without-AD-Base*: a traditional method that does not consider the expected utility in the optimization and fall-back option in the real lab match run; (2) *Cycle-Without-AD*: a new method [6] that uses the expected utility in the optimization and accounts for the fall-back option in the real lab match run. The accumulated number of transplants obtained by the two approaches with different arrival rates λ are shown in Figure 3. Generally, the accumulated number of transplants appears higher for

a larger number of arrival rate (e.g., $\lambda = 20$ in Figure 3(b) versus $\lambda = 10$ in Figure 3(a)). This implies that the more pairs participate in the kidney exchange program, the higher number of achieving matches in the KPD pool. Moreover, the accumulated number of transplants gained by the new approach (i.e., *Cycle-Without-AD*) is significantly higher than the traditional method (i.e., *Cycle-Without-AD-Base*) in the magnitude of 2–4 folds. These results indicate that the new approach is clearly advantageous to increase the number of transplants in kidney exchanges.

Next, we integrated the ADs into the new allocation strategy and investigated the role of ADs in the kidney exchanges. As discussed in Section 2, method *Cycle-With-AD* is modified by simply adding dummy edges from each donor-candidate vertex to the ADs with cycle-length size 2 and/or 3. Then we utilized the same optimization procedure as that of the *Cycle-Without-AD* method to find the optimal exchanges. Figures 4(a)-(c) display the accumulated number of transplants obtained by two strategies: (1) *Cycle-Without-AD* and (2) *Cycle-With-AD*, where the edge utility is generated by $U[10, 10]$, $U[10, 20]$, and $U[10, 30]$, and the arrival rate is assigned by $\lambda = 10$. In these panels, based on the accumulated number of transplants over 12 match runs, method *Cycle-With-AD* gives at least 10% more matches than the method without using ADs. Moreover, we plotted the results for the case of $\lambda = 20$ in Figures 4(d)-(f). Again, when more people enters, method with ADs clearly performed better than the one without ADs. In the meanwhile, we also compared accumulated utility of these two methods when the edge utility distribution changes from $U[10, 10]$ to $U[10, 30]$ in the cases of $\lambda = 10$ and $\lambda = 20$. From Figures 5(a)-(c), we noticed that the accumulated utility of the *Cycle-With-AD* method enjoys a gain between 15% to 30% over the *Cycle-Without-AD* method if $\lambda = 10$.

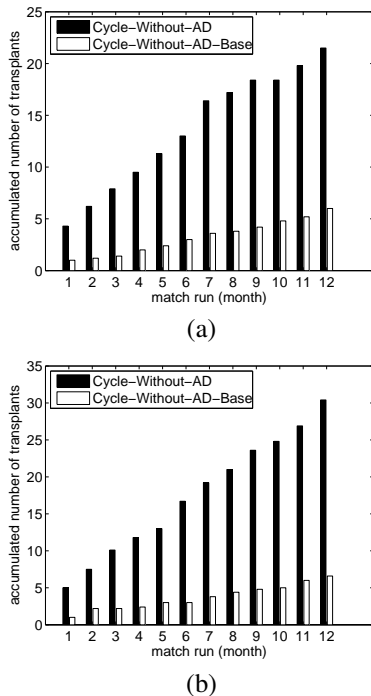


Fig. 3: Comparison of accumulated number of transplants versus month (number of match run) for *Cycle-Without-AD-Base* and *Cycle-Without-AD* methods with different arrival rate of pairs: (a) $\lambda = 10$, (b) $\lambda = 20$.

Likewise, Figures 5(d)-(f) report the accumulated utility of the method using ADs is about at least 10% higher than that of the method not using ADs if $\lambda = 20$. Therefore, it is obvious that on average the method without using ADs is consistently outperformed by the method using ADs over all match runs in terms of accumulated number of transplants and accumulated utility. As a result, using ADs in the kidney exchanges would help clinicians to achieve more number and better quality of transplants.

3.3 Software

In this paper, one of our new contributions is the development of a graphic user interface (GUI) software to visualize inputs and outputs in a kidney exchange program. Our simulation experiments above were carried out by using our GUI software developed by C++ language on a machine with Quad 3GHz Intel Core2 processors and 4GB RAM. The software offers a range of functions to create a user-friendly interface and builds appropriate configurations to support communications between inputs and outputs essential in the kidney exchanges. It includes six types of functional components associated with inputs and outputs, which are displayed in the middle panel of Figure 2: (1) reader of original data from internal and external files; (2) KPD data simulator; (3) KPD graph generator; (4) Optimizer of KPD kidney donation; (5) KPD lab match run; (6) output of graph

matching results. In addition, the input data or parameters are showed in the upper panel of Figure 2, while the output data or results are showed in the lower panel of Figure 2.

For instance, Figure 6 shows a snapshot of GUI software of kidney exchanges for five match runs by the *Cycle-With-AD* method. Relevant information is displayed in multiple-windows. *Recipient* (right-top) and *Donor* (right-middle) windows in Figure 6 show the randomly drawn kidney experimental data when the initial number of pairs, arrival rate and percentage of ADs are fixed as 50, 10 and 5% respectively. The display of data includes period (i.e., number of match run), ID, type of vertex (i.e., pair or AD), blood type or HLA type. If ID number is the same between recipient candidate and donor, it indicates a pair of originally incompatible donor-candidate, otherwise it denotes an AD. In the *Graph Builder* window (right-bottom), the corresponding directed graph is created with the edge utility and edge probability generated by $U[10, 10]$ and $P[0.1, 0.5]$, respectively. After selecting an optimization method, such as *Cycle-Without-AD-Base*, *Cycle-Without-AD*, or *Cycle-With-AD*, the center window will report the optimal graph matches between donors and recipients, including donor ID, donor type, recipient ID, recipient type, number of transplants and associated utility at each match run. Also, if desired, a further match run can be performed, leading to an evolving kidney exchange data exploration. In summary, the GUI provides a very powerful tool to help clinicians, donors and patients more easily analyze and assess the kidney exchange program.

4. Conclusions and Future Work

In this paper, we investigated a new kidney allocation strategy based on expected-utility to maximize the mutual benefits for kidney exchanges. The problem is formulated as to search for the maximum disjoint vertex sets in a weighted directed graph. First, a depth-first search algorithm is implemented to identify all cycles/chains with length size 2 and/or 3. Then, an optimal solution of maximum expected utility can be obtained by an IP. Finally, ADs are added to increase the possibility of exchanges. Through simulation studies that closely imitate the real application on computerized platform, we demonstrated that the expected-utility-based allocation strategy provides the higher quantity and quality of life than the current practising methods in the kidney exchanges. This will result in thousands of kidney patients for life-saving each year in USA.

All algorithms discussed in this paper have been fully integrated into a GUI software package, which will be released publicly through the necessary Institutional Review Board (IRB) regulations. In the future, we plan to conduct practical studies to solicit feedbacks so that the software can be improved with more user-friendly features for clinical convenience. We also intend to incorporate interaction tools for input data process, integration, and modeling, as well as

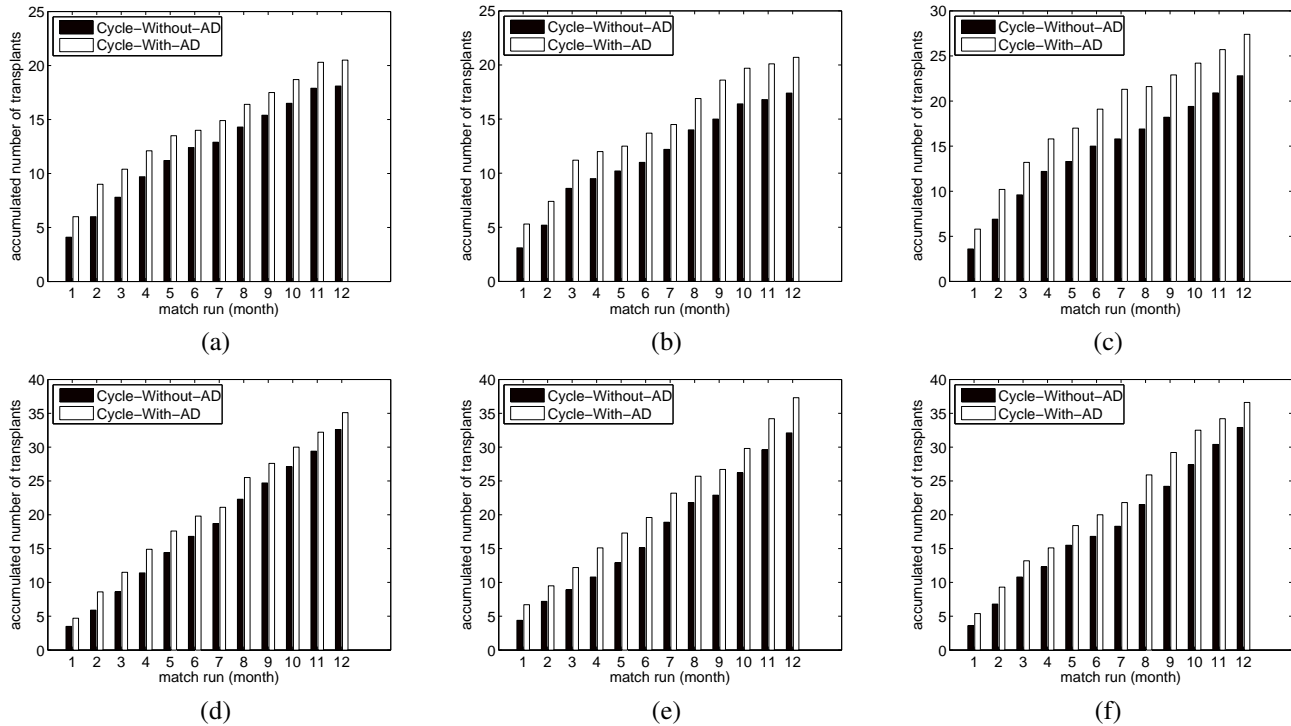


Fig. 4: Comparison of accumulated number of transplants versus month (number of match run) for *Cycle-Without-AD* and *Cycle-With-AD* methods with different arrival rate of pairs (λ) and different edge utility distributions (U): (a) $\lambda = 10$ and $U[10, 10]$, (b) $\lambda = 10$ and $U[10, 20]$, (c) $\lambda = 10$ and $U[10, 30]$, (d) $\lambda = 20$ and $U[10, 10]$, (e) $\lambda = 20$ and $U[10, 20]$, (f) $\lambda = 20$ and $U[10, 30]$.

output data graphical visualization into our existing system for its maximum flexibility of clinical practice.

Acknowledgment

We thank Dr. Alan Leichtman of Department of Internal Medicine, University of Michigan, for his invaluable support and constructive suggestions. This research was funded by U. S. NSF (National Science Foundation), CRA (Computing Research Association) and CCC (Computing Community Consortium) under sub-award CIF (Computing Innovation Fellows)-B-66 (2010-2011).

References

- [1] D. Abraham, A. Blum and T. Sandholm, "Clearing algorithms for barter exchange markets: enabling nationwide kidney exchanges," in *Proceedings of the 8th ACM conference on Electronic commerce*, pp. 295–304, 2007.
- [2] G. Ashlagi, A. Roth, and M. Rees. "Nonsimultaneous Chains and Dominos in Kidney Paired Donation – Revisited," *American Journal of Transplant*, vol. 11, pp. 1–11, 2011.
- [3] P. Biro, D. Manlove, and R. Rizzi, "Maximum weight cycle packing in directed graphs, with application to kidney exchange programs," in *Discrete Mathematics, Algorithms and Applications*, vol. 1, no. 4, pp. 499–517, 2009.
- [4] S. Gentry, R. Montgomery, B. Seihart and D. Segev, "The roles of dominos and nonsimultaneous chains in kidney paired donation," *American Journal of Transplant*, vol. 9, pp. 1330–1336, 2009.
- [5] The website: <http://www.gurobi.com/>
- [6] Y. Li, J. Kalbfleisch, P. Song, Y. Zhou, A. Leichtman and M. Rees, "Optimization and simulation of an evolving kidney paired donation (KPD) program," *Department of Biostatistics, University of Michigan, Working Paper Series, Working Paper 90*, <http://www.bepress.com/umichbiostat/paper90>, May 2011
- [7] M. Maier, L. Gragert, and W. Klitz, "High-resolution HLA alleles and haplotypes in the United States population," *Human Immunology*, vol. 68, no. 9, pp. 779–788, 2007.
- [8] R. Montgomery. "Renal transplantation across HLA and ABO antibody barriers: integrating paired donation into desensitization protocols," *American Journal of Transplant*, vol. 10, pp. 449–457, 2010.
- [9] The website: <http://optn.transplant.hrsa.gov/>
- [10] M. Rees, J. Kopke, R. Pelletier, D. Segev, M. Rutter, A. Fabrega, J. Rogers, O. Pankewycz, J. Hiller, A. Roth, T. Sandholm, M. Unver, and R. Montgomery, "A non-simultaneous extended altruistic donor chain," *New England Journal of Medicine*, vol. 360, no. 11, pp. 1096–1101, 2009.
- [11] A. Roth, T. Sonmez, and M. Unver, "Kidney exchange," *Quarterly Journal of Economics*, vol. 119, no. 2, pp. 457–488, 2004.
- [12] A. Roth, T. Sonmez, and M. Unver, "A kidney exchange clearinghouse in New England," *American Economic Review*, vol. 95, no. 2, pp. 376–380, 2005.
- [13] A. Roth, T. Sonmez, and M. Unver, "Efficient kidney exchange: Concidence of wants in a market with compatibility-based preferences," *American Economic Review*, vol. 97, no. 3, pp. 828–851, 2007.
- [14] S. Saidman, A. Roth, T. Somez, M. Unver, and F. Delmonico, "Increasing the opportunity of live kidney donation by matching for two and three way exchanges," *Transplantation*, vol. 81, no. 5, pp. 773–782, 2006.
- [15] C. Wallis, K. Samy, A. Roth and M. Rees. "Kidney paired donation," *Nephrol Dial Transplant*, vol. 0, pp. 1–9, 2011.
- [16] R. Wolfe, et al., "Calculating life years from transplant (LYFT): method for kidney and kidney-pancreas candidates," *American Journal of Transplant*, vol. 8, pp. 997–1011, 2008.

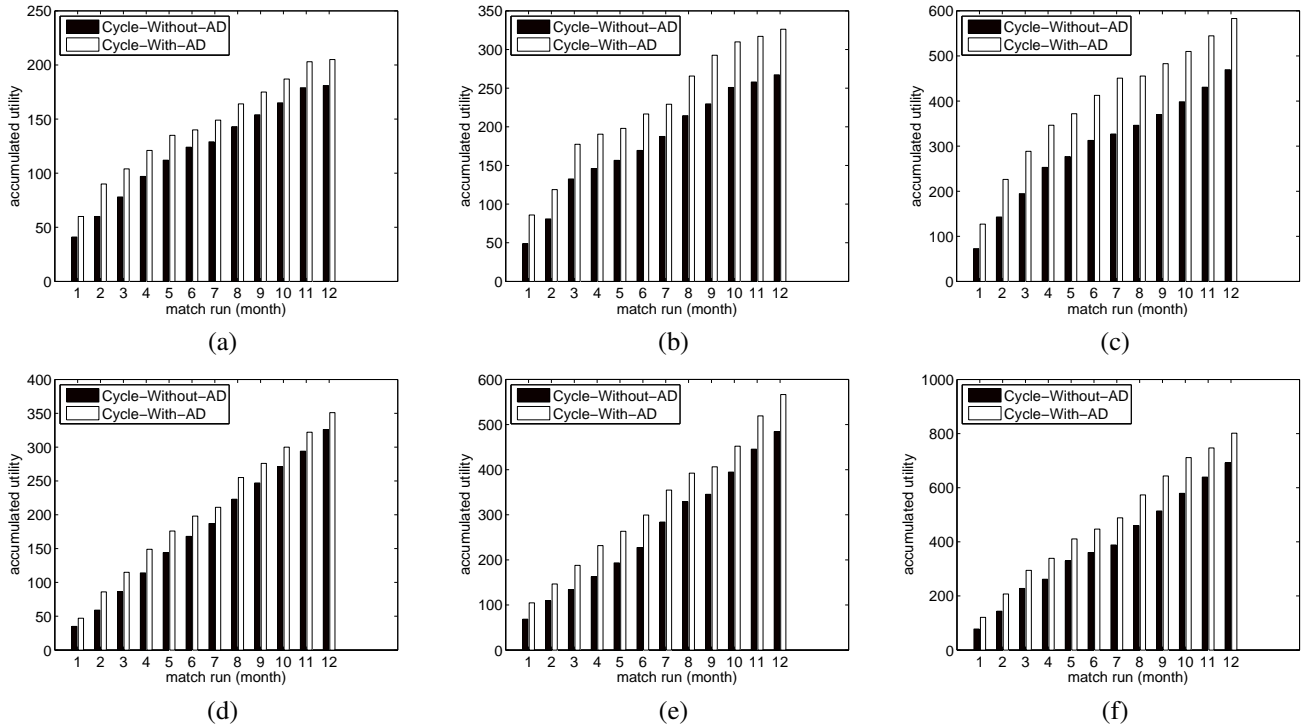


Fig. 5: Comparison of accumulated utility versus month (number of match run) for *Cycle-Without-AD* and *Cycle-With-AD* methods with different arrival rate of pairs (λ) and different edge distributions (U): (a) $\lambda = 10$ and $U[10, 10]$, (b) $\lambda = 10$ and $U[10, 20]$, (c) $\lambda = 10$ and $U[10, 30]$, (d) $\lambda = 20$ and $U[10, 10]$, (e) $\lambda = 20$ and $U[10, 20]$, (f) $\lambda = 20$ and $U[10, 30]$.

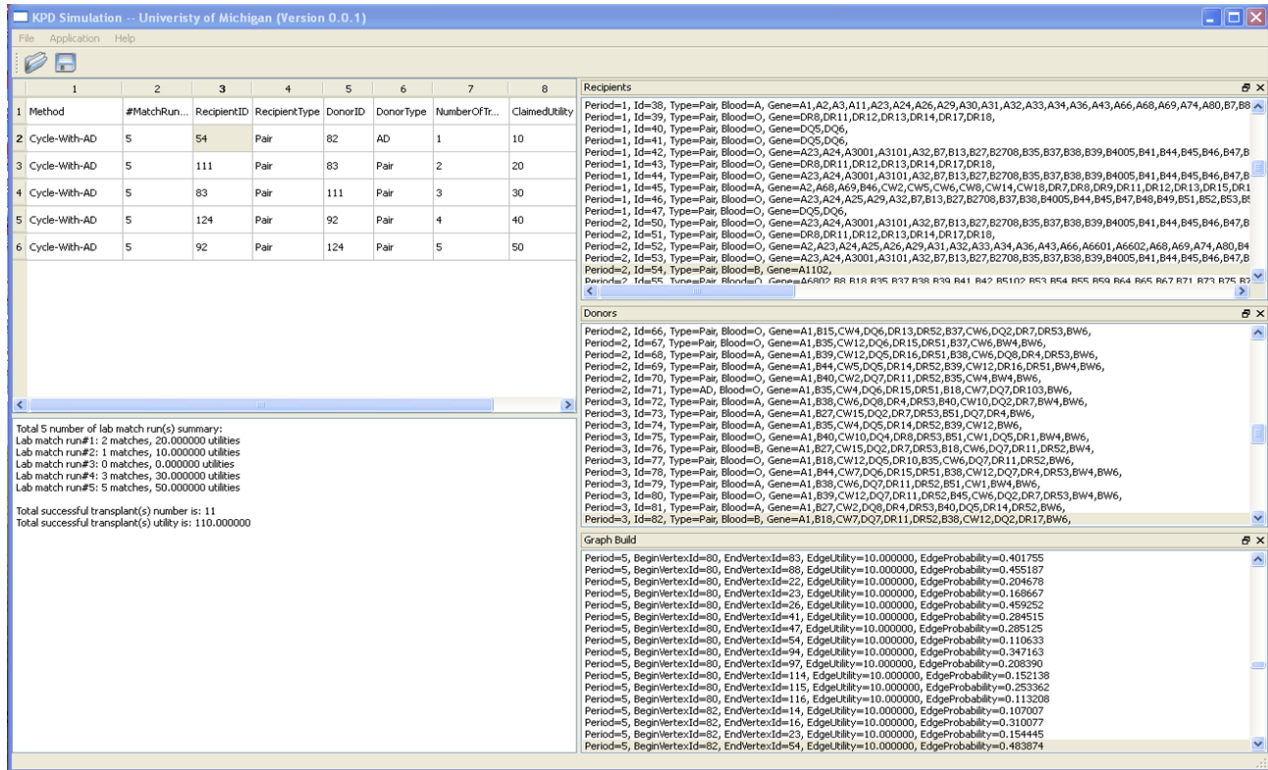


Fig. 6: A GUI example for kidney exchanges

Application of bioinformatics models to define influenza virus A subtypes

M. Ebrahimi¹, P. Agha-Golzadeh², E. Ebrahimie² and N. Shamabadi³

¹Department of Biology & Bioinformatics Research Group, University of Qom, Qom, Iran

²Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran

³Young Researcher Club, Qom Branch, Islamic Azad University, Qom, Iran

Abstract - *Influenza A viruses infect large numbers of animals and are subtyped according to their surface antigens to 16 HA subtypes and 9 NA subtypes. To identify the main prominent protein attributes representing each subtype, various clustering, screening, item set mining and decision tree models applied to dataset of 3632 HA sequences of influenza A viruses. The count of Tyr, Gln and Phe and the count of some hydrophilic – hydrophobic (such as Lys – Val, Asn – Leu and Pro – Leu) were the most important protein features. Most decision tree models used non-reduced absorption at 280nm as the main protein feature to build the trees. Parallel stump and ID3 numeric decision tree algorithms were the best tree to differentiate between HA subtypes. The results showed various bioinformatics tools may be used in this regard. For the first time, this paper showed that protein attributes can be used to differentiate between influenza A subtypes.*

Keywords: Influenza A, Bioinformatics, Modelling

1 Introduction

Influenza is a highly contagious and acute respiratory disease with a high degree of morbidity and has been in circulation for centuries [1]. The disease is caused by the influenza virus, which is a segmented, enveloped RNA virus. Within the influenza virus family, there are four genera: A, B, C virus and Thogoto virus; although only A and B cause significant disease in humans [2]. Influenza A viruses are further subtyped according to their surface antigens, HA and NA, of which 16 HA subtypes and 9 NA subtypes have been identified to date [3]. The HA and NA genes are extremely variable in sequence, and less than 30% of the amino acids are conserved among all the subtypes. New epidemic strains of influenza A arise due to point mutations within two surface glycoproteins, HA and NA. These changes in HA and NA enable emerging virus strains to evade the host's immune system and therefore necessitates the annual revision of vaccine to include the new viruses [4]. Furthermore, HA may also play a structural role in budding and particle formation. Human influenza viruses manage to cause epidemics almost every year. The circulating viruses change their surface glycoproteins by accumulating mutations (antigenic drift or

antigenic shift) which results in variant viruses of the same subtype that are able to evade the immune pressure in the population [5].

Bioinformatics represents a new field at the interface of the twentieth-century revolutions in molecular biology and computers. A focus of this new discipline is the use of computer databases and computer algorithms to analyze proteins and genes. A major challenge in biology is to make sense of the enormous quantities of sequence data and structural data that are generated by genome-sequencing projects, proteomics, and other large-scale molecular biology efforts. Fitting a model such as a decision tree or item set mining to a set of variables this large may require more time than is practical [6]. A decision tree is constructed by looking for regularities in data, determining the features to add at the next level of the tree using an entropy calculation, and then choosing the feature that minimizes the entropy impurity [7]. To better understand the features that contribute to structural differences between influenza viruses A subtypes, it is necessary to identify the main features responsible for this valuable characteristic. Herein we used various clustering, screening, item set mining and decision tree models to determine which protein attributes may be used as a marker between subtypes of influenza A viruses. All available HA sequences (3632) of influenza A viruses from Swiss-Prot database were extracted and up to 924 protein features for each HA protein sequence was generated and various bioinformatics modeling techniques applied on this.

2 Methods and Materials

Three thousand and six hundred and thirty two sequences of HA virus proteins from various species (human, bird, pig, horse, mouse, tiger, leopard, dog, and cat) were extracted from the UniProt knowledgebase database and categorized as H1 to H16, according to database classification. Nine hundred and twenty four protein features or attributes including primary and secondary protein features were extracted. A dataset of these protein features was imported into Clementine software [Clementine_NLV-11.1.0.95; Integral Solution, Ltd.], null data for subtype of virus was discarded, and this feature was set as the output variable and the other variables were set as input variables. The same database imported into RapidMiner software [RapidMiner 5.0.001, Rapid-I GmbH, Stochumer Str. 475, 44227 Dortmund, Germany] and again the subtype of virus set as target or label attribute [when Item

Set Mining model performed, no label or target attribute was set as this model requires so]. To minimize the effects of correlated features on modelling and to decrease the processing time and burden on processing facilities, the original database subjected to remove correlated features algorithm, so the number of protein attributes (variables) decreased from 924 to 486 attributes. Various algorithms such as screening models [Anomaly detection model, feature selection algorithm or attribute weighting], clustering models [K-Means, TwoStep cluster], Tree Induction models [with various criterion, C5.0, C5.0 with 10-fold cross Validation and C&RT], Item Set Mining [FPGrowth] and Rule Induction model [10 fold cross-Validation through stratified sampling] run on each dataset as described previously [8]. Whenever requested by model, data were discretized by the frequency; i.e. data were divided into 3 bins [ranges] with nearly equal the frequencies in each class [low 0-0.3, mid 0.3-0.5 and high >0.5]; and sometimes data were converted to nominal and in some cases to binominal datasets.

3 Results

The number of protein attributes gained weights higher than 0.7 in each weighting model were as follows: PCA 2, SVM 24, relief 4, uncertainty 17, gini index 280, chi squared 39, deviation 2, rule 59, gain ratio 61, info gain 350 and info gain ratio 13.

The most important feature used to build the tree was non-reduced absorption at 280nm. If the value for this protein attribute was higher than 1.180 and the value for the count of Trp – Ala was higher than 0.500 and the count of Gly was higher than 49, the viral protein was originated from H10; otherwise from H3. If the count of Trp – Ala was equal to or less than 0.500, then the count of Ala – Ala (value of 3.500), the length of protein (value of 566) and the count of Trp – Asn (value of 0.500) used to differentiate between H14, H4, H8 and H9 groups. When the count of Trp – Asn was higher equal to or less than 0.500, if the count of Ser – Cys, non – reduced absorption at 280nm and aliphatic index were higher than 1.500, 1.44 and 86.690, respectively, the protein originated from H16; otherwise from H13. With the count of Ser – Cys was equal to or less than 1.500 and the count of His – Asn was higher than 0.500 and the count of Glu – Trp was higher than 0.500, if the count of Gly was higher than 44.500, virus belonged to H2, otherwise to H5 group. With the count of Glu – Trp (< 0.500) and the count of His – Asn (<.500), the virus HA proteins belonged to H1 and H6, respectively. If non-reduced absorption at 280nm was < 1.180 and the aliphatic index was > 81.875, the protein belonged to H12 group, if not to H15 or H7.

Stump decision tree model created a very simple tree with non-reduced absorption at 280nm variable as the root feature. Decision Tree Stump (Parallel) generated a tree again with the same starting attribute. More complex tree generated by ID3 Numerical method and again tree built on non-reduced absorption attribute. Random tree started with another protein attribute, the count of His – Ala. When value for this attribute

was higher than 1.500 and the count of Ala was higher than 26.500, the protein fell into H6 group. if the count of His-Ala was higher than 1.500 and the count of Ala was less than or equal to 26.500, the virus protein identified as H16. Ten different models were used by Random Forest algorithm to induce decision trees. In the first model, the count of Met-Ala was the main feature used by this method to induce the tree and its branches was created using the count of Gly and the count of Val – Arg attributes to classify H2, H5, H10, H9, H8, H7, H1 and H11 subtypes. In the second model, the count of Gly – Ala, the frequency of Pro – Ile, the count of Asn – Cys, the frequency of Pro – Ile, the count of Met – Lys and the count of Leu – Trp to trace H6, H11, H1, H3, H5, H13, H2 and H9 subtypes. The count of Gly – Met, the count of Cys – His and the count of sulfur were the most important attributes to build the tree by the third model (H10, H3, H9H4 and H5). Random forest, the fifth model, was able to differentiate between H10, H1, H4 and H3 by inducing a tree with the frequency of Pro – Ser as the main feature and the count of Cys – Met as the other important feature. In other models the count of Gln – Phe, the count of Trp – Pro and the count of Ala – Ala (model 5), the count of His – Phe, the count of Ile – Phe, the count of Leu – Lys and the count of Ala – Gln (model 6), the count of Gln – Gln and the count of Gln – Tyr (model 7), the count of Phe – Lys, the count of Asn – His and the count of Ser – Pro (model 8), the count of Gly – Met, the count of Gly – Val, the count of Asp – Gly and the count of Pro – Ala (model 9) and the count of Trp – Met (model 10) were the most important features used to build the trees.

GRI node analysis created 100 rules with 3631 valid transactions with minimum and maximum support of 44.09 % and 44.09 %, respectively, while maximum confidence reached 100 %. When feature selection was used, minimum support, maximum support, maximum confidence, and minimum confidence were the same as previous. In both methods [with/without feature selection filtering] the count of Gln – Leu, the frequency of Gly – His and the frequency of Pro – Asn were the main features used to create the first rules.

4 Discussion

Although the numbers of attributes with weights equal to or higher than 0.70 varied from 2 (in PCA weighting) to 62 (in Info Gain Ratio and Rule Induction weighting), the percentage and the count of Tyr, the frequency and the count of Lys - Val, the percentage, the frequency and the count of Gln, the frequency and the count of Asn – Leu, the count of Pro – Leu, the percentage of Phe and the frequency of Ser – Ile chosen by 7 attribute weightings as one of the most important attributes. When the same models run on dataset with correlated remove features, only six attributes gained weights higher than 0.70; again the count of Tyr, the count of Gln, the count of Lys – Val and the count of Asn – Leu were the most important features with weights higher than 0.70. The count of Gln – Asn was the other weight higher than 0.70. More than 50% of features gained high weights in both models were hydrophobic amino acids and the rest were mainly from hydrophilic amino acids. For the first time the

importance of dipeptides in classifying the influenza virus A has been presented here. The combination of one hydrophobic amino acid such as Val, Leu or Ile and one hydrophilic amino acid such as Asn, Ser or Gln forms a strong link inside the protein and reduce the possibility of mutations in this area; but when there are hydrophobic dipeptides connections, the chance of mutation and flexibility increases.

Although some trees generated by tree induction models had just two branches, as seen in stump decision tree, the depth of trees in some models were more complicated [more than 12 branches in ID3 numeric run on information gain]. The ability of various decision tree induction models applied in this study to correctly and effectively classify influenza A subtypes based on protein attributes were very different. In some models, two or three classes were identified, showing the model was not competent in this field (as seen in decision tree stump, C5.0, C&RT, random tree and accuracy). But in some other models, such as decision tree run on removed correlated features' dataset, decision tree stump parallel and ID3 numeric, the models were able to completely classify the HA subtypes (H1 – H16) based on their protein features. So the latter models may be used as a suitable tool to classify those viral subtypes.

The results showed that various bioinformatics tools and modelling facilities can be used to identify the subtypes of influenza virus A with a precision rate up to 95%. To our knowledge, for the first time we showed that some primary or secondary attributes can be used to differentiate between various subtypes of influenza A viruses.

5 References

- [1] P. Chadha and R. H. Das, "A pathogenesis related protein, AhPR10 from peanut: an insight of its mode of antifungal activity," *Planta*, vol. 225, pp. 213-22, Dec 2006.
- [2] A. M. Ledebner, *et al.*, "Cloning of the natural gene for the sweet-tasting plant protein thaumatin," *Gene*, vol. 30, pp. 23-32, Oct 1984.
- [3] C. Kuwabara, *et al.*, "Abscisic acid- and cold-induced thaumatin-like protein in winter wheat has an antifungal activity against snow mould, *Microdochium nivale*," *Physiol Plant*, vol. 115, pp. 101-110, May 2002.
- [4] H. Breiteneder and C. Ebner, "Molecular and biochemical classification of plant-derived food allergens," *J Allergy Clin Immunol*, vol. 106, pp. 27-36, Jul 2000.
- [5] Y. Tada and T. Kashimura, "Proteomic analysis of salt-responsive proteins in the mangrove plant, *Bruguiera gymnorhiza*," *Plant Cell Physiol*, vol. 50, pp. 439-46, Mar 2009.
- [6] A. M. Casas, *et al.*, "Expression of Osmotin-Like Genes in the Halophyte *Atriplex nummularia* L.," *Plant Physiol*, vol. 99, pp. 329-37, May 1992.
- [7] D. Goel, *et al.*, "Overexpression of osmotin gene confers tolerance to salt and drought stresses in transgenic tomato (*Solanum lycopersicum* L.)," *Protoplasma*, vol. 245, pp. 133-41, Sep 2010.
- [8] A. Leone, *et al.*, "Comparative Analysis of Short- and Long-Term Changes in Gene Expression Caused by Low Water Potential in Potato (*Solanum tuberosum*) Cell-Suspension Cultures," *Plant Physiol*, vol. 106, pp. 703-712, Oct 1994.
- [9] M. Ebrahimi and E. Ebrahimie, "Sequence-based prediction of enzyme thermostability through bioinformatics algorithms," *Current Bioinformatics*, vol. 5, pp. 195-203, 2010.
- [10] M. Ebrahimi, *et al.*, "Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms," *EXCLI Journal*, vol. 8, pp. 218-233, 2009.
- [11] M. Ebrahimi, *et al.*, "Are there any differences between features of proteins expressed in malignant and benign breast cancers?," *Journal of Research in Medical Sciences*, vol. 15, pp. 299-309, 2010.
- [12] E. Ebrahimie, *et al.*, "Investigating protein features contribute to salt stability of halolysin proteins," *Journal of Cell and Molecular Research*, vol. 2, pp. 15-28, 2010.
- [13] E. Ashrafi, *et al.*, "Determining specific amino acid features in P1B-ATPase heavy metals transporters which provides a unique ability in small number of organisms to cope with heavy metal pollution " *Bioinformatics and Biology Insights*, vol. Accepted, 2011.
- [14] M. Ebrahimi, *et al.*, "Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms," *EXCLI Journal*, vol. 8, pp. 218-233, 2009.

Simulated Docking of Oseltamivir with an Avian Influenza (A/H5N1) Neuraminidase Active Site

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA

Abstract

Given the lead-time currently required for vaccine production, a widespread administration of effective anti-influenza therapeutics is the only practical defense against a 1918-scale influenza pandemic after the pandemic begins. Neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. The neuraminidase inhibitor oseltamivir is currently the most widely used anti-flu therapeutics. Oseltamivir was ineffective against the dominant H1N1 strains in the 2008 flu season and decreasingly effective against the dominant influenza H1N1 mutants in the US in the 2009 "Spring/Fall" pandemic. Several of the Influenza A/H5N1 mutants are genetically close to the 1918 pandemic strain. Here I provide a computational docking analysis of oseltamivir with the active site of the neuraminidase of an H5N1 strain. The computed inhibitor/receptor binding energy suggests that oseltamivir would not be effective against that strain. These results are consistent with the efficacy of oseltamivir observed in avian flu cases in humans.

Keywords: Influenza, H1N1, neuraminidase, oseltamivir

1.0 Introduction

The mortality rate in humans infected with Influenza A/H1N1 in the 1918 pandemic was ~50% ([2]). The 1918 mutant(s), unlike any genotype of H1N1 observed since, was easily transmitted among humans and killed ~10% of the world population within a single six-month period ([2]).

At present, no plausible public health regime could control an outbreak of a high-mortality-rate, highly infectious (HMR/Hi) H1N1 mutant. The scale of human interaction required to sustain food and fuel distribution to large urban areas would render quarantine ineffective ([5]). Currently, the lead time for vaccine development and production is at least as long as the duration of the 1918 pandemic. A widespread administration of effective anti-influenza therapeutics is therefore the only practical defense against a 1918-scale event after the pandemic begins.

Neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. The most widely used anti-influenza therapeutic, oseltamivir (Tamiflu™, [4]), was ineffective against the dominant H1N1 mutants in the 2008 flu season and was decreasingly effective against the dominant influenza mutant (Influenza A/H1N1) in the US in the 2009 "Spring/Fall" pandemic ([7]). Several of the Influenza A/H5N1 ("avian flu") mutants are genetically close to the 1918 pandemic strain. Avian flu in humans has not responded well to oseltamivir.

In the World Health Organization serotype-based influenza taxonomy, influenza type A has nine neuraminidase-related sero-subtypes, and these subtypes correspond at least roughly to differences in the active-site structures of the flu neuraminidases. The subtypes fall into two groups ([3]): group-1 contains the subtypes N1, N4, N5 and N8; group-2 contains the subtypes N2, N3, N6, N7 and N9.

Oseltamivir was designed to target the group-2 neuraminidases.

The available crystal structures of the group-1 N1, N4 and N8 neuraminidases ([1]) reveal that the active sites of these enzymes have a very different three-dimensional structure from that of group-2 enzymes. The differences lie in a loop of amino acids known as the "150-loop", which in the group-1 neuraminidases has a conformation that opens a cavity not present in the group-2 neuraminidases. The 150-loop contains an amino acid designated Asp 151; the side chain of this amino acid has a carboxylic acid that, in group-1 enzymes, points away from the active site as a result of the 'open' conformation of the 150-loop. The side chain of another active-site amino acid, Glu 119, also has a different conformation in group-1 enzymes compared with the group-2 neuraminidases ([8]).

The Asp 151 and Glu 119 amino-acid side chains form critical interactions with neuraminidase inhibitors. For neuraminidase subtypes with the "open conformation" 150-loop, the side chains of these amino acids might not have the precise alignment required to bind inhibitors tightly ([8]). The active site of the 1918 strain has the 150-loop configuration.

The difference in the active-site conformations of the two groups of neuraminidases may also be caused by differences in amino acids that lie outside the active site. This means that an enzyme inhibitor for one target will not necessarily have the same activity against another with the same active-site amino acids and the same overall three-dimensional structure ([17]).

2.0 Method

The general objective of this study is straightforward: to computationally assess the binding energy of the active site of a crystallized Influenza A/H5N1 neuraminidase with oseltamivir. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment.

Protein Data Bank (PDB) 2HU4 is a structural description of a crystallized neuraminidase of an H5N1 neuraminidase, bound to oseltamivir. 2HU4 consists of 8 identical chains, designated Chains A-H.

2HU4 was downloaded from PDB ([6]) on 31 January 2011. The ligand portion of 2HU4 was extracted using Microsoft *Word*. The automated docking suite *AutoDock Tools* v 4.2 (ADT, [9]) was used to perform the docking of oseltamivir to the receptor. More specifically, in ADT, approximately following the rubric documented in [12] -- Chains B-H, and the water in Chain A, of 2HU4 were deleted -- the ligand (oseltamivir) and Chain A's active-site was extracted (2HU4 identifies the active site of Chain A as 13 amides: ARG118, GLU119, ASP151, ARG152, TRP178, SER246, GLU276, GLU277, ARG292, TYR347, ARG371, and TYR406.)

-- the hydrogens, charges, and torsions in the ligand and active site were adjusted using ADT default recommendations, and finally, the ligand, assumed to be flexible wherever that assumption is physically possible, was auto-docked to the active site, assumed to be rigid, using the Lamarckian genetic algorithm implemented in ADT.

The ADT parameters for the docking are shown in Figure 1. Most values are, or are a consequence of, ADT defaults.

```

autodock_parameter_version 4.2      # used by autodock to validate parameter set
outlev 1                            # diagnostic output level
intelec                             # calculate internal electrostatics
seed pid time                        # seeds for random generator
ligand_types C HD OA N              # atoms types in ligand
fld 2HU4_receptor.maps.fld          # grid data file
map 2HU4_receptor.C.map             # atom-specific affinity map
map 2HU4_receptor.HD.map            # atom-specific affinity map
map 2HU4_receptor.OA.map            # atom-specific affinity map
map 2HU4_receptor.N.map             # atom-specific affinity map
elecmap 2HU4_receptor.e.map         # electrostatics map
desolvmap 2HU4_receptor.d.map       # desolvation map
move 2HU4_Ligand.pdbqt              # small molecule
about 0.5292 81.1637 109.1143       # small molecule center
tran0 random                        # initial coordinates/A or random
axisangle0 random                  # initial orientation
dihe0 random                       # initial dihedrals (relative) or random
tstep 2.0                           # translation step/A
qstep 50.0                          # quaternion step/deg
dstep 50.0                          # torsion step/deg
torsdof 7                           # torsional degrees of freedom
rmstol 2.0                          # cluster_tolerance/A
extnrg 1000.0                       # external grid energy
e0max 0.0 10000                     # max initial energy; max number of retries
ga_pop_size 150                     # number of individuals in population
ga_num_evals 2500000                 # maximum number of energy evaluations
ga_num_generations 27000             # maximum number of generations
ga_elitism 1                         # number of top individuals to survive to next
generation                           #
ga_mutation_rate 0.02                # rate of gene mutation
ga_crossover_rate 0.8                # rate of crossover
ga_window_size 10                    #
ga_cauchy_alpha 0.0                 # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0                  # Beta parameter Cauchy distribution
set_ga                               # set the above parameters for GA or LGA
sw_max_its 300                       # iterations of Solis & Wets local search
sw_max_succ 4                        # consecutive successes before changing rho
sw_max_fail 4                        # consecutive failures before changing rho
sw_rho 1.0                           # size of local search space to sample
sw_lb_rho 0.01                       # lower bound on rho
ls_search_freq 0.06                  # probability of performing local search on
individual                           #
set_pswl                             # set the above pseudo-Solis & Wets parameters
unbound_model bound                  # state of unbound ligand
ga_run 10                            # do this many hybrid GA-LS runs
analysis                             # perform a ranked cluster analysis

```

Figure 1. ADT parameters for the docking in this study

Interatomic distances between ligand and receptor in the computed form were compared to those in 2HU4.

3.0 Results

The interactive problem setup, which assumes familiarity with the general neuraminidase "landscape", took about 15 minutes in ADT; the docking proper, about

29 minutes on the platform described in Section 2.0. The platform's performance monitor suggested that the calculation was more or less uniformly distributed across the four processors at ~25% of peak per processor (with occasional bursts to 40% of peak), and required a constant 2.9 GB of

memory.

Figure 2 shows the oseltamivir/receptor energy and position summary produced by ADT. The estimated free energy of binding is ~ -8.5 kcal/mol; the estimated inhibition

constant, ~ 599 nanoMolar at 298 K. All distances between receptor and ligand atoms in the computed ligand position lie within 7% of the distances of the corresponding atoms in 2HU4.

LOWEST ENERGY DOCKED CONFORMATION from EACH CLUSTER

Keeping original residue number (specified in the input PDBQ file) for outputting.

```

MODEL          10
USER          Run = 10
USER          Cluster Rank = 1
USER          Number of conformations in this cluster = 10
USER
USER          RMSD from reference structure          = 1.083 A
USER
USER          Estimated Free Energy of Binding      = -8.49 kcal/mol  [(1)+(2)+(3)-(4)]
USER          Estimated Inhibition Constant, Ki    = 598.99 nM (nanomolar)  [Temperature =
298.15 K]
USER
USER          (1) Final Intermolecular Energy      = -10.58 kcal/mol
USER          vdW + Hbond + desolv Energy          = -6.25 kcal/mol
USER          Electrostatic Energy                 = -4.33 kcal/mol
USER          (2) Final Total Internal Energy      = -1.19 kcal/mol
USER          (3) Torsional Free Energy            = +2.09 kcal/mol
USER          (4) Unbound System's Energy  [(2)]   = -1.19 kcal/mol
USER
USER
USER          DPF = 2hu4.dpf
USER          NEWDPF move      2HU4_Ligand.pdbqt
USER          NEWDPF about    0.529200 81.163696 109.114304
USER          NEWDPF tran0    0.598137 80.588296 109.027331
USER          NEWDPF axisangle0  -0.942812 -0.318402 -0.098616 -12.108044
USER          NEWDPF quaternion0  -0.099435 -0.033581 -0.010401 -0.994423
USER          NEWDPF dihe0    -132.97 178.74 -163.16 -74.49 -77.91 6.34 21.37
USER
USER          x          y          z          vdW      Elec          q          RMS
ATOM          1  C2  G39  A  800      -1.828      80.459      110.166      +0.10      +0.08      +0.091      1.083
ATOM          2  C3  G39  A  800      -1.053      79.024      110.281      -0.32      +0.01      +0.050      1.083
ATOM          3  C4  G39  A  800       0.139      78.772      109.253      -0.19      -0.11      +0.209      1.083
ATOM          4  C5  G39  A  800       0.996      80.037      109.196      -0.15      -0.03      +0.143      1.083
ATOM          5  C6  G39  A  800       0.097      81.256      108.700      -0.14      +0.00      +0.147      1.083
ATOM          6  C7  G39  A  800      -1.218      81.494      109.394      -0.12      +0.03      +0.049      1.083
ATOM          7  O7  G39  A  800       0.965      82.478      108.693      -0.00      -0.13      -0.379      1.083
ATOM          8  C8  G39  A  800       1.066      83.449      107.573      -0.12      +0.04      +0.121      1.083
ATOM          9  C9  G39  A  800       0.655      82.959      106.157      -0.21      +0.00      +0.027      1.083
ATOM          10 C91 G39 A  800       1.669      82.075      105.411      -0.17      +0.00      +0.007      1.083
ATOM          11 C81 G39 A  800       0.247      84.645      108.019      -0.27      +0.02      +0.027      1.083
ATOM          12 C82 G39 A  800      -1.056      84.731      107.289      -0.48      +0.00      +0.007      1.083
ATOM          13 N5  G39  A  800       2.104      79.738      108.210      -0.06      -0.03      -0.352      1.083
ATOM          14 H5  G39  A  800       1.870      79.493      107.248      +0.08      +0.01      +0.163      1.083
ATOM          15 C10 G39 A  800       3.397      79.792      108.587      -0.27      +0.10      +0.214      1.083
ATOM          16 C11 G39 A  800       4.411      79.477      107.550      -0.29      +0.07      +0.117      1.083
ATOM          17 O10 G39 A  800       3.796      80.089      109.751      -0.60      -0.23      -0.274      1.083
ATOM          18 N4  G39  A  800       0.914      77.622      109.714      +0.05      +0.08      -0.073      1.083
ATOM          19 H42 G39 A  800       0.767      77.422      110.704      -0.41      -0.44      +0.274      1.083
ATOM          20 H41 G39 A  800       0.695      76.824      109.117      +0.04      -0.55      +0.274      1.083
ATOM          21 H43 G39 A  800       1.914      77.816      109.758      -0.29      -0.25      +0.274      1.083

```

ATOM	22	C1	G39	A	800	-3.098	80.703	110.809	-0.23	+0.34	+0.177	1.083
ATOM	23	O1B	G39	A	800	-3.839	81.683	110.469	-1.57	-1.96	-0.648	1.083
ATOM	24	O1A	G39	A	800	-3.463	79.919	111.732	-0.62	-1.38	-0.648	1.083

Figure 2. ADT's oseltamivir energy and position predictions.

Figure 3 is a rendering of the active-site/inhibitor configuration computed in this study.

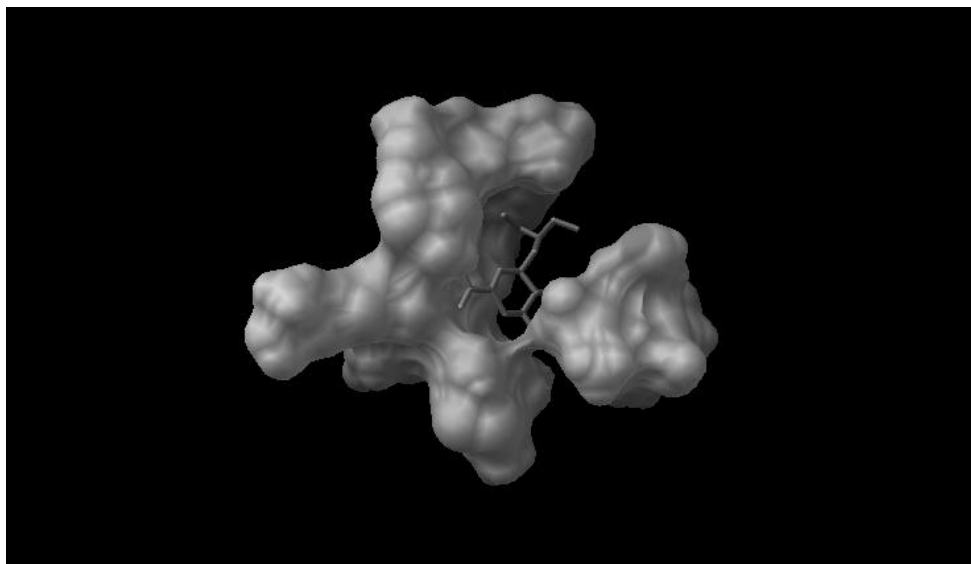


Figure 3. Rendering of oseltamivir computationally docked with the active site of Chain A of PDB 2HU4. The inhibitor is shown in stick form. Only the interior, inhibitor-containing region of the molecular surface of the active site can be compared to *in situ* data: the surface distal to the interior is a computational artifact, generated by the assumption that active site is detached from the rest of the receptor.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The inhibition constant computed in this study (~599 nanoMolar at ~298 K) is comparable to the inhibition constant of oseltamivir/neuraminidase interactions that are not clinically effective ([11], [13]). This suggests that oseltamivir would not be effective against 2HU4.

2. All distances between receptor and ligand atoms in the computed ligand position lie within 7% of the distances of the corresponding atoms in 2HU4. (For electrostatic forces, a 7% distance difference would correspond to a $(1.07^2 =)$ 14% difference in electrostatic force and potential energy. One could of course apply other statistics to the coordinate sets and provide a more comprehensive comparison of other forces/energies. Future work will address those issues.)

3. The docking study reported here assumes that the receptor is rigid. This assumption is appropriate for the binding energy computation for PDB 2HU4 per se. However, the calculation does not reflect what receptor "flexing" could contribute to the interaction of the ligand with native unliganded receptor. Future work will analyze the docking of the ligand with the native form.

4. The analysis described in Sections 2.0 and 3.0 assumes the neuraminidase is in a crystallized form. *In situ*, at physiologically normal temperatures (~310 K), the receptor is not in crystallized form. The ligand/receptor conformation *in situ*, therefore, may not be identical to their conformation in the crystallized form.

5. Minimum-energy search algorithms other than the Lamarckian genetic algorithm used in this work could be applied to this docking problem. Future work will use Monte Carlo/simulated annealing algorithms.

6. A variety of torsion and charge models could be applied to this problem, and future work will do so.

5.0 Acknowledgements

This work benefited from discussions with Tony Pawlicki. For any problems that remain, I am solely responsible.

6.0 References.

[1] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49.

[2] Johnson NP and Mueller J. Updating the accounts: global mortality of the 1918-1920 "Spanish " influenza pandemic. *Bulletin of the History of Medicine* 76 (2002), 105-115.

[3] World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization* 58 (1980), 585-591.

[4] Ward P et al. Oseltamivir (Tamiflu) and its potential for use in the event of an influenza pandemic. *Journal of Antimicrobial Chemotherapy* 55, supplement 1 (2005), i5-i21.

[5] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.

[6] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49. <http://www.pdb.org/pdb/explore/explore.do?structureId=2HU4>.

[7] US Centers for Disease Control. *Summary: Interim Recommendations for the Use of Influenza Antiviral Medications in the Setting of Oseltamivir Resistance among Circulating Influenza A (H1N1) Viruses, 2008-09 Influenza Season*. 19 December 2008. URL <http://www.cdc.gov/flu/professionals/antivirals/summary.htm>.

[8] Luo M. Structural biology: antiviral drugs fit for a purpose. *Nature* 443 (7 September 2006), 37-38. doi:10.1038/443037a, published online 6 September 2006.

[9] Morris GM, Goodsell DS, Huey R, Lindstrom W, Hart WE, Kurowski S, Halliday S, Belew R, and Olson AJ. *AutoDock* v4.2. <http://autodock.scripps.edu/>. 2010.

[10] Drug Bank. *Oseltamivir*. <http://www.drugbank.ca/drugs/DB00198>.

[11] Govorkova EA et al. Comparison of efficacies of RWJ-270201, zanamivir, and

oseltamivir against H5N1, H9N2, and other avian influenza viruses. *Antimicrobial Agents and Chemotherapy* 45 (2001), 2723-2732.

[12] Huey R and Morris GM. *Using AutoDock 4 with AutoDock Tools: A Tutorial*. 8 January 2008. <http://autodock.scripps.edu/>.

[13] Cheng Y and Prusoff WH. Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. *Biochemical Pharmacology* 22 (December 1973), 3099–3108. doi:10.1016/0006-2952(73)90196-2.

Accelerate numerical diffusion solver of 2D multi-scale and multi-resolution agent-based brain cancer model by employing graphics processing unit technology

[BIOCAMP]

Beini Jiang¹

¹Department of Mathematical Sciences
Michigan Tech University
Houghton, MI, USA
beinij@mtu.edu

Le Zhang^{1*}

¹Department of Mathematical Sciences
Michigan Tech University
Houghton, MI, USA
zhangle@mtu.edu

Wen Zhang¹

¹Department of Mathematical Sciences
Michigan Tech University
Houghton, MI, USA

Allan Struthers¹

¹Department of Mathematical Sciences
Michigan Tech University
Houghton, MI, USA

Michael E Berens²

²Cancer and Cell Biology Division
Translational Genomics Research Institute, TGen
Phoenix, AZ, USA

Xiaobo Zhou³

³Center for Bioinformatics and Department of Pathology
The Methodist Hospital
Research Institute & Weill Cornell Medical College
Houston, Texas, USA

Abstract—Diffusion model is increasingly employed to simulate diffusion of biological compounds including nutrient, oxygen and chemoattractants in the agent-based model (*ABM*). However, it takes long compute time to employ conventional numerical methods such as alternating direction implicit (*ADI*) method to approximate the exact solution of the diffusion processed by sequential computing algorithm. To overcome this limitation, our study employs cutting-edge graphics processing unit (*GPU*) technology to speed up the conventional sequential numerical solver for diffusion and incorporates our proposed parallel computing algorithms into our well developed 2D multi-scale and multi-resolution agent-based brain cancer model to break through the bottleneck of the *ABM* that it is hard to simulate the large system restricted to the limited compute resource and memory. Our simulation outputs demonstrate that *ABM* model can be used to simulate real-time actual cancer progression with relative fine grids by using *GPU* based parallel computing algorithm.

Keywords: *graphics processing unit; agent-based model; alternating direction implicit method; domain decomposition; parallel computing*

I. INTRODUCTION

Agent-based model (*ABM*) has become a popular method to describe the complex dynamic, adaptive and self-organizing cancer system. For example, Mansury and Deisboeck [1, 2] employed the *ABM* to simulate the expansion of brain tumor

in micro-macro environments. And Zhang et al. [3-6] developed multi-scale *ABMs* to model the growth of glioma and investigate incoherent relations of the tumor expansion among macroscopic environment, microscopic environment and molecular environments. A diffusion module is employed to simulate the diffusion of the chemoattractants on the macroscopic scale environment.

Though conventional finite difference numerical methods such as *ADI*, Gauss-Seidel and Jacobi methods [7-9] for diffusion module already have been used to simulate diffusion of biological compounds such as nutrients, oxygen and chemoattractants [3, 10-15] for years, they all depend on the grid size so much that a relative fine grids can better mimic the diffusion process but significantly increase the compute time. Therefore, previous studies such as the work done by Athale et al. [10, 11] and Wang et al.[16] have to employ relatively coarse grids to reduce the compute time and the work done by Dai et al.[17] and Zhang et al.[18-20] employed special numerical scheme such as preconditioned Richardson method [21, 22] to sacrifice the compute accuracy in some dimensions of coordinates to reduce the compute resource request due to the specific aim of these biomedical projects. Nonetheless, our well developed 2D multi-scale and multi-resolution *ABM* model needs such a fast diffusion module that not only can accurately model the diffusion process but also costs less compute resource. For this reason, using parallel computing algorithm to speed up the conventional numerical

solver [23, 24] is the best promising solution. Quite a few previous parallel computing algorithms employed Message Passing Interface (*MPI*) [25], a parallel computing scheme based on multiple instruction multiple data infrastructure, to parallel the sequential numerical diffusion solver. However, *MPI* is not only too expensive to be routinely used for light computing project, but also its compute speed is limited by the communication rate [26]. Since 2007, *NVIDIA* keeps releasing its graphics processing unit (*GPU*) and the novel Compute Unified Device Architecture (*CUDA*) based on single instruction multiple data infrastructure (*SIMD*). Until now, *GPU* of *NVIDIA* has been evolved into a highly parallel, multithreaded, many core processor, with dramatic compute ability and high memory bandwidth [27], especially for the recent *Fermi GPU* [28, 29]. Compared to *MPI*, *GPU* computing is more affordable, portable and suitable for the *ABM* simulation.

In general, the aim of this study is to incorporate the parallel diffusion numerical solver based on latest released *Fermi GPU* technology into our previous well developed multi-scale and multi-resolution *ABM* model [5] to resolve its compute capability shortage problem. The methods section introduces the conventional numerical scheme, alternating direction implicit (*ADI*) method [7, 30] and the *GPU* implementation [31]. And then, we show that our parallel algorithms significantly increase the performance when applied to the 2D multi-scale and multi-resolution *ABM* [5].

II. METHODS

This section gives a brief introduction to *ADI* scheme [7] with the standard domain decomposition strategy [32, 33] followed by the description of *GPU* implementations.

A. Numerical diffusion solver: *ADI* Scheme

The diffusion of the chemical cues is described by (1.a), where the D is the diffusivity for glucose ($D_G=6.7 \times 10^{-7} \text{ cm}^2 \text{ s}^{-1}$) and TGF_α ($D_T=5.18 \times 10^{-7} \text{ cm}^2 \text{ s}^{-1}$), respectively.

$$\frac{\partial u}{\partial t} = D \nabla^2 u = D \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) = D(u_{xx} + u_{yy}). \quad (1.a)$$

The Crank–Nicolson method approximates (1.a) by (1.b)

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\Delta t} = \frac{D}{2} \left(\frac{\delta_x^2}{\Delta x^2} + \frac{\delta_y^2}{\Delta y^2} \right) (u_{ij}^{n+1} + u_{ij}^n). \quad (1.b)$$

where u_{ij}^n is the numerical approximation of $u(x_i, y_j, t_n)$ and $x_i = i\Delta x, y_j = j\Delta y, t_n = n\Delta t$. δ_x and δ_y denote the central difference operators [7].

Introducing an intermediate level $u_{ij}^{n+1/2}$, the *ADI* method modifies (1.b) into two separate difference equations with implicit scheme, given by (2):

$$\frac{u_{ij}^{n+1/2} - u_{ij}^n}{\Delta t/2} = D \left(\frac{\delta_x^2}{\Delta x^2} u_{ij}^{n+1/2} + \frac{\delta_y^2}{\Delta y^2} u_{ij}^n \right). \quad (2.a)$$

$$\frac{u_{ij}^{n+1} - u_{ij}^{n+1/2}}{\Delta t/2} = D \left(\frac{\delta_x^2}{\Delta x^2} u_{ij}^{n+1/2} + \frac{\delta_y^2}{\Delta y^2} u_{ij}^{n+1} \right). \quad (2.b)$$

Writing $\mu_x = D \frac{\Delta t}{\Delta x^2}$ and $\mu_y = D \frac{\Delta t}{\Delta y^2}$ reduces (2) into the Peaceman-Rachford *ADI* scheme [7], shown as (3)

$$-\frac{\mu_x}{2} u_{i-1,j}^{n+1/2} + (1 + \mu_x) u_{ij}^{n+1/2} - \frac{\mu_x}{2} u_{i+1,j}^{n+1/2} = \frac{\mu_y}{2} u_{i,j-1}^n + (1 - \mu_y) u_{ij}^n + \frac{\mu_y}{2} u_{i,j+1}^n. \quad (3.a)$$

$$-\frac{\mu_y}{2} u_{i,j-1}^{n+1} + (1 + \mu_y) u_{ij}^{n+1} - \frac{\mu_y}{2} u_{i,j+1}^{n+1} = \frac{\mu_x}{2} u_{i-1,j}^{n+1/2} + (1 - \mu_x) u_{ij}^{n+1/2} + \frac{\mu_x}{2} u_{i+1,j}^{n+1/2}. \quad (3.b)$$

The right part of both equations of (3) is explicit formula and easily parallelized, while the left part is a symmetric and tridiagonal system of equations $Ax = b$ to be solved with the Thomas algorithm [7, 34].

Equation (3) could be written into a general form as (4.a) with $x_0 = 0$ and $x_{N+1} = 0$.

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, i = 1, 2, \dots, N. \quad (4.a)$$

The corresponding matrix form of this tridiagonal system is represented by (4.b)

$$\begin{bmatrix} b_1 & c_1 & 0 & \cdots & \cdots & 0 \\ a_2 & b_2 & c_2 & \ddots & & \\ 0 & a_3 & b_3 & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & & \\ \vdots & & & & & 0 \\ 0 & \cdots & \cdots & 0 & a_N & b_N \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ \vdots \\ d_N \end{bmatrix}. \quad (4.b)$$

B. Thomas Algorithm

The Thomas algorithm is employed to solve (4.a). It has two major steps. First is computing coefficients β_k (5.a) and ν_k (5.b) known as forward sweep. Second is using backward substitution to get solutions as (5.c).

$$\beta_k = \begin{cases} \frac{c_1}{b_1}; & k = 1 \\ \frac{c_k}{b_k - \beta_{k-1} a_k}; & k = 2, 3, \dots, N-1 \end{cases}. \quad (5.a)$$

$$\nu_k = \begin{cases} \frac{d_1}{b_1}; & k = 1 \\ \frac{d_k - \nu_{k-1} a_k}{b_k - \beta_{k-1} a_k}; & k = 2, 3, \dots, N \end{cases}. \quad (5.b)$$

$$\begin{cases} x_N = \nu_N \\ x_k = \nu_k - \beta_k x_{k+1}; & k = N-1, N-2, \dots, 1 \end{cases}. \quad (5.c)$$

The details of the deduction of (5) are described in Morton's book [7].

C. Domain Decomposition

For the boundary value problem on a large domain, the domain decomposition method decomposes the problem into smaller independent boundary value problems on smaller subdomains and then employ iterative method to resolve differences between the solutions on adjacent subdomains [32, 33]. We develop such a *GPU* based parallel computing

algorithm with classical alternating Schwarz method [33, 35] that can benefit from the advantages of *GPU* technology. Fig. 1(b) [31] exhibits the decomposition of a 10 by 10 array with an 8 by 8 inner array (green) and four vectors of boundary points (red) (Fig. 1(a) [31]) into 4 overlapping 6 by 6 sub-arrays, each of which consists of a 4 by 4 inner array (green) and four artificial internal boundaries (red). Each sub-array is iteratively solved to make the artificial boundaries converge [7, 32, 33, 35, 36]. Here, we use the data transfer between sub-matrix 1 and sub-matrix 2 as an example. The values of the four inner elements on the rightmost side in sub-matrix 1 are sent to sub-matrix 2 as the new left artificial boundary as well as the values of the four inner elements on the leftmost side in sub-matrix 2 are sent to sub-matrix 1 as the new right artificial boundary until both artificial boundaries converge.

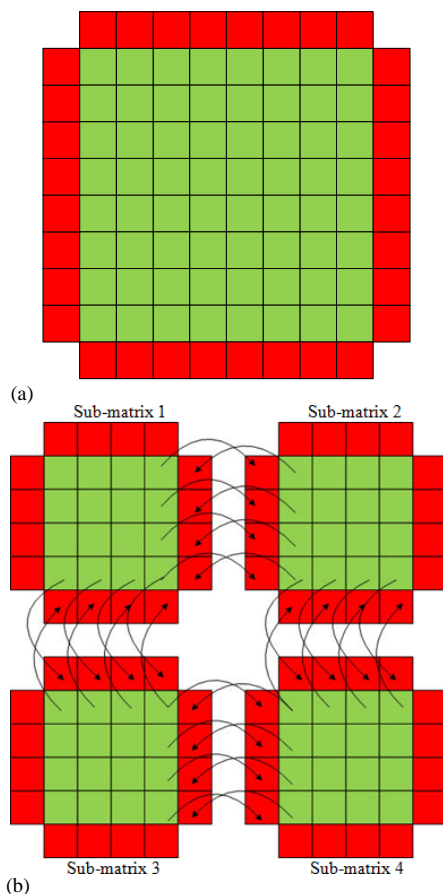


Figure 1 [31] (a) A 10 by 10 solution matrix with red to indicate boundary elements and blue to indicate inner elements and (b) Decomposition of a 10 by 10 array into 4 overlapping 6 by 6 sub-arrays with red to indicate boundary elements, green to indicate inner elements and the arrows to show how to update the boundary data.

D. Parallel Computing Algorithms to Speed up the diffusion solver

The first step of *ADI* is to set up the explicit scheme, shown as the right part of (3). Since each element could be computed independently, the explicit scheme is easy to be parallelized

with single-instruction, multiple-thread (*SIMT*) infrastructure of *CUDA*. The second step is to solve the implicit scheme of *ADI* by Thomas algorithm. As we discussed in our previous research [31], Thomas algorithm is the bottleneck to speed up the conventional numerical diffusion solver.

CUDA programming has two major steps. The first is preparing such data that can be paralleled in the host side (*CPU*). The second is processing these data in the device side (*GPU*) by kernel. *CUDA* organizes the threads into a two-level hierarchy (Fig. 2-1 of *NVIDIA CUDA Programming Guide* [27]). As shown by Fig. 2-2 of *NVIDIA CUDA Programming Guide* [27], a thread executing on the device has access to the device's (*GPU*) *DRAM* and on-chip memory through 6 different memory spaces such as registers, local memory, shared memory, global memory, constant memory, and texture memory [27, 37-40]. As a very important memory of *GPU*, global memory is in charge of exchanging the data between the host (*CPU*) and the device (*GPU*). Moreover, it plays such a role that passes the messages between the threads from different blocks, since current *GPU* infrastructure prohibits the communication of threads from different blocks [27, 29, 41]. However, as an off-chip memory, the latency of global memory is very high. As on-chip memory, shared memory, registers, and constant-memory caches are much faster with much lower latency. Nonetheless, shared memory is very limited and it is only allocated to each block. For example, the capacity of the latest version of *GPU* (Fermi) is only 64KB [28, 42]. Moreover, another on-chip memory, constant memory, is disallowed to be written to during the computation [27, 43] though it is cached.

CUDA uses a new architecture called *SIMT* to manage threads running different programs. The multiprocessor *SIMT* unit creates, manages, schedules, and executes threads in groups of 32 parallel threads we call warps [27]. We have developed three parallel computing algorithms to accelerate the numerical diffusion solver based on the new features of *GPU* technology [31]. The first is parallel computing algorithm with global memory (*PGM*), which employs only global memory to carry out parallel computing. The second is parallel computing algorithm with shared memory, global memory and *CPU* synchronization [27, 29, 41, 44] (*PSGMC*) and the third is parallel computing algorithm using shared memory, global memory and *GPU* synchronization [29, 41, 45] (*PSGMG*). *PSGMC* and *PSGMG* employ "tiles" strategy to partition the data and take advantages of both global memory and shared memory with the classical alternating Schwarz domain decomposition method [7, 32, 33, 35, 36]. The details of these three implementation methods are presented in our recent publication [31]. Here, we incorporate our fastest parallel diffusion solver into 2D multi-scale and multi-resolution *ABM* [5] to speed up the computation of *ABM*.

III. RESULTS

Our source code is implemented by C [46, 47] and *NVCC* [48] programming language and running on the recent Fermi

GPU card (GeForce GTX 480) [42, 49, 50] with CUDA standard.

In the beginning, let us briefly show how to use parallel computing algorithms [31] based on GPU technology to accelerate the numerical diffusion solver as following.

First, we employ PGM to compute the diffusion on the lattice with different number of grid points and compare the computing time with the sequential computing. Fig. 2 shows PGM computing time is not always faster than sequential algorithm for the lattice with small point number but dramatically faster than sequential algorithm for the lattice with large point number [31].

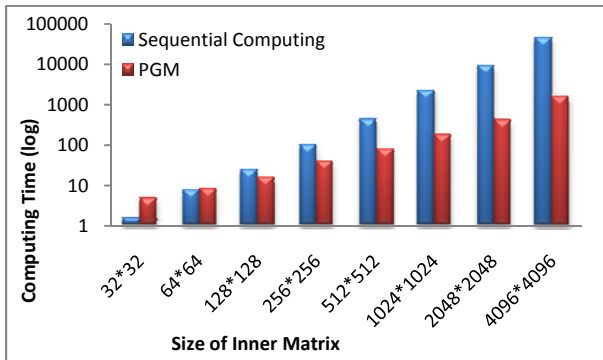


Figure 2 [31]. Computing time of PGM and sequential computing by logarithmic scale. The x axis represents the inner matrix size (number of inner grid points) and y axis represents the computing time (logarithmic scale with base 10) in millisecond. The blue bar represents the computing time of sequential computing and the red bar represents the computing time of PGM.

Second, we compare the compute time between PSGMC and PGM, when simulating the diffusion on a 4098 by 4098 lattice. Fig. 3 shows PSGMC improves the performance by 58% compared with PGM [31].

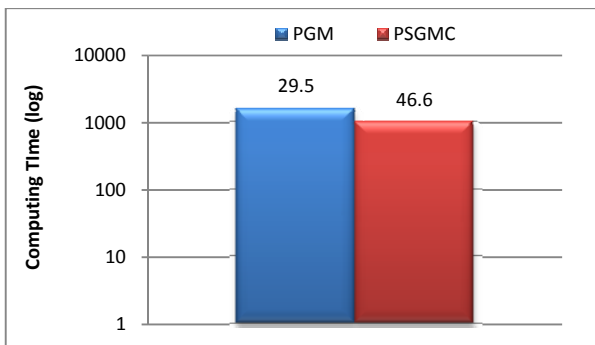


Figure 3 [31]. Computing time of PSGMC and PGM by logarithmic scale. The y axis represents the computing time (logarithmic scale with base 10) in millisecond. The blue bar represents the computing time of PGM and the red bar represents the optimal computing time of PSGMC. The number on each bar indicates the multiple of acceleration to the sequential computing.

Third, we compare the performance of PSGMC and PSGMG. Fig. 4 exhibits PSGMG improves the performance by 11% compared with PSGMC, when processing the diffusion on a 4098 by 4098 lattice [31].

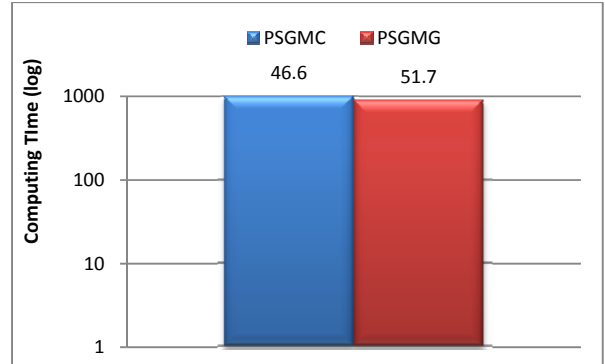


Figure 4 [31]. Computing time of PSGMG and PSGMC by logarithmic scale. The y axis represents the computing time (logarithmic scale with base 10) in millisecond. The blue bar represents the computing time of PSGMC and the red bar represents the computing time of PSGMG. The number on each bar indicates the multiple of acceleration to the sequential computing.

Next, we incorporate the fastest parallel computing method (PSGMG) into the well developed multi-scale and multi-resolution ABM model [5]. The multi-resolution model is designed based on two different resolution lattices, namely low-resolution lattice and high-resolution lattice. The low-resolution lattice is set up with a grid size of about $62.5 \mu m$, on each grid point of which, a 6 by 6 high-resolution lattice with a grid size of approximately $10 \mu m$ is superimposed, described by Fig. 5 [5]. To demonstrate the advantages of the parallel computing algorithm, we scale up the lattice size of the previous multi-scale and multi-resolution ABM model [5]. Current low-resolution lattice is changed from 100 by 100 to 683 by 683 and high-resolution lattice is upgraded from 600 by 600 to 4098 by 4098.

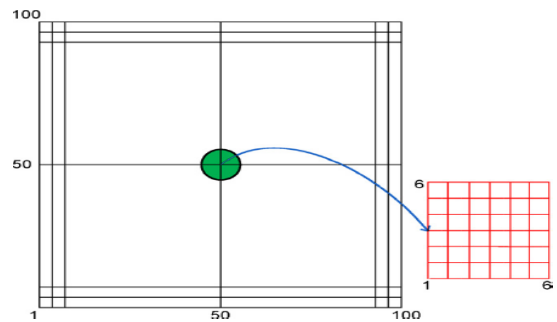


Figure 5 [5] Configuration of multi-resolution lattice.

The diffusion of the chemical cues is observed on the high-resolution lattice, with a grid size of approximately $10 \mu m$, namely both Δx and Δy in the ADI scheme (2) are equal to $10 \mu m$. Δt is set to 1s to make $\max\{\mu_x, \mu_y\} \leq 1$ regarding to the maximum principle [7], thus the ADI scheme needs to be computed 3600 times for each time step, which is equivalent to $1h$.

And then, Fig. 6 exhibits that parallel computing can significantly increase the performance of the compute time 37.5 folders than sequential computing for multi-scale and multi-resolution ABM model [5].

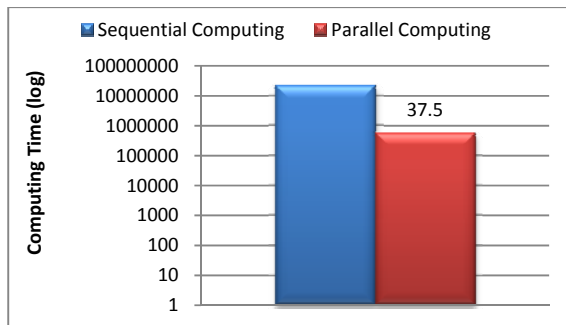


Figure 6. Computing time of parallel and sequential computing by logarithmic scale. The y axis represents the computing time (logarithmic scale with base 10) in millisecond. The blue bar represents the computing time of sequential computing and the red bar represents the optimal computing time of parallel computing. The number on the red bar indicates the multiple of acceleration to the sequential computing.

IV. CONCLUSIONS

This study demonstrates that it is possible to simulate the real-time actual tumor progression in a 2D lattice with relative fine grids by using *GPU* based parallel computing algorithms. Our extension research will develop a *GPU* based parallel *ODE* solver to speed up the molecular pathway module of our well developed multi-scale and multi-resolution agent-based model [5].

References

- [1] Y. Mansury, M. Kimura, J. Lobo, and T. S. Deisboeck, "Emerging patterns in tumor systems: simulating the dynamics of multicellular clusters with an agent-based spatial agglomeration model," *J Theor Biol* vol. 219, pp. 343-370 2002.
- [2] Y. Mansury and T. S. Deisboeck, "The impact of "search precision" in an agent-based tumor model," *J Theor Biol* vol. 224, pp. 325-337, 2003.
- [3] L. Zhang, C. A. Athale, and T. S. Deisboeck, "Development of a three-dimensional multiscale agent-based tumor model: simulating gene-protein interaction profiles, cell phenotypes and multicellular patterns in brain cancer," *J Theor Biol*, vol. 244, pp. 96-107, Jan 7 2007.
- [4] L. Zhang, Z. Wang, J. A. Sagotsky, and T. S. Deisboeck, "Multiscale agent-based cancer modeling," *J Math Biol*, vol. 58, pp. 545-59, Apr 2009.
- [5] L. Zhang, L. L. Chen, and T. S. Deisboeck, "Multi-scale, multi-resolution brain cancer modeling," *Math Comput Simul*, vol. 79, pp. 2021-2035, Mar 2009.
- [6] L. Zhang, C. Strouthos, Z. Wang, and T. S. Deisboeck, "Simulating brain tumor heterogeneity with a multiscale agent-based model: Linking molecular signatures, phenotypes and expansion rate," *Mathematical and Computer Modelling*, vol. 49, pp. 307-319, 2009.
- [7] k. Q. Morton and D. F. Mayers, *Numerical solution of partial differential equations*, 2nd ed. New York: Cambridge University Press, 2008.
- [8] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, 2nd ed. Philadelphia, PA: SIAM: Society for Industrial and Applied Mathematics, 2004.
- [9] R. L. Burden and J. D. Faires, *Numerical analysis*, 8th ed. Belmont, CA: Thomson Higher Education, 2008.
- [10] C. Athale, Y. Mansury, and T. S. Deisboeck, "Simulating the impact of a molecular 'decision-process' on cellular phenotype and multicellular patterns in brain tumors," *J Theor Biol*, vol. 233, pp. 469-81, Apr 21 2005.
- [11] C. A. Athale and T. S. Deisboeck, "The effects of EGF-receptor density on multiscale tumor growth patterns," *J Theor Biol*, vol. 238, pp. 771-9, Feb 21 2006.
- [12] K. R. Swanson, E. C. Alvord, Jr., and J. D. Murray, "A quantitative model for differential motility of gliomas in grey and white matter," *Cell Prolif*, vol. 33, pp. 317-29, Oct 2000.
- [13] K. R. Swanson, E. C. Alvord, Jr., and J. D. Murray, "Virtual brain tumours (gliomas) enhance the reality of medical imaging and highlight inadequacies of current therapy," *Br J Cancer*, vol. 86, pp. 14-8, Jan 7 2002.
- [14] K. R. Swanson, C. Bridge, J. D. Murray, and E. C. Alvord, Jr., "Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion," *J Neurol Sci*, vol. 216, pp. 1-10, Dec 15 2003.
- [15] K. R. Swanson, R. C. Rostomily, and E. C. Alvord, Jr., "A mathematical modelling tool for predicting survival of individual patients following resection of glioblastoma: a proof of principle," *Br J Cancer*, vol. 98, pp. 113-9, Jan 15 2008.
- [16] A. X. Cong, H. O. Shen, W. X. Cong, and G. Wang, "Improving the Accuracy of the Diffusion Model in Highly Absorbing Media," *International Journal of Biomedical Imaging*, vol. 2007, 2007.
- [17] W. Dai, A. Bejan, X. Tang, L. Zhang, and R. Nassar, "Optimal temperature distribution in a three dimensional triple-layered skin structure with embedded vasculature," *Journal of Applied Physics*, vol. 99, 2006.
- [18] L. Zhang, W. Dai, and R. Nassar, "A Numerical Method for Optimizing Laser Power in the Irradiation of a 3-D Triple-Layered Cylindrical Skin Structure," *Numerical Heat Transfer*, vol. 48, pp. 21 - 41, 2005.
- [19] L. Zhang, W. Dai, and R. Nassar, "A Numerical Method for Obtaining an Optimal Temperature Distribution in a 3-D Triple-Layered Cylindrical Skin Structure Embedded with a Blood Vessel " *Numerical Heat Transfer*, vol. 49, pp. 765 - 784, 2006.
- [20] L. Zhang, W. Dai, and R. Nassar, "A numerical algorithm for obtaining an optimal temperature distribution in a 3D triple-layered cylindrical skin structure," *Computer Assisted Mechanics and Engineering Sciences*, vol. 14, pp. 107-125, 2007a.
- [21] B. Bialecki, "Preconditioned Richardson and Minimal Residual Iterative Methods for Piecewise Hermite Bicubic Orthogonal Spline Collocation Equations," *Siam Journal on Scientific Computing*, vol. 15, pp. 668-680, May 1994.
- [22] W. H. Dai and R. Nassar, "A preconditioned Richardson method for solving three-dimensional thin film problems with first order derivatives and variable coefficients," *International Journal of Numerical Methods for Heat & Fluid Flow*, vol. 10, pp. 477-487, 2000.
- [23] B. Barney, "Introduction to Parallel Computing," 2010.
- [24] K. Asanovic, R. Bodik, B. C. Catanzaro, and J. J. Gebis, "The Landscape of Parallel Computing Research:A View from Berkeley," 2006.
- [25] Y. Aoyama and J. Nakano, "RS/6000 SP: Practical MPI Programming," IBM, 1999.
- [26] C. Rosul, "Message Passing Interface (MPI) Advantages and Disadvantages for applicability in the NoC Environment," 2005.
- [27] NVIDIA, "NVIDIA CUDA Programming Guide," NVIDIA, 2009a.
- [28] NVIDIA, "NVIDIA's Next Generation CUDA Compute Architecture: Fermi": NVIDIA, 2009b.
- [29] W. C. Feng and S. C. Xiao, "To GPU Synchronize or Not GPU Synchronize?," in *International Symposium on Circuits and Systems* Paris, France, 2010.
- [30] R. McOwen, *Partial Differential Equations: Methods and Applications*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 2002.
- [31] B. Jiang, A. Struthers, L. Zhang, Z. Sun, Z. Feng, X. Zhao, W. Dai, K. Zhao, X. Zhou, and M. Berens, "Employing graphics processing unit technology, alternating direction implicit method and domain decomposition to speed up the numerical diffusion solver for the biomedical engineering research," *International Journal for Numerical Methods in Biomedical Engineering*, vol. (in press), 2011.

- [32] B. Smith, P. Biqrstad, and W. Gropp, *Domain Decomposition: Parallel multilevel methods for elliptic partial differential equation*, 1st ed. New York: Cambridge University Press, 2004.
- [33] A. St-Cyr, M. J. Gander, and S. J. Thomas, "Optimized Restricted Additive Schwarz Methods," in *16th International Conference on Domain Decomposition Methods*, New York 2005.
- [34] W. Dai, "A Parallel Algorithm for Direct Solution of Large Scale Five-Diagonal Linear Systems," in *Proceedings of the Seventh SIAM Conference on Parallel Processing for Scientific Computing*, San Francisco, CA, 1995, p. 875.
- [35] X. C. Cai and M. Sarkis, "A restricted additive Schwarz preconditioner for general sparse linear systems," *Siam Journal on Scientific Computing*, vol. 21, pp. 792-797, Oct 26 1999.
- [36] J. P. Zhu, *Solving Partial Differential Equations On Parallel Computers*. London: World Scientific Publishing Co. Pte. Ltd., 1994.
- [37] V. Volkov and J. Demmel, "Benchmarking GPUs to Tune Dense Linear Algebra," in *Conference on High Performance Networking and Computing archive Proceedings of the 2008 ACM/IEEE conference on Supercomputing* Austin, TX: IEEE Press Piscataway, NJ, USA 2008.
- [38] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable Parallel Programming with CUDA," in *ACM Queue*. vol. , 2008, pp. 42-53.
- [39] M. Guevara, C. Gregg, K. hazelwood, and K. Skadron, "Enabling Task parallelism in the CUDA Scheduler," in *Proceedings of the Workshop on Programming Models for Emerging Architectures (PMEA)* Raleigh, NC, 2009.
- [40] S. Che, M. Boyer, J. Y. Meng, D. Tarjan, J. W. Sheaffer, and K. Skadron, "A performance study of general-purpose applications on graphics processors using CUDA," *Journal of Parallel and Distributed Computing*, vol. 68, pp. 1370-1380, Oct 2008.
- [41] S. C. Xiao, A. M. Aji, and W. C. Feng, "On the Robust Mapping of Dynamic Programming onto a Graphics Processing Unit," in *International Conference on Parallel and Distributed Systems* Shenzhen, China, 2009.
- [42] NVIDIA, "Tuning CUDA Applications for Fermi," NVIDIA, 2010.
- [43] D. Kirk and W. M. Hwu, *Programming Massively Parallel Processors*, 1st ed. Burlington, MA: Morgan Kaufmann, 2010.
- [44] M. Boyer, M. Sarkis, and W. Weimer, "Automated Dynamic Analysis of CUDA Programs," in *Third Workshop on Software Tools for MultiCore Systems in conjunction with the IEEE/ACM International Symposium on Code Generation and Optimization (CGO)* Boston, MA: , 2008.
- [45] S. C. Xiao and W. C. Feng, "Inter-Block GPU Communication via Fast Barrier Synchronization," in *In Proc. of the IEEE International Parallel and Distributed Processing Symposium* Atlanta, GA 2010.
- [46] B. W. Kernighan and D. M. Ritchie, *The C Programming Language*, 2nd ed. Englewood Cliffs, New Jersey: Prentice Hall, 1988.
- [47] S. G. Kochan, *Programming in C*, 3rd ed. Indianapolis, Indiana: Sams, 2004.
- [48] NVIDIA, "The CUDA Compiler Driver NVCC," NVIDIA, 2007.
- [49] P. N. Glaskowsky, "NVIDIA's Fermi: The First Complete GPU Computing Architecture " 2009.
- [50] T. R. Halfhill, "Looking Beyond Graphics," 2009.

EpiGraph: A Scalable Simulation Tool for Epidemiological Studies

Gonzalo Martín, Maria-Cristina Marinescu, David E. Singh and Jesús Carretero

Computer Science Department
Carlos III University of Madrid
28911 Leganés, Spain

Abstract—*This paper presents a novel approach to modeling the propagation of the flu virus throughout a realistic interconnection network based on actual individual interactions which we extract from social networks. We allow the individual interconnections to change during the propagation by making them time-dependent. We have implemented a scalable, fully distributed simulator and we validated the epidemic model by comparing the simulation results against those of another epidemic simulator, with similar prediction values and better performance. We then performed an extensive analysis of the effects of the new features of our approach on the results of the simulations.*

Keywords: simulation, epidemiology, social networks, distributed algorithms

1. Introduction

Modeling the evolution of an epidemics involves both modeling the specific infectious agent as well as the actual social structure of the population under study. The purpose of the work we present in this paper is to accurately model the evolution of an epidemics in specific populations over a short to medium time span. Using an actual social model as input for the epidemic model promises more accurate results then either using probability distributions or synthetically generating the interaction graphs. Our approach approximates an actual social model by a realistic model based on real demographic information and actual individual interactions extracted from social networks. To the extent of our knowledge ours is the first attempt to model the connections within a population at the level of an individual based on information extracted from virtual social networks such as Enron or Facebook. Additionally, we allow modeling the characteristics of each individual as well as customizing his daily interaction patterns based on the time and the day.

We implemented EpiGraph, a simulator which takes as inputs the social model and an epidemic model specific to the influenza virus. The implementation is distributed and fully parallel; this allows simulating large populations of the order of millions of individuals in execution times of the order of minutes. We compared the results for our simulations in terms of the effects of the epidemics with the results obtained by InFluSim in [1]. We show that the simulators predict similar results. We further perform an extensive study of the effects of the features specific to our approach on the disease

propagation. For instance, we study how different social models affect the disease propagation and we investigate the effects of introducing different vaccine or quarantine programs at different stages of the epidemic.

Our contributions: The specific contributions of this work are the following: (1) We use real demographic data to model group types with different characteristics. We leverage data extracted from social networks to model the interaction patterns between individuals pertaining to the same social group; (2) We allow modeling individual characteristics such as profession, age, gender, etc. We also allow customizing individual behavior based on the time of day for every type of interaction between individuals; (3) We implement a scalable, fully distributed simulator and we evaluate its performance on two platforms; (4) We validate the results of the simulation against another epidemic simulator. We additionally perform an extensive analysis of the effects of the features specific to our approach on the results of the simulations.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 describes the modeling task and the simulation algorithm. Section 4 presents a study of the performance and simulation results of EpiGraph. Section 5 summarizes the paper with the conclusions and some directions for future work.

2. Related work

Interconnection networks: The majority of human-transmitted infectious diseases use physical contact as the main transmission mean. For this reason the dynamics of the propagation is tightly related to the structure and the characteristics of the network of connections between the individuals within a population [2], [3], [4], [5], [6]. Typically epidemiological models are compartmental in the sense that they model the dynamics of the epidemics by nonlinear differential equations and do not model the topology of the contact network. The assumption is that individuals in a population are homogeneously connected, which means that all individuals have the same probability of infecting other individuals [5]. In reality each person has specific, possibly very different, interaction patterns. This makes the interconnection network be heterogeneous [7], [5]. Additionally, there tend to be few people who have many connections, some strong but most of them weak—these are

the “core groups”—while most of the individuals have few connections [8], [9].

The typical way to approximate a heterogeneous contact network is to build a contact graph in which the individuals are nodes and edges represent connections [10], [11], [12]. A straightforward model implements the graph as an adjacency matrix. We use a more sophisticated model in which each matrix cell holds a value that represents the type of social interconnection: study, work, leisure, or family. The patterns of interactions depend on whether they occur between individuals within the same group or from different groups. We additionally allow the type of interconnection to change depending on a time parameter to reflect the fact that we may interact with individuals from different group types at different times during the day. This approach allows to more accurately model the heterogeneity of the actual contact network.

Work such as HPCgen and EpiGrass [13], [14] take the approach of modeling actual populations; FastGen and CL-model [15], [16] choose instead to generate a random adjacency matrix. HPCgen uses actual demographic data from census data and interviews, and introduces the idea of generating the contact network based on social structures with arbitrary degree distributions following a Poisson distribution. To work well HPCgen requires a very high accuracy when modeling the social contacts for a specific population. The contact network is fully static in the sense that the interconnections between individuals cannot change during simulation. Experiments have shown that such a model is accurate in the case that the propagation rate of the infection is high relative to the rate with which the interconnections may change in the network [17], but would break down otherwise.

[18] presents a large-scale simulator based on a stochastic model for influenza. It uses a molecular dynamic algorithm for modeling the interactions between individuals. Their approach is computationally expensive, requiring extended simulation times and a large number of processors to complete. In contrast, EpiGraph has lower computational requirements and can simulate single individuals with specific characteristics and dynamically evolving interactions.

A different approach is followed by BioWar [19]. BioWar is a multiagent network model for simulating the effects of epidemic outbreaks due to bioterrorism attacks. It takes into account several input models such as disease, geography, weather, attack and communication technology, also it models the population behavior distributed in social group types with real census data. InflaSim [1] extends the SEIR epidemic model. It uses demographic information from real census data and it models the social structure based on different age groups. InflaSim uses differential equations to model the transmission of the disease and does not take into account time-dependent individual interactions, such as EpiGraph does.

Epidemic models: The typical mathematical model for simulating epidemics is the SIR model [20]. The SIR model is usually appropriate for infectious diseases which confer immunity to recovered individuals and it works best if demographic effects may be neglected. Our work focuses on the propagation of the influenza virus over short to medium time spans. Work in [21] extends the mathematical model with latent, asymptomatic, and dead states, as well as the possibility of introducing a vaccine program. The latent state corresponds to the incubation state in which an individual is infected but has not yet developed symptoms. A relatively small percent of the population will never develop them, passing into an asymptomatic state. All asymptomatic individuals, together with a high percentage of infected individuals recover and become immune. The rest of them pass to the dead state. EpiGraph builds on this model and extends it to introduce a new hospitalized state.

[22] proposes a more detailed model for the dissemination of the influenza virus. In their approach the susceptible cases first go to a latent stage that is non-infective. This can transition either to an asymptomatic stage which leads to removal, or to a second latent stage with some contagion degree, followed by two contagious stages with different contagion degrees. Treatment is applied only during the first infective stage.

3. The modeling task

EpiGraph consists of two main components: (1) a model for the population under study with the patterns of contact between individuals within this population, and (2) a model of how the participating agents spread the disease. This work focuses on the dissemination of the flu virus over a short to medium length time span. Our goal is to facilitate the understanding and prediction of how the virus spreads within specific populations with possibly dramatically different interaction patterns over short and medium time spans. We do *not* focus on extended time periods during which qualitatively different parameters may make a difference. For instance, in our model there is no entry into or departure from the population, except possibly through death from the disease. This is a reasonable hypothesis in case of short to medium time spans. On the other hand we are modeling interaction features that may have a large impact in the case of a single epidemic outbreak but whose effects level out over time. Generally diseases transmitted by viral agents confer immunity so the assumption is that if an infected individual recovers he will acquire immunity for a time period at least as extended as the simulation time for the infection.

In the social model each graph node models a single individual and may have specific characteristics such as gender, age, role, as so on. Each graph edge represents an interaction between two individuals and depends on the time of the day. That is, EpiGraph can capture heterogeneity

features at the level of both the individual and each of his interactions.

The social model is based on two data sources: actual demographic information, as well as a realistic model of social interactions. These are used to build graphs for both intra- and inter-group interactions. A group is a collection of individuals of the same group type as extracted from the demographic information. The complete graph is then used as an input for the epidemic model. This model captures the characteristics that are important in the process of spreading a contagious agent, is specific to the agent under study, and needs to make assumptions such as what is the subset of susceptible individuals that an infected individual may pass the agent to. Rather than assuming a distribution or generating synthetic interaction graphs, we use real information from social networks to model the social interaction patterns. The interaction network is built statically to reflect the existence of communication between individuals but abstracts away the timing for these interactions. To recover the dynamic nature of these interactions we introduce a time component depending on which an individual may interact with any number of other individuals following his own patterns.

3.1 Modeling the population

To most faithfully simulate the effects of an infectious agent spreading through a specific population we decided to use real instead of synthetic data. We use real demographic information obtained from the Primary Metropolitan Statistical Area of Philadelphia [23] to determine the distribution of the population in group types; these typically show different patterns in terms of social interactions. The group types which we extracted from the census and which we are modeling are the following: (1) school-age children and students, (2) workers, (3) stay-home parents, and (4) retired individuals. The population is split into many groups of each of these types—a structure which reflects the way individuals tend to associate with each other in terms of social contacts. Each individual has contacts within his own group as well as with individuals from other groups. Let's take the example of a worker. She's going to interact frequently with people from the same work group during work hours, with friends during leisure hours, and with family during evening/night hours. We therefore model three kinds of interactions: (1) between individuals of the same group, (2) between individuals of different groups, and (3) between members of the same family. Each of these kinds of interactions is assigned to a specific daily time frame depending on the schedule for the main activity—work, study, etc—, for leisure activities, and for family time. This makes the simulation more realistic, particularly over short time periods.

Intra-group connections: Which specific group an individual belongs to determines the actual number and patterns of interactions with other individuals from his own group. One of the contributions of our work is that we model

intra-group communications by scaling down real interaction graphs extracted from the Social Networks (SN) of Enron and Facebook. The idea is to exploit the connectivity that exists in real business and leisure SNs. The graph extracted from the Enron email database consists of 70,578 nodes and 312,620 connections, while Facebook has 250,000 nodes and 3,239,137 connections. We use Enron's SN to model the worker and retired groups and Facebook's to create the school and stay-home groups. Note that the SNs are bigger than the generated groups. We scale each down by selecting as many random entries of the SN as group members, then connecting the nodes following the same patterns as those in the SN. The selection of random entries of the SN allows us to create different structures for each group. This approach is more realistic than either synthetically generating the interaction graphs or using discrete probability distributions to approximate the number of individual interactions.

Inter-group and family connections: We create a number of intergroup contacts per individual based on the group characteristics which the individual belongs to. Mostly the inter-group contacts occur in the hours between finishing one's main daily activity—such as work or study—and going home in the evening, or during weekends. These reflect daily activities which occur in public places such as parks, gyms, public transport, coffee shops, where one generally interacts with unknown people or friends pertaining to a different group. The connections of inter-group contacts are generated at the level of the group based on a set of percentages which reflect the degree to which groups of specific types are connected. There are two types of connections between pairs of groups: strong and weak. Probabilistic parameters decide whether two groups are strongly connected, weakly connected, or are not connected at all. In addition to intra- and inter-group contacts we also model a different type of social interaction: the contacts one has with members of his family. These may be pertaining to the same or to a different group and one has contacts with them from late night to morning, and during the weekends.

Strong vs. weak ties: Interactions between groups may be either strong or weak. This reflects the degree to which the connection may serve as a channel for spreading the infectious agent. Strongly coupled groups tend to be the ones who spend many hours in contact, either for affinity, family, or work related reasons. On the other hand, weak connections are between groups that only share a few contacts. It reflects occasional or casual contacts between individuals.

Data structures: EpiGraph models interactions between individuals via a graph. To represent it we are using sparse matrices in Compressed Sparse Column format which enables both optimized matrix operations and an efficient way to distribute and access the matrices in parallel.

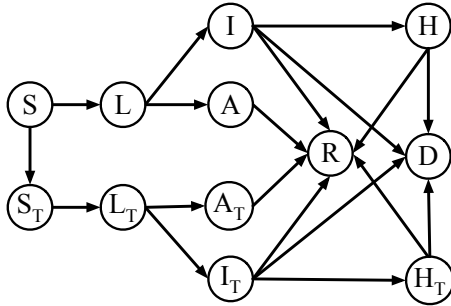


Fig. 1: State diagram for the epidemic model.

3.2 Modeling the infectious agent

The basic epidemic model is based on the principles of the SIR model as it is described in [20] and extended for the case of the flu virus by [21]. The extended model consists of a set of additional states—latent, asymptomatic, and dead—which reflect real possible stages during the development of the infection within a host. We further enhance the model with a hospitalized state in which an individual's contacts are severed. Having such a state is important when simulating realistic cases where hospitalization may be needed in order to curb the effects of the epidemics.

Figure 1 consists of two symmetrical subgraphs; the upper part has states with non-subscripted names, the lower part consists of subscripted states. Let's focus on the non-subscripted subset of the states for the time being. A susceptible individual in state S may be infected by another individual and pass to the latent—or incubating—state L . From here he normally goes to the infective state I , but may also become asymptomatic and go to state A . Individuals which are asymptomatic will always recover and go to state R ; infective individuals may recover, get hospitalized, or die. A hospitalized individual in state H either recovers or dies. In the case of the flu virus we assume that recovery implies immunity over short and medium time spans such that a recovered individual will not get infected again during the time of the simulation.

The epidemic model for influenza has many parameters, some of the most important being the basic reproduction number R_0 (average number of secondary cases of infections which produces an infected individual), the time an individual spends in each of the states, the probability that an individual will take a transition from a source state into each of the target states, and so on. The time each individual spends in a given state is generated following a Gaussian distribution to faithfully simulate the time ranges which are specific to the stages of a flu infection. The probability of infecting another individual while incubating depends on whether the specific connection is low or high risk. A high risk interaction reflects a contact between individuals which has high probability to transmit the infection. For instance, these may be interactions between members of the same work team in a company or between friends in

a classroom. On the other hand, low risk connections are related to contacts that have a low probability for disease transmission. For instance, these may be contacts between members of different work teams in the same company.

We adopted most of the concrete values for the model parameters from the existing literature on flu epidemics [21], [24], [25]. The epidemic model also receives as an input the social model constructed in the previous step.

Vaccination: Our simulator provides for the possibility of vaccinating a subset of individuals either before the outbreak of the epidemics or at any other point during the outbreak. The lower half of Figure 1 consists of subscripted states which reflect the susceptible, latent, asymptomatic, infectious, and hospitalized states for the case of vaccinated individuals. The figure contains a transition from state S to state S_t which reflects the adoption of a vaccination policy for susceptible individuals. Since in case of the flu virus no symptoms are evident during the latent period it is in reality possible to vaccinate individuals either in the latent or in the asymptomatic state. We assume that getting vaccinated when are states L or A does not make any difference with respect to the individual's response to infection. Vaccination has specific implications such as: reducing the susceptibility of getting infected at the time of contact with an infected individual, reducing the probability of infecting another individual, reducing the recovery time, and reducing the possibility of becoming symptomatic. Vaccination is implemented such that it is possible to control the number of vaccines available and the probability of it succeeding when applied to a specific individual. Due to the fact that only part of the population is susceptible as result of a vaccination program we now use for the subscripted cases a control reproduction number R_v instead of the basic reproduction number R_0 .

In case of an epidemics the period of time between its onset and the time when a vaccine becomes available is usually problematic because of the lack of understanding of the effects of the timing when the vaccine is administrated. Our simulator allows analyzing the effects of implementing a vaccination program at different times throughout the dissemination of the infectious agent. One of the advantages of our epidemic model is that it is possible to monitorize the effect of interventions such as vaccination or hospitalization for each individual. It is therefore possible to simulate various scenarios like vaccinating or insulating a specific collective, for instance, the members of a specific company or school, or a given city area.

3.3 The simulation algorithm

Our simulation algorithm uses as inputs both the social model as well as the epidemic model. The social model provides the intra-group connections for each individual; these are the paths through which the infectious agent may propagate and they may be either low-risk or high-risk. The

epidemic model captures the states that each individual goes through during an epidemics and the probabilities for taking transitions from a given source to a specific destination state. The simulation algorithm processes each connection of every individual to generate a probability with which the connection will serve for transmitting the infection. This probability depends on: (1) The connection type and current time: the connection types are intra-group, inter-group, and family, and each of them corresponds to a specific daily time slice; and (2) The current state of the individual: this is the current state in the epidemic model plus other factors like the group which he belongs to, age, etc.

3.4 Performance issues

EpiGraph has been designed as a fully parallel application. It employs MPI [26] to perform the communication and synchronization for both components of the simulator: the contact network model and the epidemic model. This approach has two main advantages. First, it can be executed efficiently both on shared memory architectures—for instance multicore processors—and on distributed memory architectures—such as clusters. On both platforms EpiGraph successfully exploits the hardware resources and achieves a significant reduction in execution time relative to a sequential implementation. The second advantage is that the simulator scales with the available memory. Given that all the data structures are evenly distributed, the size of the problems that can be simulated grows with the number of computational resources.

4. Results

Our main simulation scenario is the population of the Primary Metropolitan Statistical Area of Philadelphia, U.S. We used [23] to extract statistical demographic data for the city and we created a *basic scenario* with the following characteristics. The city has 3,849,647 inhabitants with the following distribution: 27.95% school-age children, 43.62% workers, 14.52% stay-home parents, and 13.92% retired individuals. The interconnection graph has 160 millions of contacts, on average 41 per inhabitant. Working hours are from 9am to 5pm, leisure time is from 5pm to 7pm, and time spent at home—family and sleep time—is from 7pm to 9am. We consider 13,181 groups of workers; 8,513 groups of school-age children corresponding to classrooms; 4,192 groups of stay-home parents corresponding to friends that share activities such as shopping or walking; and 4,314 groups of retired individuals. We use Gaussian distributions to assign a size to each group; the mean size for each of the four group types is 261, 39, 12 and 8.

Figure 2 displays in logarithmic scale the number of individuals in each epidemic state during a simulation of 200 days for our basic scenario. This scenario includes the following parameters extracted from [24]: the basic reproduction number $R_0 = 1.373$, the factor by which the

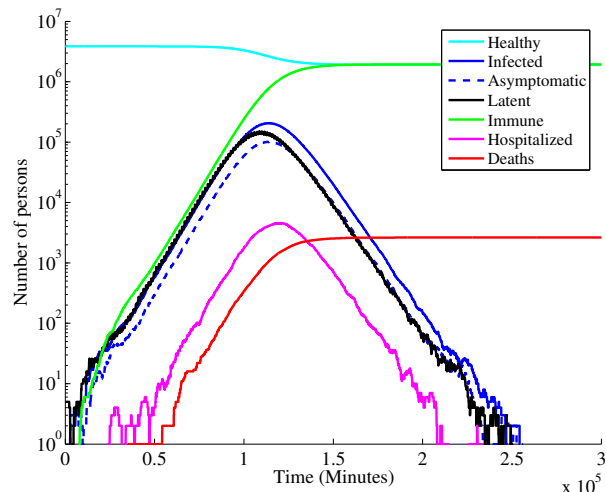


Fig. 2: Epidemic propagation for the basic scenario and a 200-day simulation.

infectivity of asymptomatic individuals is reduced $\delta = 0.5$, the probability that susceptibles become asymptomatic $p = 0.33$, the latent period for influenza 1.9 days, the infective period for influenza 4.1 days, and the hospitalization period 3 days. We can observe that the infection lasts approximately 175 days and its peak is around day 82.

We have performed a number of experiments in order to evaluate the strengths of EpiGraph. These experiments address three different properties of the simulator: (1) the prediction accuracy of the mathematical epidemic model, (2) the ability to accurately model highly heterogeneous scenarios where each individual and her connections may be customized, and (3) the performance and scalability of the simulator.

4.1 Validation of the EpiGraph model

In order to evaluate the accuracy of our mathematical model we compare the simulation results of EpiGraph with those of InflaSim [1]. In order to perform a comparison we used in both simulators the population and epidemic parameters of the basic scenario. Table 1 shows the number of susceptible, immune and dead individuals for each simulator. Results show deviations of 3.30%, 2.97% and 8.04% in the number of susceptible, immune, and dead individuals. Another aspect that we have considered is the numerical stability of EpiGraph under different conditions. More specifically, we have analyzed the variability of the results for two cases: when EpiGraph is executed several times with the same input parameters and when it is executed using different time step durations.

Table 1: InflaSim and EpiGraph results.

State	InflaSim	EpiGraph	Deviation
Susceptible	2,023,187	1,930,773	3.30%
Immune	1,837,305	1,916,226	2.97%
Deaths	2,362	2,647	8.04%

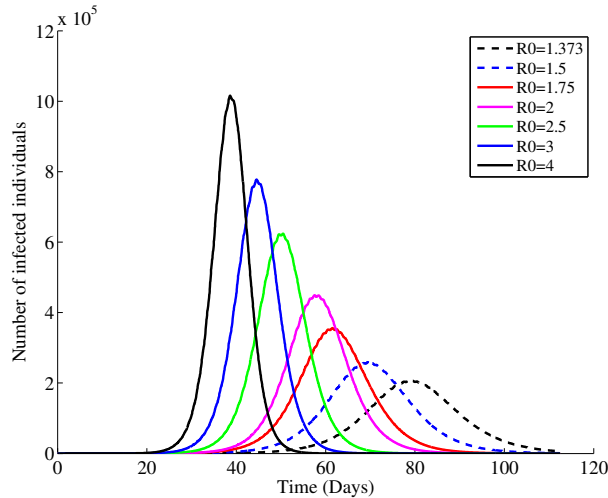


Fig. 3: Impact of different basic reproduction numbers on the number of infected. Basic scenario.

The time step determines the frequency of computations for each individual. By default we use a 10 minute step, which means that we apply the propagation model and update the system state six times per hour. A smaller time step implies a more detailed simulation at the expense of a longer execution time. We execute the basic scenario using the following time steps: 1, 5, 10, 30, 60 and 120 minutes. We observe that the loss of accuracy when using larger steps is not important. More specifically, the peak of infected individuals for all of these executions reaches a mean value of 205,168 with a standard deviation (in percentage of the mean value) of 1.17% and confidence interval of 1.21%. This peak is reached at the simulation time of 118,770 minutes, with a standard deviation of 2.83% and confidence interval of 2.97%.

To evaluate the variability of EpiGraph we run ten times the same scenario with the same initial conditions, including the same set of individuals that are initially infected. After repeatedly simulating the epidemics for 200-day intervals, results show a variability in the number of immune individuals of 0.28%. Similar results are obtained for susceptible and dead individuals. Based on these results we conclude that EpiGraph is able to precisely model the epidemic with a small variability in the results.

4.2 Exploiting the features of EpiGraph

EpiGraph employs a highly detailed social model which allows customizing the interactions of each individual as well as the effect of time on the individual relationships. These features allow the simulation of infection and transmission process for individual cases.

We have performed experiments aimed at evaluating the effect of different basic reproduction numbers and different graphs structures on the epidemic propagation. Figure 3 evaluates the effect of different reproduction numbers. We can see that the epidemic propagation is faster and the

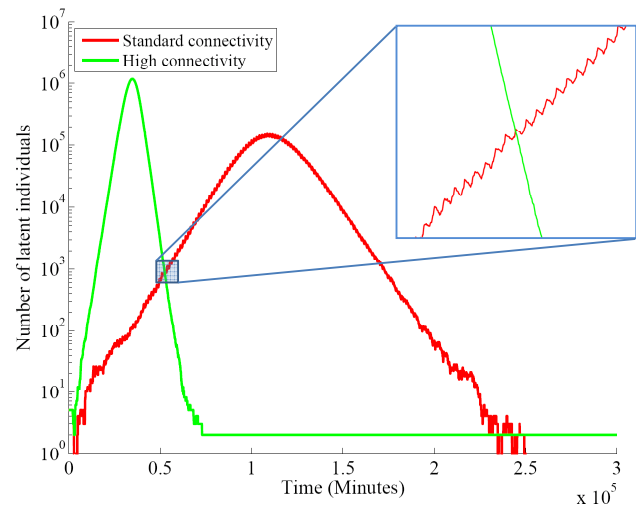


Fig. 4: Effect of different graph configurations on the latent cases. 200-day simulation.

number of infected individuals is larger when the basic reproduction number grows. For instance, for values of R_0 of 1.373, 2, and 4 the overall numbers of infected individuals are 1,933,901 and 2,783,435 and 3,597,751, respectively.

We evaluated two different graph structures called standard connectivity and high connectivity. Standard connectivity corresponds to the basic scenario; high connectivity corresponds to a scenario where the graph is flattened. Specifically we are considering only the graph connections corresponding to workers and we assume that the working hours are from 9am to 9am of the next day. That is, in this case we are considering a global graph that contains only one group type which is active during the whole day. Figure 4 illustrates the evolution of the latent cases for the scenarios of standard connectivity and high connectivity; infected cases exhibit a similar behavior. The figure shows that differentiating between social groups has a significant impact on the evolution of the epidemics. We can observe that when we assume standard connectivity there exists a periodic variation of the latent cases. This is related to the existence of different daily time slices that exhibit different propagation patterns. In the case of high connectivity this pattern doesn't appear due to the unique time interval, that of working hours.

We have evaluated the effect of different vaccination policies on the basic scenario. Figure 5 shows the evolution of the infected cases for five different strategies: no vaccination (reference), vaccination at the beginning of the outbreak, before reaching the peak of the outbreak (day 52), at the peak of the outbreak (day 82) and after the peak (day 97). For each of these cases 28% of the population is vaccinated and the reproduction number for vaccinated people is $R_v = 0.047$ [25]. We can observe the following behavior: vaccinating at day 0 is the most efficient approach in terms of minimizing the number of infected individuals. When vaccinating at day 52 there is a large number of

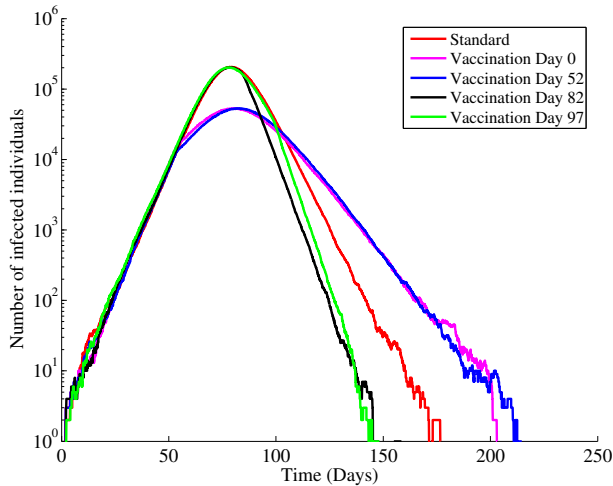


Fig. 5: Impact of different vaccination strategies. Basic scenario, 200-day simulation.

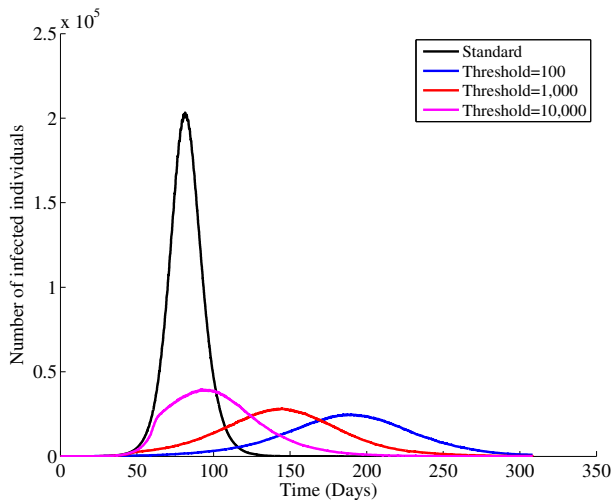


Fig. 6: Impact of different quarantine policies. Basic scenario, 300-day simulation.

individuals in infected and latent stages; the vaccine reduces the number of infected cases but also delays its propagation, thus increasing the duration of the outbreak. This effect is also manifested when vaccinating at day 0. In contrast, for the vaccination campaigns at days 82 and 97 the peak of infected cases has already been reached; vaccination thus contributes to an early end of the outbreak.

Lastly, we evaluated different quarantine policies. For the basic scenario we specify a given threshold in number of infected cases. When this threshold is reached all the school and work activities are cancelled, keeping only two leisure hours per day; during the rest of the day all the individuals stay at home with their family. Figure 6 shows the simulation results when quarantine is applied based on different threshold values. We observe that there is a decrease in the number of infected at the expense of a larger propagation time.

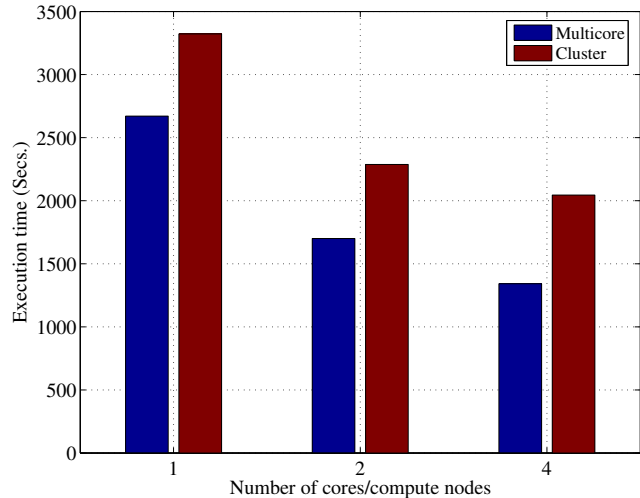


Fig. 7: EpiGraph execution time on a multicore processor and a cluster. Basic scenario, 200-day simulation.

4.3 Performance evaluation

We measured the execution time of EpiGraph on two different parallel architectures: a multicore processor and a cluster. The multicore is an Intel Xeon X7350 quadcore processor with a frequency of 2.93 GHz, 3 MB of cache and 16GB of RAM. The cluster consists of 4 computers connected with a GigaBit network, each of them with a single Intel Xeon E5405 at 2GHz with 6MB of cache and 4GB of RAM. Figure 7 shows the EpiGraph execution time for the basic scenario when simulating 200 days of epidemic outbreak. Given the faster interconnection system of the multicore architecture, this achieves better performance than the cluster system. We can observe that in both cases EpiGraph reduces its execution time when more processors are used.

5. Conclusions

This paper presents a novel approach to modeling the propagation of the flu virus via a realistic interconnection network based on actual individual interactions extracted from social networks. We have implemented a scalable, fully distributed simulator and we present an extensive analysis of the effects of the new features of our approach on the results of the simulations. Work in progress and future work involve studying the effects of introducing new states in the epidemic model and making use of the individual values such as age and gender in implementing different social and medical propagation characteristics. We are also interested in investigating the characteristics of our social models—such as clustering, node distance, and so on—and estimate to what degree disease propagation occurs differently for different types of real social networks.

Acknowledgements

The work has been performed under the HPC-EUROPA2 project (project number: 228398) with the support of the European Commission-Capacities Area-Research Infrastructure and the Spanish Ministry of Science and Education under the MEC 2011/00003/001 contract.

References

- [1] M. Eichner, M. Schwehm, H. P. Duerr, and S. Brockmann, "The influenza pandemic preparedness planning tool *influsim*," *BMC Infectious Diseases*, vol. 7, no. 17, pp. e-pub, 2007.
- [2] M. J. Keeling and K. T. D. Eames, "Networks and epidemic models," *Journal of The Royal Society Interface*, vol. 2, no. 4, pp. 295–307, Sept. 2005. [Online]. Available: <http://dx.doi.org/10.1098/rsif.2005.0051>
- [3] I. Doherty, N. Padian, C. Marlow, and S. Aral, "Determinants and consequences of sexual networks as they affect the spread of sexually transmitted infections," *The Journal of Infectious Diseases*, vol. 191, no. S1, p. 42–54, 2005.
- [4] K. T. D. Eames and M. J. Keeling, "Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases," *Proc Natl Acad Sci U S A*, vol. 99, no. 20, pp. 13 330–13 335, Oct. 2002. [Online]. Available: <http://dx.doi.org/10.1073/pnas.202244299>
- [5] S. Bansal, B. T. Grenfell, and L. A. Meyers, "When individual behaviour matters: homogeneous and network models in epidemiology," *Journal of The Royal Society Interface*, vol. 4, no. 16, pp. 879–891, Oct. 2007. [Online]. Available: <http://dx.doi.org/10.1098/rsif.2007.1100>
- [6] R. M. Christley, G. L. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, R. Bennett, and J. Turner, "Infection in social networks: Using network analysis to identify High-Risk individuals," *American J. of Epidemiology*, vol. 162, no. 10, pp. 1024–1031, 2005.
- [7] L. A. Meyers, B. Pourbohloul, M. E. Newman, D. M. Skowronski, and R. C. Brunham, "Network theory and SARS: predicting outbreak diversity," *Journal of theoretical biology*, vol. 232, no. 1, pp. 71–81, Jan. 2005. [Online]. Available: <http://dx.doi.org/10.1016/j.jtbi.2004.07.026>
- [8] M. E. J. Newman, *The spread of epidemic disease on networks*, Apr. 2002. [Online]. Available: <http://arxiv.org/abs/cond-mat/0205009>
- [9] J. M. Read, K. T. Eames, and W. J. Edmunds, "Dynamic social networks and the implications for the spread of infectious disease," *Journal of the Royal Society, Interface / the Royal Society*, vol. 5, no. 26, pp. 1001–1007, Sept. 2008. [Online]. Available: <http://dx.doi.org/10.1098/rsif.2008.0013>
- [10] A. Vazquez, "Spreading dynamics on heterogeneous populations: Multitype network approach," *Phys. Rev. E*, vol. 74, no. 6, p. 066114, Dec 2006.
- [11] F. Harary, *GRAPH THEORY*. Addison Wesley Longman Publishing Co, 1969. [Online]. Available: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/B000OLF0P0>
- [12] D. B. West, *Introduction to Graph Theory*, 2nd ed. Prentice Hall, Sept. 2000.
- [13] T. Zhang, S. H. Soh, X. Fu, K. K. Lee, L. Wong, S. Ma, G. Xiao, and C. K. Kwok, "Hpcgen a fast generator of contact networks of large urban cities for epidemiological studies," in *International Conference on Computational Intelligence, Modelling and Simulation*, 2009, pp. 198–203.
- [14] F. C. Coelho, O. G. Cruz, and C. T. Codeco, "Epigrass: a tool to study disease spread in complex networks," *Source code for biology and medicine*, vol. 3, no. 1, Feb. 2008. [Online]. Available: <http://dx.doi.org/10.1186/1751-0473-3-3>
- [15] S. Eubank, A. V. S. Kumar, M. V. Marathe, A. Srinivasan, and N. Wang, "Structural and algorithmic aspects of massive social networks," in *SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2004, pp. 718–727. [Online]. Available: <http://portal.acm.org/citation.cfm?id=982792.982902>
- [16] F. Chung and L. Lu, "Connected components in random graphs with given expected degree sequences," *Annals of Combinatorics*, vol. 6, pp. 125–145, 2002, 10.1007/PL00012580. [Online]. Available: <http://dx.doi.org/10.1007/PL00012580>
- [17] E. Volz and L. A. Meyers, "Susceptible-infected-recovered epidemics in dynamic contact networks," *Proc Biol Sci*, vol. 274, no. 1628, 2007.
- [18] T. C. Germann, K. Kadau, I. M. Longini, and C. A. Macken, "Mitigation strategies for pandemic influenza in the united states," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5935–5940, Apr. 2006. [Online]. Available: <http://dx.doi.org/10.1073/pnas.0601266103>
- [19] K. Carley, D. Fridsma, E. Casman, A. Yahja, N. Altman, L.-C. Chen, B. Kaminsky, and D. Nave, "Biowar: scalable agent-based model of bioattacks," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 36, no. 2, pp. 252 – 265, 2006.
- [20] R. M. Anderson, R. M. May, and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control*, new ed ed. Oxford University Press, USA, Sept. 1992. [Online]. Available: <http://www.worldcat.org/isbn/019854040X>
- [21] F. Brauer, P. v. d. Driessche, and J. Wu, Eds., *Mathematical Epidemiology*, 1st ed. Springer, June 2008. [Online]. Available: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/3540789103>
- [22] M. E. Alexander, C. S. Bowman, Z. Feng, M. Gardam, S. M. Moghadas, G. RÅúst, J. Wu, and P. Yan, "Emergence of drug resistance: implications for antiviral control of pandemic influenza," *Proceedings of the Royal Society B: Biological Sciences*, vol. 274, no. 1619, pp. 1675–1684, 2007.
- [23] *U. S. Census Bureau*, <http://www.census.gov/>.
- [24] I. M. Longini, E. M. Halloran, A. Nizam, and Y. Yang, "Containing pandemic influenza with antiviral agents," *Am. J. Epidemiol.*, vol. 159, no. 7, pp. 623–633, Apr. 2004. [Online]. Available: <http://dx.doi.org/10.1093/aje/kwh092>
- [25] L. R. Elveback, J. P. Fox, E. Ackerman, A. Langworthy, M. Boyd, and L. Gatewood, "An influenza simulation model for immunization studies," *American Journal of Epidemiology*, vol. 103, no. 2, pp. 152–165, 1976. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/814808>
- [26] *MPI: A Message-Passing Interface Standard*, Message Passing Interface Forum, 1995.

Altered Gliclazide Metabolic Pathway and its Implications on Increased Therapeutic Response in CYP2C9*2: Molecular Dynamics simulation and Autodock Studies

N. Renuka¹, Hussaina Banu¹, and Geetha Vasanthakumar^{1*}

¹Department of Biotechnology, Manipal University, P. O. Box 345050, Dubai, United Arab Emirates

*Corresponding author

Abstract – Among sulfonylureas, gliclazide is prescribed to 80% of the diabetic population and mainly metabolized by CYP2C9 in Caucasians. Our data shows that the orientation of the substrate is changed and therefore, the site of oxidation with respect to heme-Fe is altered in *2. This leads to an altered metabolic pathway in *2 and it is a rate limiting step in gliclazide metabolism. Our results also show that the position of 7-propionate side chain of ring A and 6-propionate side chain of ring D is flipped in *2 and thus, the stability of heme and oxidative potential of substrate in the active binding pocket are reduced. In summary, the altered pathway, and instability of heme and the substrate in the active site are contributing to decreased metabolic activity consistent with greater therapeutic response observed in patients carrying CYP2C9 *2 allele.

Keywords: CYP2C9*2, gliclazide, therapeutic response, molecular dynamics, docking simulation, metabolic pathway

1 Introduction

Among sulfonylureas, gliclazide is dispensed almost 4 million prescriptions in UK [1] and 1.2 million prescriptions in Australia [2]. It is also given in combination with metformin to keep successful control of the disease [3, 4]. Comparing with other hypoglycemic agents, the incidence of hypoglycemia is relatively low in gliclazide and may have beneficial effects beyond reduction of blood glucose [5]. In Caucasians, gliclazide is extensively metabolized by CYP2C9. Pharmacokinetic clearance of gliclazide revealed the existence of two major metabolites due to the oxidation of methyl carbon of tolyl-group that constitutes ~60% of metabolites and hydroxylation at a specific site in the azabicyclo-octyl ring represent ~40% of metabolite observed in urine [6,7] as shown in Fig. 1.

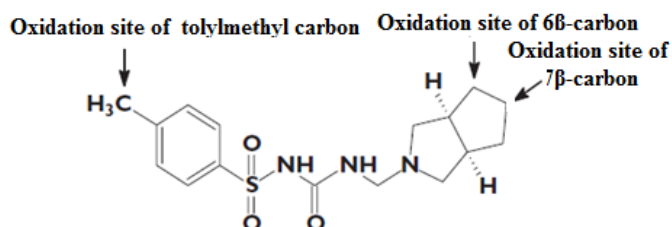


Fig. 1. Location of the hydroxylation sites in gliclazide

CYP2C9 is the major human enzyme of the cytochrome P450 2C subfamily and it is responsible for metabolism of ~10% of therapeutic drugs in the market. This gene is highly polymorphic [8,9] and so far twenty four alleles have been identified [10]. Two alleles, *2/*2 (R144C) and *3/*3 (I359L) genotype carriers had a lower gliclazide clearance, with reductions of 25 and 57%, respectively, relative to those carrying the wild type [11-13]. Crystallographic data confirmed that the I359L variation is located in proximity to the active center in the substrate recognition site (SRS) 5 and therefore, explain the loss of functional activity in the variant allele. However, the codon 144 amino acid substitution is located outside the active center and therefore, the loss of activity observed in this allele is not clear. Minor differences in frequencies of these genotypes between different ethnic subgroups of the Caucasians population have been reported and the variant CYP2C9*2 (*2) is almost absent in Africans and Asian population [13]. Pharmacogenetic study conducted in larger population of 1073 patients with type-2 diabetes recruited between 1992 and 2007 demonstrated that the loss-of-function alleles *2 are robustly associated with greater response to sulfonylureas and approximately 80% of the patients treated only with gliclazide in this study population [14]. The influence of CYP2C19 polymorphism in the pharmacokinetics of gliclazide has been reported in healthy Chinese population [15, 16]. This small discrepancy may be due to the ethnic differences and also due to the selection of smaller population for shorter periods of treatment.

Pharmacokinetic studies show that 6β- 7β-, and tolylmethyl- hydroxylation represents the rate-limiting pathway of gliclazide elimination [7]. Our previous molecular docking of gliclazide on *2 studies indicate that 6β- and 7β- carbon

atom is closer to heme-Fe [17]. Based on this, our hypothesis is that β -hydroxylation may be the preferred route of metabolism and this may lead to the reduced metabolic clearance of gliclazide observed in *2. Therefore, in this study we are proposing to use molecular dynamic simulation and automated molecular docking tools to better understand the altered substrate orientation, proton – heme distance, binding pocket, heme and gliclazide stabilization, and regioselectivity of metabolism in *2 allelic variant and thus it leads to the altered route of metabolism

2 Materials and methods

2.1 Computational methods of CYP2C9*1 & *2

With the X-ray crystal structure of human CYP2C9/flurbiprofen (PDB code 1R9O) [18] as a model, substrate free computational models of CYP2C9 *1 (wild, *1) and *2 (R144C) were constructed using the software tools VMD and NAMD [19, 20]. The missing amino acid residues 38 - 42 and 214 - 220 were also included in the computational models using Modeller [21]. The generated models were validated for their structural quality using Procheck [22, 23].

2.2 Molecular dynamics simulation

The generated computational models were further processed for MD simulation. The intermolecular hydrogen atoms were added and the complexes were solvated in a layer of TIP3 water molecules of 10Å radius, ionized at a physiological pH of 7 and subjected to energy minimization for 2000 steps. The minimized protein complex was simulated using NAMD for 600 picoseconds without any restrains at a constant temperature of 300 K. In each model, the lowest potential energy state was chosen for further analysis, whose stability was examined by calculating the root mean square deviation of the protein backbone.

2.3 Flexible docking

The simulated protein was further processed using the molecular modeling program CHIMERA [24] to remove water and ions, and add Gastegier charges and hydrogen atoms. Gliclazide was docked with the above models using the grid-based docking program AutoDock 4.2 [25-27], in which some of the key residues of the active site were kept flexible. The best ten clusters having the lowest energies and <2Å RMSD values were chosen for analysis.

3 Results and discussion

Modeller was used to generate *1 and *2 models and the Procheck program was used to check the stereochemical quality of a protein structure within the allowed Ramachandran region. The results show that 92% and 94% of

residues in 3D structure of *1 and *2 lie in the most favored regions and 0.7% and 0.5% of residues lie in disallowed regions of the Ramachandran plot (Fig. 1a & 1b). The docking results indicate that gliclazide is located nearby heme and surrounded by SRS residues. The location of SRS residues in *1 and *2 are in consistence with the results of the crystal structure of CYP2C9 [18] and confirms the validity of our docked models.

The automated molecular docking using Autodock was performed to further validate the reliability of the conformation of the SRS in *1 and *2 models (Fig. 2a & 2b).

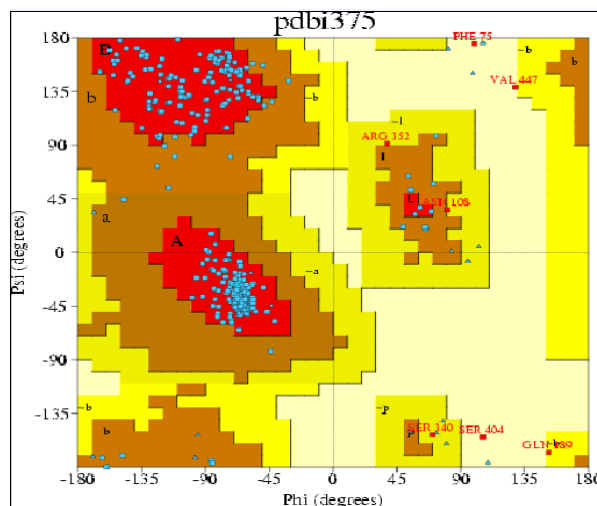


Fig. 1a. Ramachandran plot of *1^a

^aRamachandran Plot statistics

	No. of residues	%-tage
Most favoured regions [A,B,L]	373	92.1%
Additional allowed regions [a,b,l,p]	25	6.2%
Generously allowed regions [~a,~b,~l,~p]	4	1.0%
Disallowed regions [XX]	3	0.7%
<hr/>		
Non-glycine and non-proline residues	405	100.0%
End-residues (excl. Gly and Pro)	3	
Glycine residues	27	
Proline residues	31	
<hr/>		
Total number of residues	466	

^aBased on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20.0 a good quality model would be expected to have over 90% in the most favoured regions [A,B,L]

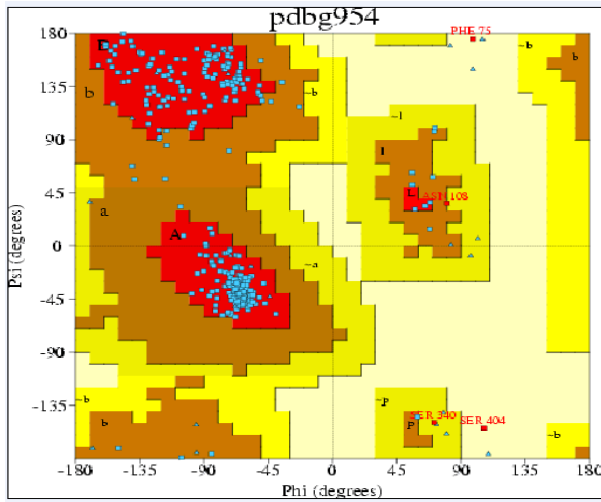


Fig. 1b. Ramachandran plot of *2^a

^aRamachandran Plot statistics

	No. of residues	%-tage
Most favoured regions [A,B,L]	380	93.8%
Additional allowed regions [a,b,l,p]	21	5.2%
Generously allowed regions [~a,~b,~l,~p]	2	0.5%
Disallowed regions [XX]	2	0.5%
<hr/>		
Non-glycine and non-proline residues	405	100.0%
<hr/>		
End-residues (excl. Gly and Pro)	3	
Glycine residues	27	
Proline residues	31	
<hr/>		
Total number of residues	466	

^aBased on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20.0 a good quality model would be expected to have over 90% in the most favoured regions [A,B,L]

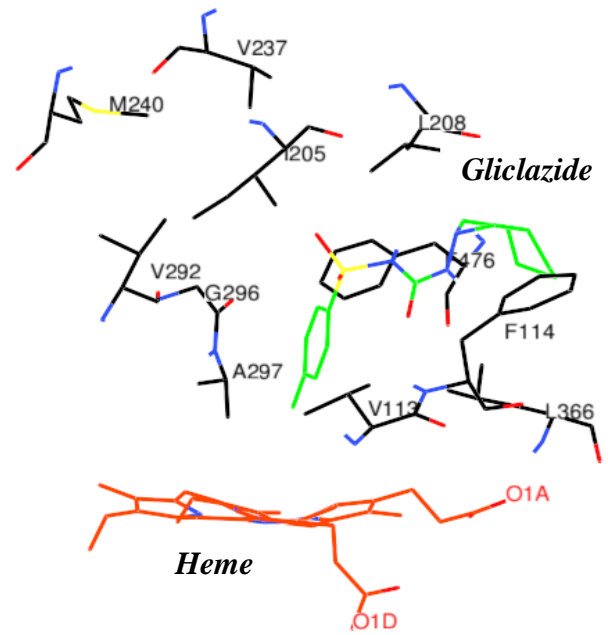


Fig. 2a. Substrate recognition site of *1

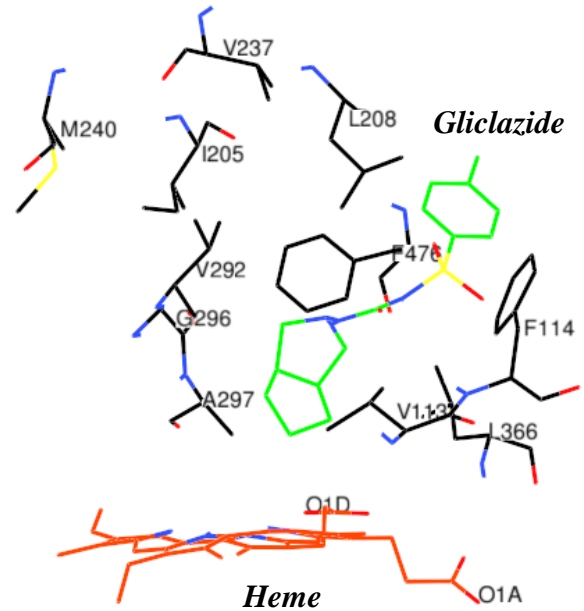


Fig. 2b. Substrate recognition site of *2

In CYP2C9 crystal structure study, the heme and active binding pocket are buried deep into the protein molecule and the substrate should access the binding pocket for the occurrence of catalysis. Since the substrate access channel and the binding pocket near the heme-Fe play an important role in the determination of the orientation of the substrate towards heme-Fe, we have examined these factors in this study. Our previous study shows that the number of amino acids forming the hydrophobic cage is not changed in *2 but the size of the substrate access channel is reduced from 10.3Å (*1) to 9.3Å (*1) [17], and this may change the orientation of the substrate entering into the binding pocket and thus alter the position of the substrate in the binding pocket. Since the size of the substrate access channel is smaller in *2, azabicyclo- group may enter first rather than the bulky methyl-phenyl group. We believe that this resulted in the complete change in the orientation of gliclazide in the binding pocket (Fig. 3a & 3b).

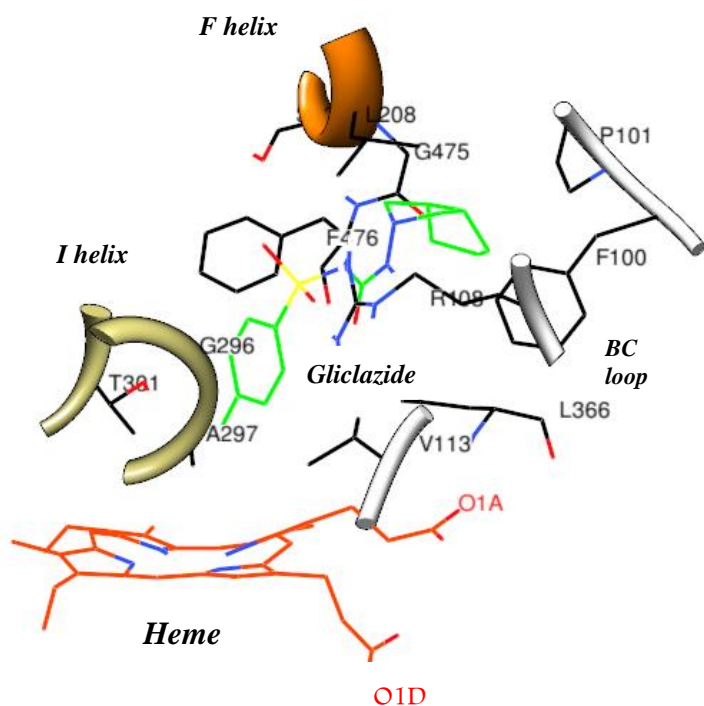


Fig. 3a. Binding pocket of *1 after docking gliclazide

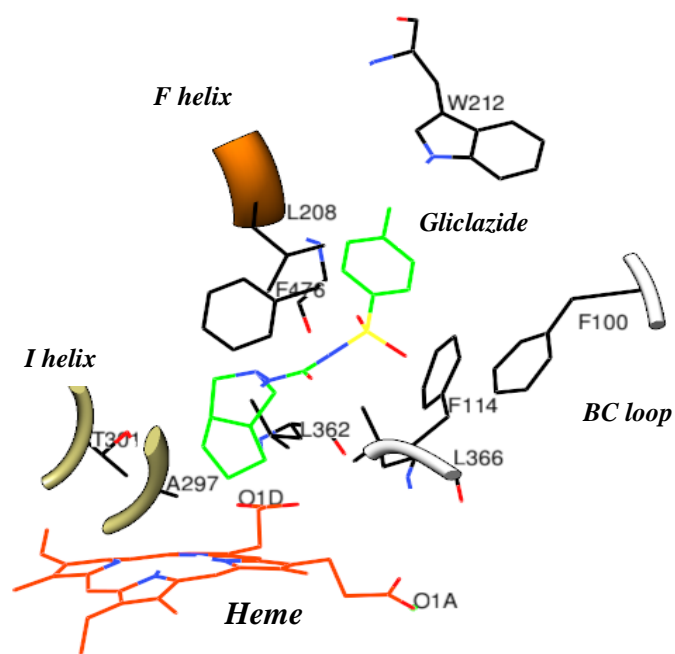


Fig. 3b. Binding pocket of *2 after docking gliclazide

Our previous studies show that the distance between tolyl methyl carbon atom of gliclazide and the heme is 4.1Å and 6β-carbon atom and the heme is 10.3Å in *1 [17]. These results correlate with the pharmacokinetic data which shows that the tolylmethyl hydroxylation is the major pathway responsible for metabolic clearance of gliclazide in *1 [6] and this constitutes ~60% of the metabolites detected in urine. While in *2, tolylmethyl carbon atom is located at 14.7Å & hydroxyl group is located at 4.5Å from heme and this suggests that β-hydroxylation may be favored route of metabolic clearance of gliclazide. According to pharmacokinetic data, this route of metabolism constitutes only ~40% of the metabolite and therefore, it may explain the reduced activity observed in diabetic patients carrying *2 allele. Previous study [28] shows that the differences in the distance between substrate proton to heme-Fe play a key role in the observed differences in catalytic activity. NMR derived T_1 relaxation studies conducted with the probe substrate flurbiprofen and co-incubation of flurbiprofen with dapsone show that the movement of flurbiprofen protons closer to the heme iron partially explains heteroactivation observed in CYP2C9 allelic variants [28,29].

Amino acids in the binding pocket of both *1 and *2 are similar, except BC loop amino acids V113 and R108 are not present in *2 (Fig. 3a & 3b). R108 stabilizes the gliclazide by binding to the acidic group of gliclazide and formation of hydrogen bonds (Table 1). R108 itself is stabilized by the formation of hydrogen bond with D293, thus gliclazide positions itself in proximity with heme prosthetic group for subsequent oxidation in *1 as represented in Table 1. While

gliclazide and R108 stabilization by the hydrogen bonding are lacking in *2 (Table 1). These results are consistent with the proposed catalysis model for P450 [30, 31]. Hydrophobic amino acids, G296 and G475, that stabilizes the binding pocket are absent in *2 (Fig. 3b).

Table 1: Gliclazide stabilization in *1 compared to *2 by hydrogen bond formation

Donor	Acceptor	Distance (Å) ^a		Angle (°) ^b	
		*1	*2	*1	*2
R108	Gliclazide Carboxyl O1	2.0	5.6	153	116
R108	Gliclazide Carboxyl O1	1.9	6.5	160	102
R108	D293 OD2	2	4.4	124	124

The donor and acceptor distance is $<3.5 \text{ \AA}$ ^a [30] and $180^\circ \pm 45^\circ$ ^b [31], a hydrogen bond is defined to be formed

Many factors are known to affect the heme redox potential, including proximal heme ligand and propionate and substrate orientations and interactions with the immediate protein environment. In *1 complex, the A ring propionate is stabilized by the formation of hydrogen bonding with S365, L366, and R97, whereas the D ring is stabilized by R124, R433, and W120 by the formation of hydrogen bonds (Fig. 4a ; Table 2).

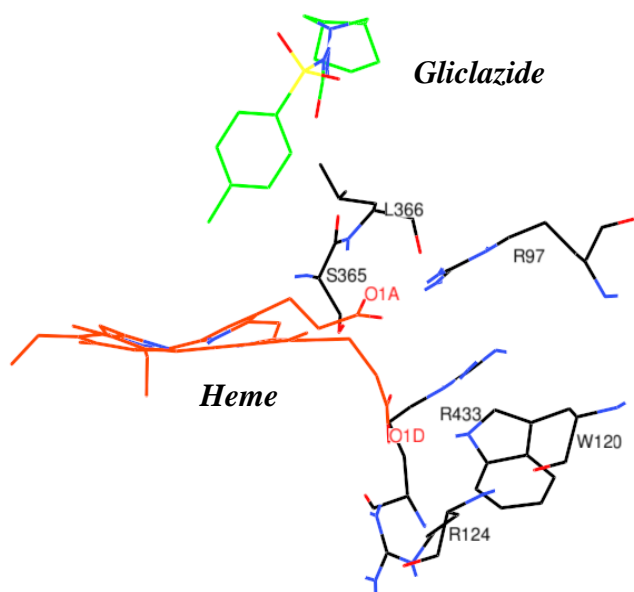


Fig. 4a. Heme stabilizing amino acids of *1

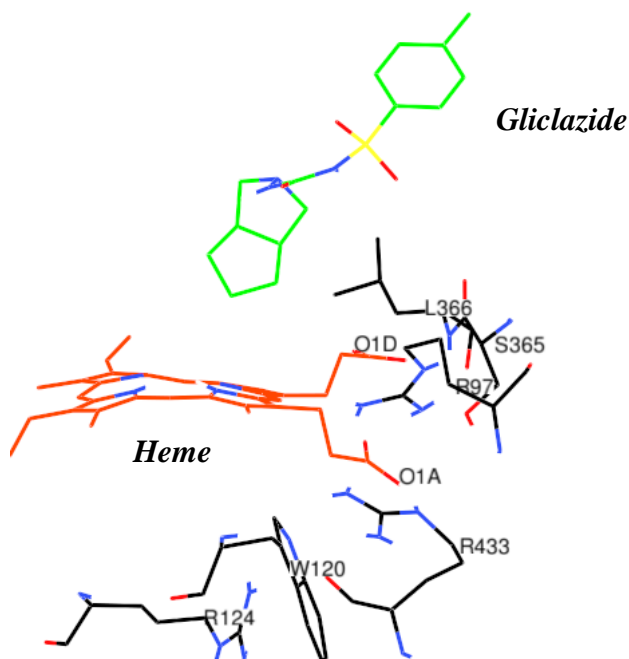


Fig. 4b. Heme stabilizing amino acids of *2

Table 2: Heme stabilization in *1 compared to *2 by hydrogen bond formation

Donor	Acceptor	Distance (Å) ^a		Angle (°) ^b	
		*1	*2	*1	*2
W120 NE1	Heme O1D	1.9	15.7	168	158
R124 NH1	Heme O1D	1.8	14.6	162	77
S365 NG1	Heme O1A	1.6	4.1	164	108
L366 NH	Heme O1A	1.9	4.4	157	149
R433 NH1	Heme O2D	2.4	14.2	143	58

The donor and acceptor distance is $<3.5 \text{ \AA}$ ^a [30] and $180^\circ \pm 45^\circ$ ^b [31], a hydrogen bond is defined to be formed.

This structure is consistent with the closed form of 2C enzyme reported earlier [32]. In contrast to this conformation, the position of rings A and D are flipped in *2 and the stabilization of both propionate rings are lacking (Fig. 4b; Table 2). These alterations in heme coordination may affect the heme redox potential. Mutagenesis and structural studies indicate the importance of the movement of the ring A towards substrate for the occurrence of oxidation [33]. Replacement of the D-ring resulted in the loss of enzyme activity and confirms the importance of this propionate in catalytic activity [33].

4 Conclusions

In summary, our present study shows that the orientation of gliclazide is altered significantly and changes the nature of the functional group located closer to heme-Fe and therefore, the site of oxidation is changed in *2. Since tolyl-methyl group is closer to heme-Fe, tolylmethyl-hydroxylation of gliclazide is the preferred route of metabolism in *1. While

azabicyclo-octyl ring is closer to heme-Fe and therefore, 6 β - and 7 β -hydroxylation is the preferred route of metabolism in *2. The reduced catalytic activity in *2 is consistent with pharmacokinetic data where the detection of 6 β -hydroxylation metabolite is only ~40%. The position of SRS amino acid residues are not altered but in the binding pocket, B-C loop amino acid residues are missing in *2. In addition, we show that the substrate access channel and a significant change in the structural link between substrate binding and the binding of redox partners would also partly explain the reduced catalytic efficiency.

5 Acknowledgements

NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. Authors are indebted to Dr. Fiaz from SZABIST, Dubai Campus for the use of their computer facility. We are grateful to Dr. Ramjee, Director and Dr. Firdos Alam Khan, Chairperson, Department of Biotechnology, Manipal University, Dubai for their continuous support and encouragement to conduct this research.

6 References

- [1] <http://www.ic.nhs.uk/pubs/precostanalysis2005/final/file>.
- [2] http://www.medicareaustralia.gov.au/providers/health_statistics/statistical_reporting/pbs.html
- [3] A. Ward, M. Salas, J.J. Caro, D. Owens, "Health And Economic Impact Of Combining Metformin With Nateglinide To Achieve Glycemic Control: Comparison Of The Lifetime Cost Of Complications In The U.K", *Cost Effectiveness and Resource Allocation*, 2, 2-10, 2004.
- [4] J.K. DiStefano, R.M. Watanabe, "Pharmacogenetics Of Anti-Diabetes Drugs, Pharmaceuticals", 3, 2610-2646, 2010.
- [5] K.J. Palmer, R.N. Brogden, "Gliclazide. An update of its pharmacological properties and therapeutic efficacy in non-insulin-dependent diabetes mellitus", *Drugs*, 46, 92-125, 1993.
- [6] T. Oida, K. Yoshida, A. Kagemoto, Y. Sekine, T. Higashijima, "The metabolism of gliclazide in man", *Xenobiotica*, 15, 87-96, (1985).
- [7] D.J. Elliot, M.S. Suharjono, B.C. Lewis, E.M. Gillam, D.J. Birkett, A. Gross, J.O. Miners, "Identification Of The Human Cytochromes P450 Catalysing The Rate-Limiting Pathways Of Gliclazide Elimination", *Br. J. Clin. Pharmacol.* 4, 64, 450-457, 2007.
- [8] J. Blaisdell, L.F. Jorge-Nebert, S. Coulter, S.S. Ferguson, S.J. Lee, B. Chanas T. Xi, H. Mohrenweiser, B. Ghanayem, J.A. Goldstein, "Discovery Of New Potentially Defective Alleles Of Human Cyp2c9", *Pharmacogenetics*, 14, 527-537, 2004.
- [9] L.J. Dickmann, A.E. Rettie, M.B. Kneller, R.B. Kim, A.J. Wood, C.M. Stei, G.R. Wilkinson, U.I. Schwarz, "Identification And Functional Characterization Of A New CYP2C9 Variant (CYP2C9*5) Expressed Among African Americans", *Mol. Pharmacol.*, 60, 382-387, 2001.
- [10] <http://www.imm.ki.se/CYPalleles/cyp2c9.htm>.
- [11] J. Kirchheiner, J. Brockmöller, I. Meineke, S. Bauer, W. Rohde, C. Meisel, I. Roots, "Impact Of CYP2C9 Amino Acid Polymorphisms On Glyburide Kinetics And On The Insulin And Glucose Response In Healthy Volunteers", *Clin. Pharmacol. Ther.*, 71, 286-296, 2002.
- [12] A.E. Rettie, J.P. Jones. "CYP2C9: Clinical And Toxicological Relevance", *Ann. Rev. Pharmac. Toxic.*, 45, 477-494, 2005.
- [13] J. Kirchheiner, M. Tsahuridu, W. Jabrane, I. Roots, J. Brockmöller, "The CYP2C9 Polymorphism: From Enzyme Kinetics To Clinical Dose Recommendations", *Future Medicine*, 1, 63-84, 2004.
- [14] K. Zhou, L. Donnelly, L. Burch, R. Tavendale, A.S.F. Doney, G. Leese, A.T. Hattersley, M.I. McCarthy, A.D. Morris, C.C. Lang, C.N.A. Palmer, E.R. Pearson, "Loss Of Function CYP2C9 Variants Improve Therapeutic Response To Sulfonylureas In Type 2 Diabetes: A Go-Darts Study", *Clin. Pharmacol. Ther.*, 87, 52-56, 2009.
- [15] Y. Zhang, D. Si, X. Chen, N. Lin, Y. Guo, H. Zhou, D. Zhong, "Influence Of CYP2C9 And CYP2C19 Genetic Polymorphisms On Pharmacokinetics Of Gliclazide MR In Chinese Subjects", *Br. J. Clin. Pharmacol.*, 64, 67-74, 2007.
- [16] H. Shao, X. M. Ren, N.F. Liu, G.M. Chen, W.L. Li, Z.H. Zhai, D.W. Wang, "Influence Of CYP2C9 And CYP2C19 Genetic Polymorphisms On Pharmacokinetics And Pharmacodynamics Of Gliclazide In Healthy Chinese Han Volunteers", *J. Clin. Pharm. Ther.*, 35, 351-360, 2010.
- [17] Hussaina Banu, N. Renuka, Geetha Vasanthakumar, "Reduced Catalytic Activity Of Human Cyp2c9 Natural Alleles For Gliclazide: Molecular Dynamics Simulation And Docking Studies", *Biochimie*, Article in press, 2011.
- [18] M.R. Wester, J.K. Yano, G.A. Schoch, C. Yang, K.J. Griffin, C.D. Stout, E.F. Johnson, "The Structure Of Human Cytochrome P450 2c9 Complexed With Flurbiprofen At 2.0 Å Resolution", *J. Biol. Chem.*, 279, 35630-35637, 2004.
- [19] W. Humphrey, A. Dalke, K. Schulten, "VMD - Visual Molecular Dynamics", *J. Molec. Graphics*, 14, 33-38, 1996.
- [20] C. James, Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kale, Klaus Schulten, "Scalable Molecular Dynamics With NAMD", *J. Comput. Chem.*, 26, 1781-1802, 2005.
- [21] A. Sali, T. L. Blundell, "Comparative Protein Modelling By Satisfaction Of Spatial Restraints", *J. Mol. Biol.*, 234, 779-815, 1993.
- [22] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, "PROCHECK - A Program To Check The Stereochemical Quality Of Protein Structures", *J. App. Cryst.*, 26, 283-291, 1993.
- [23] R.A. Laskowski, R.A. Rullmann, M.W. MacArthur, R. Kaptein, J.M. Thornton, "AQUA and PROCHECK-NMR: Programs For Checking The Quality Of Protein Structures Solved By NMR", *J. Biomol. NMR*, 8, 477-486, 1996.

- [24] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, "UCSF Chimera-A Visualization System For Exploratory Research And Analysis", *J. Comput. Chem.*, 25, 1605-1612, 2004.
- [25] D.S. Goodsell, A.J. Olson, "Automated Docking of Substrates to Proteins by Simulated Annealing", *Proteins: Struct., Funct. & Genetics*, 8, 195-202, 1990.
- [26] G.M. Morris, D.S. Goodsell, R. Huey, A.J. Olson, "Distributed Automated Docking of Flexible Ligands to Proteins: Parallel Applications of AutoDock 2.4", *J. Comput. Aided Mol. Des.*, 10, 293-304, 1996.
- [27] G. M. Morris, D. S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, "Automated Docking Using Lamarckian Genetic Algorithm And An Empirical Binding Free Energy Function", *J. Comp. Chem.*, 19, 1639-1662, 1998.
- [28] M.A. Hummel, P.M. Gannett, J. Aguilar, T.S. Tracy, "Substrate Proton To Heme Distances In CYP2C9 Allelic Variants And Alterations by the Heterotropic Activator, Dapsone", *Arch. Biochem. Biophys.*, 475, 175-183, 2008.
- [29] M.A. Hummel, P.M. Gannett, J. Aguilar, T.S. Tracy, "Effector-Mediated Alteration Of Substrate Orientation In Cytochrome P450 2C9", *Biochemistry*, 43, 7204-7214, 2004.
- [30] E. Sano, W. Li, H. Yuki, X. Liu, T. Furihata, K. Kobayashi, K. Chiba, S. Neya, T. Hoshino, "Mechanism Of The Decrease In Catalytic Activity Of Human Cytochrome P450 2C9 Polymorphic Variants Investigated By Computational Analysis", *J. Comput. Chem.*, 31, 2746-2758, 2010.
- [31] G.D. Szklarz, R.L. Ornstein, J.R. Halpert, "Application of 3-Dimensional Homology Modeling Of Cytochrome P450 2B1 For Interpretation Of Site-directed Mutagenesis Results", *J Biomol Struct Dyn* 12, 61-78, 1994.
- [32] E.E. Scott, M. A. White, Y. A. He, E.F. Johnson, C.D. Stout, J.R. Halpert, "Structure Of Mammalian Cytochrome P450 2B4 Complexed With 4-(4-chlorophenyl)imidazole At 1.9-Å Resolution", 279, 27294-27301, 2004.
- [33] D. Fishelovitch, S. Shaik, H.J. Wolfson, R. Nussinov, "How Does The Reductase Help To Regulate The Catalytic Cycle Of Cytochrome P450 3A4 Using The Conserved Water Channel", *J. Phys. Chem. B*, 114, 5964-5970, 2010.

Optimizing a Cost Matrix to Solve Rare-Class Biological Problems

Mark J. Lawson¹, Lenwood S. Heath², Hai Zhao³, and Liqing Zhang²

¹Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA

²Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

³Department of Computer Science, Shanghai Jiaotong University, Shanghai, China

Abstract—*In a binary dataset, a rare-class problem occurs when one class of data (typically the class of interest) is far outweighed by the other. Such a problem is typically difficult to learn and classify and is quite common, especially among biological problems such as the identification of gene conversions. A multitude of solutions for this problem exist with varying levels of success. In this paper we present our solution, which involves using the MetaCost algorithm, a cost-sensitive “meta-classifier” that requires a cost matrix to adjust the learning of an underlying classifier. Our method finds this cost matrix for a given dataset and classification algorithm, creating a final classification model. Through a detailed description, a basic evaluation, and the application to the problem of identifying gene conversions, we show the effectiveness of this approach. Our novel approach to generating a cost matrix has proven to be quite effective in the identification of gene conversions and represents a robust way to tackle the rare-class data problem.*

Keywords: Rare class, cost matrix, gene conversion

1. Introduction

Gene conversion, an important biological process, refers to the exchange of DNA sequence information between two genes [1]. Caused through DNA strand breaks, one gene (the donor) donates part or all of its sequence to another gene (the acceptor). This can lead to two types of evolutionary processes: gene conservation and genetic diversity. By having two genes repeatedly “convert” each other for the entire sequence, they can remain identical or highly similar in sequences, despite the fact that they were duplicated a long time ago. This has been observed in genes such as ribosomal RNA genes and genes on the human X-Chromosome [2]. On the other hand, if two genes exchange only part of their sequences, it can lead to the creation of new sequences, creating the potential for genetic diversity. This has been observed in gene families where diversity is important, such as immunoglobulin genes [3] and human major histocompatibility complex genes [4].

The identification of gene conversions is important for understanding the evolution of duplicated genes and the cause of certain genetic diseases. However, current gene conversion identification software has poor performance [2], with high false negative rates due to the fact that prediction of gene conversion is a rare-class problem.

Rare-class prediction (also referred to as “imbalanced” data prediction) is a common problem in classification [5]. In this type of problem, one class of data is far outweighed by other classes, thus making it difficult for a classification algorithm to accurately predict this class after learning. This is typically confronted in a binary class problem, in which there are two classes, often referred to as the minority and majority classes. Typically the minority class is the class of interest but the created classifier performs poorly in identifying those data members. A typical result is that the classifier classifies all data members as being majority class members, due in part to the concept of Occam’s razor [6], in which the simplest hypothesis is used to create the classifier. These classification algorithms are also designed to maximize predictive accuracy, which favors the majority class.

Many approaches exist for solving the rare-class problem. These are typically one of two types: data-level approaches and algorithm-level approaches. Data-level approaches consist of two main ideas: oversampling, in which minority class members are increased through re-use, and undersampling, in which majority class members are filtered out. Both strive to attain a balanced dataset, thus allowing the classifier the ability to better differentiate between the two classes. However they both suffer from shortcomings: oversampling can easily lead to overfitting and undersampling is likely to remove relevant data objects from the training set. Recent approaches have attempted to rectify these shortcomings: SMOTE (Synthetic Minority Oversampling TEchnique) creates synthetic minority class data members based on existing ones [7] and a recent undersampling approach uses clustering to filter out irrelevant majority class data members [8].

The other methods consist of algorithm-level approaches. The most common is cost-sensitive learning in which the learning of an underlying classifier is adjusted based on pre-determined misclassification costs. One of these approaches is MetaCost [9]. MetaCost takes in training data and a classification algorithm and adjusts the learning by taking into account a given cost matrix that assigns punishments for misclassifications and rewards for correct classifications. The advantage of MetaCost is that it has a “black box” approach, any classification algorithm can be used and there are no limits on types of training data. However, the cost matrix must be known in advance [10], which is usually impossible.

Currently there is no systematic way to determine an ideal cost matrix for a given dataset and classification algorithm. We propose a greedy-based approach for determining a cost matrix. Based on the given training data and the given classification algorithm, our approach incrementally searches for a cost matrix, returning the best one it finds to the user. At worst, the returned cost matrix and classification model perform as well as an unaltered classification algorithm, but our evaluations show general improvement in rare-class datasets. In this paper, we will present a formal, detailed description of this approach and illustrate why it is effective. In addition we will show its power through the classification of a basic, example dataset and how it performs in the prediction of gene conversions.

2. Methods

2.1 MetaCost

The classification problem is to take a classification algorithm L and train it on a set of training data S , thereby creating a model M . M is then used to predict the classes of additional data, based on a learned hypothesis. A training set S consists of a set of samples, each having a vector of attributes and an assigned label. An optimal model would sufficiently learn S so that it can correctly identify every x in the test data T . However, an optimal model is typically not possible, so we seek to create an approximation that achieves the best results.

An attempt that is focused on approximating this optimal model, especially in regards to rare-class problems, is MetaCost [9]. The basic idea of MetaCost is to take a normal, unaltered classifier and adjust the learning with a cost matrix. This is done through a series of steps. The first step is to take the training data and create multiple bootstrap samples of the data. These bootstrap samples are then used for training to create an ensemble of classifiers. The ensemble of classifiers are then combined through a majority vote to determine the probability of each data object x belonging to each class label. Next, each data object in the training data is relabeled based on the evaluation of a *conditional risk* function, and a final classifier is then produced after applying the classification algorithm to the relabeled training data.

The key aspect in the MetaCost learning process is to minimize *conditional risk*,

$$R(i|x) = \sum_j P(j|x)C_{i,j}. \quad (1)$$

$R(i|x)$ defines the cost of predicting that data object x belongs to class label i instead of class label j , $P(j|x)$ is the probability that data object x belongs to class label j , and $C_{i,j}$ is the cost for making such a classification. $C_{i,j}$ corresponds to entries in the cost matrix, essentially a variant of the confusion matrix (Table 1) where $i \in \{0,1\}$ and

$j \in \{0,1\}$. The cost matrix allows one to punish misclassifications and reward correct classifications, for example, by negative and positive values, respectively. Clearly, the success of the evaluation of the *conditional risk* function and thereby the performance of the MetaCost prediction rests on the cost matrix. Imaginably a bad cost matrix can distort the learning and produce a bad classifier. Therefore, it is imperative to identify a high quality cost matrix.

Table 1: Confusion Matrix

TP True Positive	FP False Positive
FN False Negative	TN True Negative

$$C = \begin{bmatrix} C_{0,0} & C_{0,1} \\ C_{1,0} & C_{1,1} \end{bmatrix} \quad (2)$$

2.2 Cost Matrix Optimization

The MetaCost algorithm has input values m , n , and p that are essentially tweaks or givens of the algorithm once the type of classifier is determined. To simplify the function call, we can fix some default values for them, thus, the call to the MetaCost algorithm becomes a function of S , L , and C and returns a classification model M . Let us define an evaluation function $\text{Eval}(M, T)$ that takes as input a generated model M based on a cost matrix C and produces an evaluation of its performance on test set T . This evaluation function can be based on any of the metrics for rare-class predictions such as F-measures, ROC curves, and G-mean. Assuming that we have access to the set of all possible cost matrices ($C_i, i \in N^*$), we can then search for the cost matrix that achieves the highest evaluation value,

$$C_{best} = \arg \max_C \text{Eval}(M_{C_i}, T), \quad (3)$$

and denote C_{best} as the *optimal* cost matrix for the given data and classification algorithm.

So the problem is how to find the *optimal* cost matrix computationally. While an exhaustive search of all possible cost matrices can guarantee that we find the *optimal* cost matrices, it is not possible. Here we propose a greedy approach to heuristically find a matrix that produces a high evaluation value.

Shown in Algorithm 1, the basic idea of the search is to start with an initial cost matrix and to increment its costs to find a cost matrix that achieves a better evaluation value. An initial cost matrix is typically a cost matrix that will create a model that is the same as the model created by an unaltered classifier. Our positive class is the minority class.

Starting with this initial cost matrix, the method creates seven new ones ($A_0, A_1, A_2, A_3, A_4, A_5, A_6$). Each of these cost matrices represents a different combination of incrementing/decrementing the costs (correct classifications are

Algorithm 1 Greedy-Based Search

Input:
S is the training set
T is the test set
L is a classification algorithm
5: *n* is the number of iterations to run the algorithm

{0,1} is the set of classes
Let $\text{Eval}(M, T)$ return an evaluation value on how Model *M* performed on test set *T*

10: **Function** GreedyCost(*S*, *T*, *L*, *n*)

Let *I* be the initial cost matrix where all punishments/rewards are 0
Let *C* be the current best cost matrix, initialized to *I*
Let *M_C* be the current best model, initialized to MetaCost(*S*, *L*, *C*)

15: Let *O* be the overall best cost matrix, initialized to *I*
Let *M_O* be the overall best model, initialized to *M_C*

for *i* = 1 to *n* **do**
Let *A* be a set of cost matrices

20: where

$$A_0 \leftarrow C + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$A_1 \leftarrow C + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

$$A_2 \leftarrow C + \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$A_3 \leftarrow C + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

25: $A_4 \leftarrow C + \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}$

$$A_5 \leftarrow C + \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix}$$

$$A_6 \leftarrow C + \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix}$$

Set *C* and *M_C* to null

30: **for** *j* = 0 to 6 **do**
M = MetaCost(*S*, *L*, *A_j*)
if $\text{Eval}(M, T) > \text{Eval}(M_C, T)$ **then**
C = *A_j* and *M_C* = *M*
end if

35: **end for**

if $\text{Eval}(M_C, T) > \text{Eval}(M_O, T)$ **then**
O = *C* and *M_O* = *M_C*
end if

40: **end for**

return *O*, *M_O*

decremented by one and misclassifications are incremented by one). Of note here is that the cost for correct classifications of the majority class (negative class) is not adjusted and left at zero. This is due to the fact that typically a poor classifier will order most (if not all) data members as belonging to the majority class. Therefore, there is no need to reward such behavior and our method has fewer cost matrices to test. After creating these seven new cost matrices, each one is used to create a new model through MetaCost, using the given training data *S* and classification algorithm *L*. After these models have been created, they are evaluated on the given test set *T* using the evaluation function $\text{Eval}(M, T)$. The model that has the highest evaluation value is kept and

its cost matrix is used to initialize the next iteration of cost matrix creation.

As can be seen in the algorithm, the method keeps track of two cost matrices, a “current best” cost matrix and an “overall best” cost matrix. This was done in order to overcome one of the common problems with greedy searches, that of finding a local maximum that is lower than the global maximum. So when a potential poor local maximum is reached, it can be stored as the overall best and the method can essentially “look ahead” to see if a better cost matrix can be found. If a better one is found, the overall best cost matrix is updated. So conceivably, we can continue to generate new cost matrices while still keeping track of a good one. While this does not guarantee that a global maximum will be found, it does allow for a more comprehensive search than a typical greedy search.

The parameter *n* is passed into the function to give a count of how many times the creation of new cost matrices occurs. A simple check of whether the overall best matrix is the same as the current best cost matrix serves as an indication of whether a maximum (local or global) has been reached. If not, the number of iterations can be increased. A possible modification to this algorithm would be to have a set number of iterations to run after a maximum has been reached.

The search for the best cost matrix can only improve upon a base classifier. At worst, the method will work as well as an unaltered classifier. This is due to the fact that a model that is built by the MetaCost algorithm with the initial cost matrix is identical to a model that was built using only the base classification algorithm. So if no better cost matrix is found, the initial cost matrix will be returned as the best.

One final note is in regards to the use of training and test data. While it is ideal if they are different, it is not necessary and training data can be used for both the creation of the model and evaluation of the cost matrix. Having a separate set of test data gives the learning process more breadth as using only one set of training data does bias the classification model towards this training data. So for evaluation purposes of the final generated classification model, one must have an additional set of test data to use that was not part of the learning process and cost matrix search.

3. Experiment

3.1 Gene Conversion Data and Classification Programs

Because actual gene conversion data is difficult to obtain, we created simulated gene conversion data similar to an approach developed by Marais [11]. Essentially we simulated the creation of a gene family from a root sequence (through mutation along a simulated phylogenetic tree) and inserted a gene conversion event between two of the genes. This way we could create recent gene conversion events (by having mutations take place mostly before the gene conversion

event) and more ancient gene conversion events (by having more mutations occur after the event). We then created two sets of data: SET1 which consisted of multiple recent gene conversions and SET2 which consisted of multiple ancient gene conversions. Each of these datasets consisted of multiple gene families (consisting of six genes each) and one (or no) gene conversion event. For each of these datasets, we created a large set of training data, a set of evaluation test data to be used in the greedy-based search, and a set of final test data to evaluate the final generated classifiers. The results shown in the next section are of how the classifiers performed on this final test data, data that was not used in the learning process.

For our experiments, we used two gene conversion prediction programs, GENECONV [12] and Partimatrix [13]. GENECONV is a program designed for the identification of gene conversions that gives a prediction of what sequence fragments have the highest, unique similarity between two sequences, ranking these predictions by p -value. Partimatrix uses bipartitions to determine if DNA sequences show evidence of anomalous phylogenetic history, giving support and conflict scores for each prediction.

For classification, we represented each pair of genes within a gene family through a feature vector. In this representation, we can see that gene conversion is a rare-class data problem. A set of six genes represents 15 gene pair combinations and at most one of these gene pairs will have a gene conversion event. Each of these feature vectors consists of the following attributes: average GC content, overall sequence similarity, GENECONV prediction global and pairwise p -values, and Partimatrix conflict and support scores.

Classification was done through the greedy-based search for a cost matrix that we detailed in the methods section. We used the following classification algorithms as the underlying classifiers: NaiveBayes (as implemented by John and Langley [14]), J4.8 (an implementation of the C4.5 decision tree learner [15]), PART (a combination of rule-based learning and C4.5 [16]), and JRip (a rule-based learner based on RIPPER [17]). All classification algorithms were implemented in weka [18], a collection of machine learning algorithms.

3.2 Results

In Table 2 we can see the classification results. For our purposes the positive class is when a gene pair has a gene conversion and the negative class is when it does not. For the learning of each set, we created separate training and test data and then evaluated the final model on a second set of unique test data.

In SET1, one can see that GENECONV performs quite well. “GENECONV Strict” has a high accuracy, even higher than the “Just Say No approach”. However, through our method we are able to increase the amount of true positives,

Table 2: Simulation Results

SET1				
Classifier	TP	FP	Accuracy	F-measure
<i>Perfect</i>	139	0	1	1
<i>Just Say No</i>	0	0	0.937	UNDEF
<i>GENECONV Strict</i>	102	4	0.975	0.840
<i>GENECONV LP</i>	123	57	0.955	0.776
<i>Partimatrix</i>	9	137	0.833	0.064
<i>G-or-P</i>	128	191	0.874	0.561
<i>NaiveBayes</i>	122	58	0.954	0.770
<i>PART</i>	107	5	0.978	0.859
<i>J4.8</i>	109	11	0.975 8	0.848
<i>JRip</i>	111	9	0.978	0.864
SET2				
Classifier	TP	FP	Accuracy	F-measure
<i>Perfect</i>	150	0	1	1
<i>Just Say No</i>	0	0	0.933	UNDEF
<i>GENECONV Strict</i>	1	8	0.930	0.014
<i>GENECONV LP</i>	5	68	0.905	0.045
<i>Partimatrix</i>	15	135	0.880	0.100
<i>G-or-P</i>	19	197	0.854	0.104
<i>NaiveBayes</i>	8	75	0.904	0.069
<i>PART</i>	35	214	0.854	0.175
<i>J4.8</i>	23	160	0.872	0.138
<i>JRip</i>	40	265	0.833	0.176

This table represents the performance of the various classification methods on datasets SET1 and SET2. The upper half represents the basic classifiers that do not use the greedy-based approach. *Perfect* represents a theoretical optimal classifier and is included for comparison. *Just Say No* represents a classifier that classifies all data elements as majority class. *GENECONV Strict* uses only global p -values for predictions, whereas *GENECONV LP* uses local pairwise p -values (with 0.05 being used as the threshold for positive classification). *Partimatrix* represents a prediction based on the lowest conflict score between a gene pair within a gene family. *G-or-P* is a basic unification of *GENECONV LP* and *Partimatrix* predictions. The lower half represents the classification algorithms predictions after using the greedy-based search for a cost matrix.

increase the accuracy, and most importantly, increase the F-measure. The best performers are JRip and PART, which is not surprising as they are rule-based classifiers and rule-based classifiers are known to perform well on rare-class data [19]. Both have a higher F-measure than “GENECONV Strict”, a higher accuracy, and both identify more true positives. J4.8 does well too and identifies more true positives than PART, but more false positives as well. Of all the cost matrix classifiers, NaiveBayes identifies the most true positives, but is hindered by the number of false positives it identifies.

In SET2, one can see that ancient gene conversions are far more difficult to accurately detect, as the mutations after the conversion makes some difficult to differentiate. GENECONV performs quite poorly, both in Strict and LP. Partimatrix identifies more gene conversions and G-or-P has the best F-measure of these basic classifiers. This set also shows the shortcoming of using accuracy as a metric as the “Just Say No” approach would appear to be the best classifier. Among the cost matrix classifiers, the NaiveBayes classifier performs quite poorly. It has an F-measure lower than G-or-P, so it shows no improvement over a basic classifier (it does not identify more gene conversions

correctly either). But the rule-based classifiers again perform quite well, with both identifying more gene conversions and having higher F-measures than any of the basic classifiers. In fact, aside from NaiveBayes, all classifiers exhibit both a higher recall and a higher precision than the basic classifiers, showing a definite improvement.

Table 3: Final Generated Cost Matrices

	SET1	SET2								
NaiveBayes	<table border="1"><tr><td>-3</td><td>2</td></tr><tr><td>2</td><td>0</td></tr></table>	-3	2	2	0	<table border="1"><tr><td>-2</td><td>2</td></tr><tr><td>1</td><td>0</td></tr></table>	-2	2	1	0
-3	2									
2	0									
-2	2									
1	0									
PART	<table border="1"><tr><td>-4</td><td>3</td></tr><tr><td>5</td><td>0</td></tr></table>	-4	3	5	0	<table border="1"><tr><td>-3</td><td>1</td></tr><tr><td>19</td><td>0</td></tr></table>	-3	1	19	0
-4	3									
5	0									
-3	1									
19	0									
J4.8	<table border="1"><tr><td>-2</td><td>3</td></tr><tr><td>4</td><td>0</td></tr></table>	-2	3	4	0	<table border="1"><tr><td>0</td><td>1</td></tr><tr><td>4</td><td>0</td></tr></table>	0	1	4	0
-2	3									
4	0									
0	1									
4	0									
JRip	<table border="1"><tr><td>-4</td><td>1</td></tr><tr><td>6</td><td>0</td></tr></table>	-4	1	6	0	<table border="1"><tr><td>0</td><td>1</td></tr><tr><td>7</td><td>0</td></tr></table>	0	1	7	0
-4	1									
6	0									
0	1									
7	0									

In Table 3, we can see the cost matrices that were determined for each classifier by the greedy-based approach and subsequently used to make gene conversion predictions. From this table it is quite clear that a cost matrix is highly dependent on both the classifier and the data being used. No classifier has the same cost matrix across both datasets and no dataset has a cost matrix that is best for more than one classifier. In fact, all cost matrices that were determined by our approach are unique. All final classification models were generated after 25 iterations of the greedy-based search method.

3.3 Additional Analysis

In order to further analyze the improvement our greedy search method has over an “unaltered classifier,” i.e. a classifier whose learning has not been altered by a cost matrix, we generated 10 samples for each gene conversion dataset and compared the performance of each classification algorithm. These samples were created by taking each dataset, SET1 and SET2, and splitting up the data as 1/2 training, 1/4 evaluation test data (for the evaluation in the greedy algorithm), and 1/4 for the final test data, which was not involved in the learning process. For the unaltered classification the evaluation test data was added back to training data, so 3/4 of the data was used for training and 1/4 for testing. Since we are dealing with datasets in which the amount of positive examples is few and the ratio between positive and negative examples is important, the creation of these samples was not entirely random. First the dataset was ordered into positive and negative examples. Then 1/2 of the positive samples were put into the training data, 1/4 in the evaluation test data, and 1/4 into the final test data. The same is then done with the negative data. This ensures that each set contains positive data members and that the ratio is conserved. After running the simulations, we used a

Wilcoxon signed rank test [20] to determine the significance of improvement.

In Tables 4 and 5, we see the resulting F-Measures, summarized as average, maximum, and minimum. In the SET1 simulation results, we can see improvement for each classification algorithm except NaiveBayes. The others see improvement in their average, maximum, and minimum F-Measures (however JRip has a lower maximum). Unfortunately, only PART shows significant improvement by using the greedy method, generating a p -value of 0.024.

In the SET2 simulation (Table 5), we see F-Measure improvements for all classification algorithms. However, the improvement for JRip is misleading. In its unaltered form, the generated classifiers made no positive predictions, hence the F-Measure of 0. The classifiers generated with the greedy method made ONLY positive predictions, generating an F-Measure of 0.125 each time. Clearly, this is not a “better” classifier. The other three classification algorithms showed significant improvement, each generating a p -value of 0.0027.

Table 4: SET1 Simulation

Classifier	F-Measure		
	Average	Max	Min
<i>Unaltered</i>			
NaiveBayes	0.770	0.796	0.744
JRip	0.874	0.904	0.846
PART	0.858	0.880	0.823
J4.8	0.861	0.883	0.839
<i>Greedy</i>			
NaiveBayes	0.766	0.793	0.738
JRip	0.876	0.892	0.857
PART	0.875	0.885	0.862
J4.8	0.873	0.900	0.848

Table 5: SET2 Simulations

Classifier	F-Measure		
	Average	Max	Min
<i>Unaltered</i>			
NaiveBayes	0.105	0.119	0.089
JRip	0.000	0.000	0.000
PART	0.014	0.031	0.000
J4.8	0.000	0.000	0.000
<i>Greedy</i>			
NaiveBayes	0.135	0.174	0.125
JRip	0.125	0.125	0.125
PART	0.176	0.185	0.161
J4.8	0.151	0.169	0.140

3.4 Real-World Data

In order to see how the generated classification models perform on real world data, we used the “Transcription Elongation Factor A” gene family on the X-Chromosome that has been shown to exhibit gene conversions [2]. The three gene family members are located in a large syntenic region that is conserved between primates and rodents, indicating that these genes were generated/duplicated before

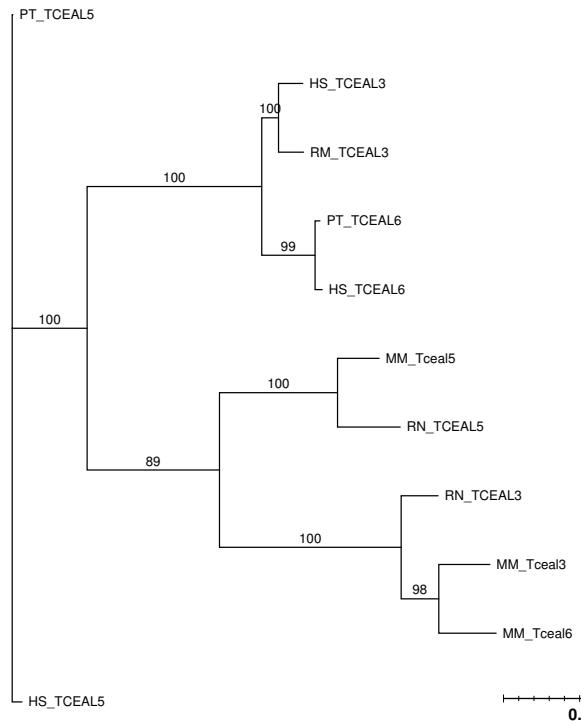


Fig. 1: Transcription Elongation Factor A phylogenetic tree
 HS = Homo sapien, PT = Pan troglodytes, RM = Rhesus Monkey
 (Macaca mulatta)
 MM = Mus musculus, RN = Rattus norvegicus

the split of primates and rodents. Thus, the phylogenetic tree (Figure 1) provides strong evidence for gene conversions within biological orders (primates and rodents). Interestingly, gene conversion seems to occur independently in both primates and rodents after their split but before the further splits within primates and within rodents.

We used the classification models from SET1 that were generated with the PART, J4.8, and JRip classification algorithms (NaiveBayes was left out due to its poor performance). Both PART and J4.8 made the same 3 predicted gene conversions: MM_Tceal3 and MM_Tceal5, MM_Tceal6 and MM_Tceal5, and RN_TCEAL3 and RN_TCEAL5. JRip made the same predictions, with the addition of a gene conversion between HS_TCEAL5 and PT_TCEAL6 that is a false positive due to the fact that gene conversion only occurs between genes from the same species. These predictions are consistent with the phylogenetic evidence.

Using the threshold of a p -value of 0.05 as sufficient evidence that two genes have undergone gene conversion GENECONV Strict gives evidence for 13 gene conversions and GENECONV LP for 44. While some gene conversions do correspond with what is seen in the graph, others do not, for instance gene conversions between primates and rodents.

Partmatrix does not provide guidelines for a threshold to be used for predicting gene conversions. However those with the highest support scores also involve conversions between primates and rodents.

Unlike the simulated cases where we can use the F-measure to compare the performance of gene conversion prediction programs with our ensemble method, it is difficult to perform this analysis on real data because we do not know the exact numbers of true or false positives and negatives. The challenge of the difficulty in performance evaluation on real world data can be addressed in future work by manual compilation of a carefully monitored set of genes for which exact numbers of true or false positives and negatives can be accurately inferred.

4. Discussion

Due to the complexity and uniqueness of datasets, as well as the differing performance of classification algorithms, the best performance can be achieved with a cost-sensitive classification method when a best cost matrix is found for both the given data and the given classification algorithm. Theoretical research on the rare-class problem has shown that aspects of data that are difficult to quantify (such as the “complexity of concept”) play a role in classification [10] and our own results have shown that a cost matrix that achieves good performance is dependent on both the given training data and the given classification algorithm. Thus a cost matrix must be found taking these two entities into consideration.

A greedy search is efficient but not optimal. While it cannot be proven that the eventual “overall best” cost matrix is one that achieves optimal classification results, we have shown that it will improve upon an unaltered classifier. At worst, the resulting classification model will perform as well as a classifier that was generated without MetaCost. This is more than can be said about other methods for dealing with rare-class data that can cause overfitting and/or eliminate relevant data and achieve even poorer results.

One thing we were able to recreate with this method, was the “black box” approach that MetaCost used. Of importance was the fact that the details are hidden from the end-user, with inputs being passed in and a final model being returned, with little user interaction. Our approach requires only the same inputs with the simple addition of a value being given for the number of iterations. At the end of these iterations, the best model and cost matrix found will be returned to the user. In addition, our method only requires a “meta-classifier” that takes in a cost matrix and adjusts the learning of a classification algorithm according to it. While MetaCost is a great method for accomplishing this, it can easily be replaced with a method that might be better suited for a specific problem domain.

Our future work will focus on improving the search for a best cost matrix. Simulated annealing [21] and genetic

algorithms [22] will be experimented with to see if they achieve better performance in terms of classification. While these methods can achieve better results than greedy search, they do require more time as they generate many more possible solutions. Therefore, we will investigate whether there is a trade-off between performance gain and increased searching time when compared to the greedy-based solution. In addition, we will also look into any improvements to the MetaCost algorithm that may increase performance (for instance, using boosting instead of bagging to determine probabilities as suggested in [23]). Finally, although our current analysis shows that MetaCost with the greedy search of a cost-matrix made some improvement in predicting gene conversion over GENECONV and Partimatrix, it is based on simulated data and rather limited real data. Future work will involve the curation and application of more real data on gene conversion to train and test models in order to further improve the performance of the prediction programs.

5. Acknowledgments

We thank Naren Ramakrishnan for helpful suggestions. The work was supported by NSF grant IIS-0710945 to L.Z.

References

- [1] J.-M. Chen, D. N. Cooper, N. Chuzhanova, C. Ferec, and G. P. Patrinos, "Gene conversion: Mechanisms, evolution and human disease," vol. 8, pp. 762–775, 2007, *nature Reviews Genetics*.
- [2] M. J. Lawson and L. Zhang, "Sexy gene conversions: Locating gene conversions on the X-chromosome," *Nucl. Acids Res.*, vol. 37, no. 14, pp. 4570–4579, 2009. [Online]. Available: <http://nar.oxfordjournals.org/cgi/content/abstract/37/14/4570>
- [3] N. Maizels, "Immunoglobulin gene diversification," *Annual Review of Genetics*, vol. 39, pp. 23–46, 2005.
- [4] N. Takahata and Y. Satta, "Selection convergence, and intragenic recombination in HLA diversity," *Genetica*, vol. 103, pp. 157–169, 1998.
- [5] N. V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.
- [6] P. Murphy and M. Pazzani, "Exploring the decision forest: An empirical investigation of Occam's razor in decision tree induction," *Journal of Artificial Intelligence Research*, pp. 171–187, 1994.
- [7] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [8] K. Yoon and S. Kwek, "A data reduction approach for resolving the imbalanced data issue in functional genomics," *Neural Computing & Applications*, vol. 16, pp. 295–306, 2007.
- [9] P. Domingos, "MetaCost: A general method for making classifiers cost-sensitive," *Advances in Neural Networks, International Journal of Pattern Recognition and Artificial Intelligence*, pp. 155–164, 1999.
- [10] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–450, 2002.
- [11] G. Marais, "Biased gene conversion: Implications for genome and sex evolution," *Trends Genet.*, vol. 19, no. 6, pp. 330–8, 2003.
- [12] S. Sawyer, "Statistical tests for detecting gene conversion," *Molecular Biology and Evolution*, vol. 6, no. 5, pp. 526–538, 1989.
- [13] I. B. Jakobsen, S. R. Wilson, and S. Easteal, "The partition matrix: Exploring variable phylogenetic signals along nucleotide sequence alignments," *Molecular Biology and Evolution*, vol. 14, no. 5, pp. 474–484, 1997.
- [14] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.
- [15] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
- [16] E. Frank and I. H. Witten, "Generating accurate rules sets without global optimization," *Fifteenth International Conference on Machine Learning*, pp. 144–151, 1998.
- [17] W. W. Cohen, "Fast effective rule induction," *Machine Learning: Proceedings of the Twelfth International Conference (ML95)*, 1995.
- [18] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Elsevier, 2005.
- [19] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction To Data Mining*. Addison-Wesley, 2006.
- [20] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.
- [21] B. Suman and P. Kumar, "A survey of simulated annealing as a tool for single and multiobjective optimization," *Journal of the Operational Research Society*, vol. 57, no. 10, pp. 1143–1160, 2006.
- [22] C. R. Reeves and J. E. Rowe, *Genetic Algorithms — Principles and Perspectives*. Kluwer Academic Publishers, 2003.
- [23] K. M. Ting, "An Empirical Study of MetaCost using Boosting Algorithms," In: *Proceedings of the Eleventh European Conference on Machine Learning*, pp. 413–425, 2000.

Flow and particle deposition simulations with heat-transfer in the nine-generation lung airways

B. Soni, N. Arra, and S. Aliabadi

Northrop Grumman Center for High Performance Computing, Jackson State University, Jackson, MS, USA

Abstract - *The flow fields in lower lung airways are rich in secondary flows in form of vortices. These vortices are known to impact the micro-particle transport in lung airways. The complex geometry of lung airways plays a significant role on the secondary flows generation. Inhaled air temperature may also have an impact on the bronchial tube flows and therefore to the particle deposition. The steady-state inspiratory air flow with and without heat-transfer were simulated in a nine-generation lung airway model using our in-house flow-solver. Particle traces were simulated using our Lagrangian based particle tracking software. The flow and particle trace simulation results with and without heat-transfer were compared. The effects of heat-transfer on the flow fields and particle deposition in the lung model were found to be insignificant suggesting that, the thermal effects can be overlooked when simulating the flow and particle transport in the small lung airways.*

Keywords: *biomechanics, lung airways, particle deposition, lung flow, computational Fluid Dynamics (CFD)*

1 Introduction

The complexity of lung air flow fields exists mainly due to the presence of secondary flows in form of vortices. These vortices are generated as a result of the bifurcating geometry of the bronchial tubes. The secondary flows are known to play a crucial role in particle transport from inhaled air. The study of bronchial tube flows can increase an understanding of the effects of the inhalation of harmful particles as well as the pulmonary drug delivery to improve human health. Particles suspended in the atmosphere are of various sizes and shapes [1]. Most of the particles from the atmosphere found inside the human lungs range in size from $2\text{-}10\mu\text{m}$, corresponding to coal dust, asbestos fiber, pollen, bacteria, etc [1]. There have been some studies identifying the health risks related to inhalation of micro- or nano-particles[2-9].

The bronchial tube geometry is characterized by bifurcations that produce multiple generations with asymmetric and nonplanar branching. There are a total of 18 generations (excluding the alveolus) [10] of airways in the human airway tree that consists of 2^{17} distinct tubes. The out-of-plane branch angles defining nonplanarity are randomly distributed to fill the chest cavity without any overlap. The effects of nonplanarity for asymmetric three-generation bronchial tube flow fields were investigated for three-

generation bronchial tube models by Soni et al. [11]. They also demonstrated significant difference between the particle deposition in the planar and nonplanar three-generation bronchial tube models [12].

The flows in small bronchial tubes are laminar with a Reynolds number less than 1000 [13]. However, the presence of vortices makes these flows quite complicated. The effects of secondary flow on particle dispersion were demonstrated by Soni et al. [12]. They used particle destination and Finite Time Lyapunov (FTLE) maps [14] to visualize particle deposition. The flow becomes more complex further down the tree due to the accumulative effects of nonplanarity and multiple generations. The bronchial tube flows can be categorized as primary and secondary flows. The flow in the direction of the local axis of the tube is called primary flow. The flows perpendicular to the local axis of the tube are called secondary flows. Figure 1 demonstrates the primary and secondary flows in the bronchial tubes. Figure 1(a) shows the primary velocity vectors at various cross-sections in the nine-generation bronchial tube model. The vectors are colored by dimensionless velocity magnitude. The secondary flows in form of a vortex pair in the second generation are shown in Figure 1(b). The cross-flow velocity vectors are plotted on the cutting plane which is colored by total velocity magnitude.

There have been some studies to simulate flows in bronchial tubes with more than just few generations in effort to achieve flow simulation of fully resolved bronchial tree. Nowak et al. [15] presented flow fields and particle transport simulations on lung airways with multigenerational symmetric planar model for up to 23 generations and a CT-scan model with nine generations. Ertbruggen et al. [16] described a lung airway model with eight generations containing 17 bifurcations and simulated steady-state flow with micro-particle transport. Gemci et al. [17] presented a simulation of 17 generations of the human lung based on the anatomical model of Schmidt et al. [18]. The geometry was only partially resolved, containing only 1453 bronchi as opposed to 2^{17} branches. In more recent efforts, Walters and Luke [19] proposed a Flow Path Ensemble (FPE) model to simulate flows in a nine-generation model with the model truncated so that the overall size of simulation was significantly reduced.

Thermal effects of the inhaled air temperature on the flow fields and particle deposition may become important when cold air or hot vapor is being inhaled. There are few experimental and numerical studies addressing heat-transfer in the lung airways [20-23]. The heat-transfer and mass-transfer was simulated for hot vapor by Zhang et al. [22]. They also

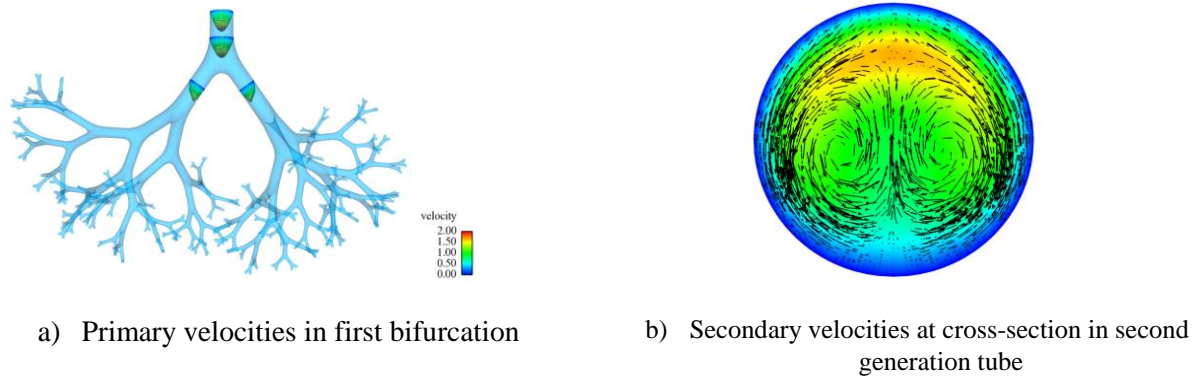


Figure 1. Bronchial tube flow fields.

studied the effect of cold weather on the steady state flow fields and particle deposition. The impact of temperature difference was found to be pronounced in the mouth to trachea geometry, whereas the thermal effects were found to be deteriorating in downstream generations. The effects were evaluated for three-generation, symmetric, planar bronchial tube model along with airway from mouth to trachea. In later study [23], they investigated heat-transfer and mass-transfer for hygroscopic droplets with unsteady flow conditions in same lung airway model.

2 Numerical modeling

The nine-generation geometry as shown in Figure 1(a), given by Walters and Luke [19], based on Weibel's [24] morphology for generations 4-12 of the human bronchial tree was employed in this study. A single parent tube being one-generation, two daughter tubes diverging from a parent tube defines two generations. A general expression to obtain total number of exits based on number of generations can be given as $N_{exit} = 2^{N-1}$, where N is the number of generations and N_{exit} is the number of exits. The parent tube diameter was taken to be $0.0057m$. The out-of-plane angles defining nonplanarity are randomly distributed between 0° to 180° . There are in total $2^8 = 256$ exits in this model.

Steady-state inhalation with Reynolds number of 319 corresponding to an inlet volumetric flow rate of $20.83 \text{ cm}^3/\text{s}$ were simulated with and without heat-transfer. The air flows were simulated by using the CaMEL flow solver.²⁵ CaMEL is an advanced computational fluid dynamics flow solver specifically developed for large scale simulations at the Northrop Grumman Center for High Performance Computing at Jackson State University. CaMEL is a highly scalable, incompressible, non-dimensional code. CaMEL is a hybrid finite volume/element solver, which takes advantage of the merits of both the Finite Volume and Finite Element methods and avoids their shortcomings. The buoyancy force was included in the momentum equation to capture the thermal effects while simulating the heat-transfer in the nine-generation model. A Grashof number of 1388 corresponding to the temperature difference of 47° C was utilized. Grashof number is dimensionless parameter which provides a ratio of buoyancy forces to the viscous forces. The bronchial tube

walls were assumed to be at the normal body temperature (37°C) and inhaled air temperature during cold weather condition (-10°C). At the inlet, a parabolic velocity and uniform temperature profiles were applied. No-slip condition with isothermal temperature was applied on the tube walls. At the exits, zero static pressure was specified. The nondimensional temperature is given by $T^* = (T - T_{wall}) / (T_{in} - T_{wall})$, where T is the temperature, T_{in} is the inlet temperature and T_{wall} is the bronchial tube wall temperature. A fully unstructured mesh was employed for discretization of the nine-generation bronchial tube model. The commercial software package Gridgen [26] was utilized to generate high quality mesh. The final mesh consisted of approximately 40 million tetrahedral elements.

The particle traces were simulated as a post-processing step using a Lagrangian method. Water droplets with a diameter of $10\mu\text{m}$ were released from the inlet of the model. Approximately 34000 particles were released at the inlet. The particles were released from the nodes of the uniform triangular mesh at the inlet. The initial velocities of the particles were kept the same as the inlet fluid velocities. Since impaction plays an important role for micro-particle transport, drag and gravitational forces were included in the equation of motion. The fourth-order Runge-Kutta method was used to integrate the equation of motion.

3 Results

The results of the flow fields and particle trajectory simulations for the nine-generation bronchial tube model with and without heat-transfer are shown and compared in this section. The comparison of the various metrics is made to investigate the thermal impact on the flow fields and particle deposition. The localized flows in second-generation in terms of primary and secondary velocities are shown in figure 1. A symmetric vortex pattern is observed in figure 1(b) due to the fact that the second-generation is symmetric with respect to the bifurcation plane. Figure 2 shows the primary and secondary velocities in the third-generation branching. In figure 2 (b) asymmetric pattern of vortices are observed since the third-generation branching is locally nonplanar. Now we focus on one of the eighth-generation branches. Primary and secondary flows are shown at a cross-section of one of the

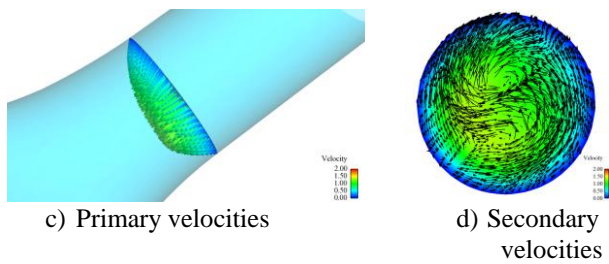


Figure 2. Flows in third-generation.

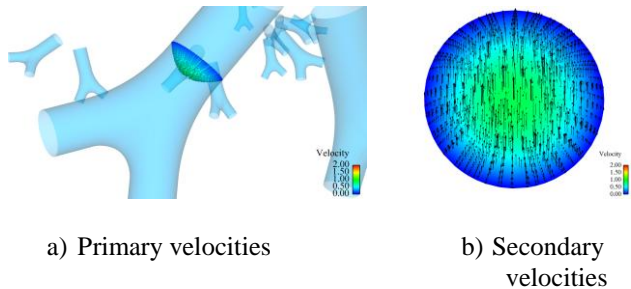


Figure 3. Flows in eighth-generation.

branches of the eighth-generation in figure 3. It can be observed here that, the secondary flows are not dominated by the vortices as the flow rate is not large enough to generate vortices.

In figure 4, dimensionless temperature variation at the cutting plane located in the first bifurcation is shown. The temperature profiles are also shown at various cross-sections perpendicular to the local axis of the tubes. Here, the dimensionless inlet temperature is one and at the bronchial tube walls it is zero. The temperature profiles are skewed towards the center of the bifurcation similar to the velocity profiles. This is because high velocity fluid carries the inlet temperature towards the center of the bronchial tubes.

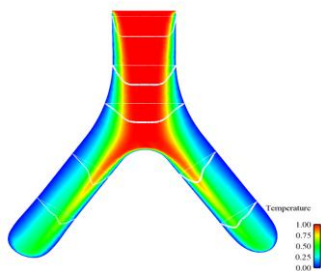


Figure 4. Temperature distribution in first bifurcation

Now to investigate the impact of heat-transfer on the bronchial tube flows, mass flow rate and secondary flow intensity comparisons are made as shown in Figure 5. Results are extracted at cross-sections located in generation-2, generation-3, and generation-8. As the number of generations increases, the tube diameter decreases and therefore the mass flow entering each branch also decreases as observed in Figure 5(a). The secondary velocities at these cross-sections can be quantified by measuring intensity of the secondary flows. The intensity of the secondary flows is defined as the ratio of the averaged local secondary velocities with respect to

the averaged local primary velocity in a given cross-section. The intensities of secondary velocities were also plotted at each cross-section location for both with and without heat-transfer cases. Figure 5(b) shows the variation of the intensity of secondary flows. The values of mass flow and intensity of secondary flows with and without heat-transfer show minimal differences (see figure 5).

To further assess the effects of heat-transfer, primary velocity profiles were plotted on the line segment passing through the middle of the cross-sections to compare the results for both cases. In Figure 6(a), the comparison between the primary velocities in simulation with and without heat-transfer cases in the second-generation is made. The dimensionless velocities along the length of the line segment are plotted. Similarly, velocity distributions in third- and eighth-generations are shown in figure 6(b) and 6(c), respectively. From figures 6(a)-(c) it can be observed that, the differences between the two velocity profiles for all three plots are insignificant. This suggests negligible thermal effects on local flow fields.

Since flows in the bronchial tube are unaffected by the temperature difference between inlet and tube walls, it can be indirectly implied that the particle deposition will also show minimal sensitivity to the temperature differences. However, we simulate particle trajectories to study particle deposition in order to explicitly investigate the heat-transfer effects on the particle deposition. We compare the particle deposition efficiencies in each generation for the simulation with and without heat-transfer in figure 7. Particle deposition efficiency is defined as the ratio of percent particle deposition to the incoming particles in each generation. As it was predicted, the thermal effects on particle deposition are minimal as the particle deposition efficiencies are quite close to each other for both cases. In general, the particle deposition efficiency increases as we go further down the generations, except in the eight-generation. The reason being, most of the particles entering the eighth generation are exited from the outlets and that result in to the low particle deposition efficiency.

Figure 8 shows particle deposition in nine-generation bronchial tubes in terms of particle destination and FTLE maps. Particle destination map shows the scalar values at particles' release location equal to the generation number it deposited to. Figure 8(a) shows the color map of the generations for particle destination map. Figures 8(b) and 8(c) show the particle destination and FTLE maps, respectively. For example, particles released from the dark blue region of the particle destination map shows that they deposit in the second generation. FTLE map shows the deposition behavior of closely seeded particles inside the generation. Higher FTLE values (red region) implies that the particles released from here are being more dispersed than the particles released at lower FTLE values (blue region). This phenomenon can be explained in detail from figure 9. We release particles from the high FTLE values (purple traces) and low FTLE values (magenta traces) and follow their path and how they are being affected by the geometry and secondary flows of the nine-generation bronchial tube as shown in figure 9(a). The particle

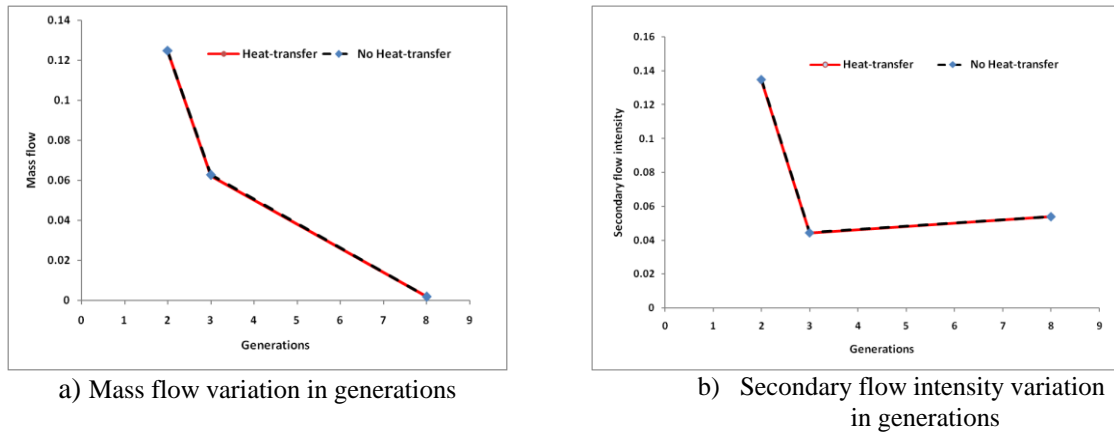


Figure 5. Mass flow and secondary flow intensity comparison.

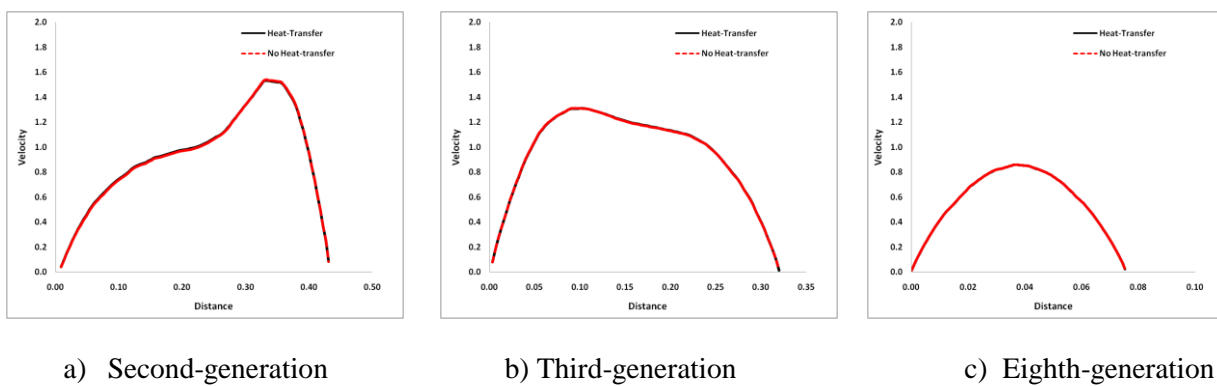
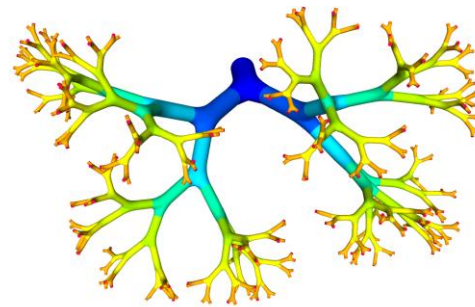


Figure 6. Comparison of dimensionless velocity profiles in generations-2, 3, and 8.

release locations are shown in detail in figure 9(b). In second generation, particles interact with the vortices here as shown in figure 9(c). Since purple particles are passing through the higher cross-flow velocity region, they get more affected by the vortex and get dispersed. Eventually, due to the combined effects of the vortex and geometry they diverge their paths and go to different tube after the first bifurcation. In the third-generation tube, the magenta particle traces are being more affected by one of the vortices and being dispersed (see figure 9(d)). However, the vortex is not strong enough to diverge the



a) Color map for destination map

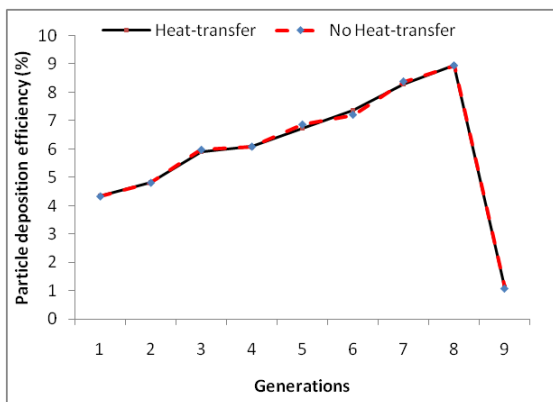
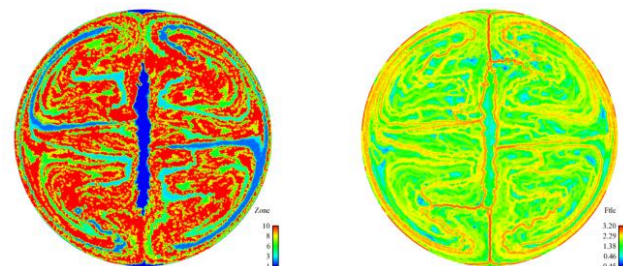
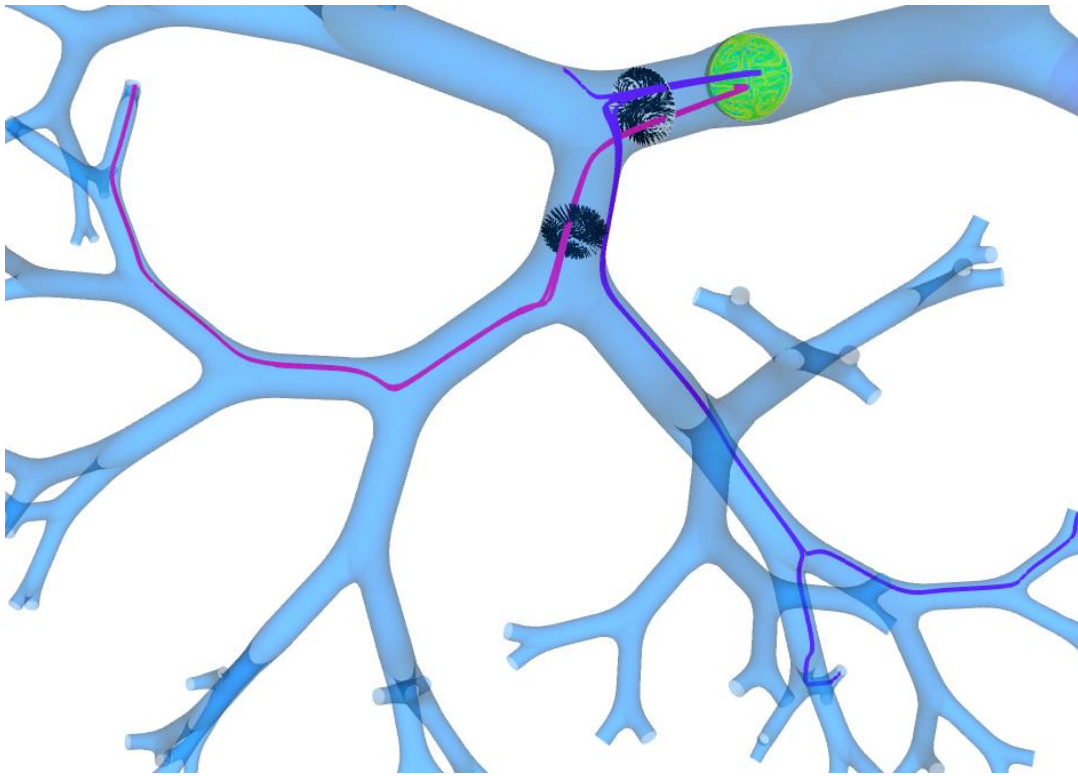


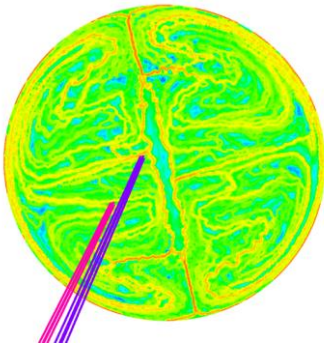
Figure 7. Comparison of particle deposition efficiency in each generation.



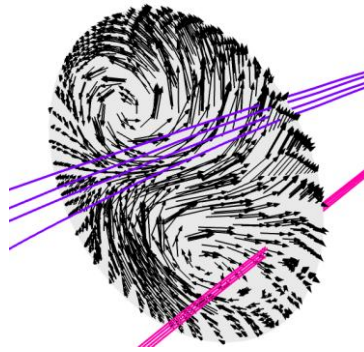
b) Particle destination map c) FTLE map
Figure 8. Particle deposition in bronchial tube model.



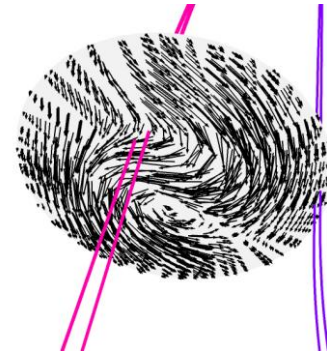
a) Particle traces in nine-generation bronchial tube model



b) FTLE map and particle release location



c) Vortices in the second-generation interacting with particles



d) Vortices in the third-generation interacting with particles

Figure 9. Particle traces in the nine-generation bronchial tube model.

particles' path into different next generation tubes. As a result, they still travel to the same next generation tube. After the third-generation, particles are being more affected by the nonplanar, multigenerational geometry than the vortices as the vortices get weaker with increasing number of generations.

4 Conclusions

Main objective of this paper is to understand and identify the importance of thermal effects of the inhaled air temperature on the flow and particle deposition in the small bronchial tubes. The results suggest that, the effects of the

differences between inhaled air temperature and lung tube wall temperature have little impact on the flow fields and particle deposition in the small bronchial tubes corresponding to the 4-12 generations of lung airway network. Therefore, simulating flow and particle transport with heat-transfer do not appear to be meaningful for the small bronchial tube models, based on this study.

We also try to understand the effects of bronchial tube geometry with multiple bifurcations on the flows and particle deposition. The results showed that particle transport in nine-generation bronchial tubes is mainly driven by the vortices and the nine-generation geometry. It was observed that the impact

of vortices is prominent when particles are in first few generations where the vortices are stronger compared to the ones in further generations. After that, the particle paths are influenced mainly by the nonplanarity of the bronchial tube geometry.

5 Acknowledgments

We would like to thank National Science Foundation (NSF) for funding this research through grant No. EPS-0903787. We would also like to thank Dr. Keith Walters and William Luke for providing us with the geometry of the nine-generation bronchial tube model.

6 References

- [1] American Society of Heating, R. and Engineers, A.-C., *Fundamentals Handbook*, Vol. Chapter 11: Air contaminants, ASHRAE, 1977.
- [2] Higenbottam, T., Siddons, T., and Demoncheaux, E., "The direct and indirect action of inhaled agents on the lung and its circulation: Lessons for clinical science," *Environmental Health Perspectives*, Vol. 109, No. 4, 2001, pp. 559–562.
- [3] Wang, X. and Christiani, D., "Respiratory symptoms and functional status in workers exposed to silica, asbestos, and coal mine dusts," *Journal of Occupational and Environmental Medicine*, Vol. 42, 2000, pp. 1076–1084.
- [4] Inglesby, T., O'Toole, T., and Henderson, D., "Anthrax as a biological weapon, 2002: updated recommendations for management," *Journal of American Medical Association*, Vol. 287, 2002, pp. 2236–2252.
- [5] Lane, H., Montagne, J., and Fauci, A., "Bioterrorism: a clear and present danger," *Nature Medicine*, Vol. 7, 2001, pp. 1271–1273.
- [6] Dockery, D., Pope, C., Xu, X., Spengler, J., Ware, J., Fay, M., Ferris, B., and Speizer, F., "An association between air pollution and mortality in six U.S. cities," *New England Journal of Medicine*, Vol. 329, 1993, pp. 1759–1759.
- [7] Toren, K., Bergdahl, I., Nilsson, T., and Jarvholm, B., "Occupational exposure to particulate air pollution and mortality due to ischaemic heart disease and cerebrovascular disease," *Occupational Environmental Medicine*, Vol. 64, 2007, pp. 515–519.
- [8] Peters, A., Dockery, D., Muller, J., and Mittleman, M., "Increased particulate air pollution and the triggering of myocardial infarction," *Circulation*, Vol. 103, 2001, pp. 2810–2815.
- [9] Pope, C., Muhlestein, J., May, H., Renlund, D., Anderson, J., and Horne, B., "Ischemic heart disease events triggered by short-term exposure to fine particulate air pollution," *Circulation*, Vol. 114, 2006, pp. 2443–2448.
- [10] Warrell, D. A., Cox, T. M., Firth, J. D., and Benz, E. J., *Oxford Textbook of Medicine*, Vol. 2, Oxford University Press, 2003.
- [11] Soni, B., Lindley, C., and Thompson, D., "The Combined Effects of Nonplanarity and Asymmetry on Primary and Secondary Flows in the Small Bronchial Tubes," *International Journal for Numerical Methods in Fluids*, Vol. 59, 2009, pp. 117-146.
- [12] Soni, B., Thompson, D., and Machiraju, R., "Visualizing Particle/Flow Structure Interactions in the Small Bronchial Tubes," *IEEE Transactions on Visualization and Computer Graphics (Proceedings of Visualization/Information Visualization 2008)*, Vol. 14, No. 6, 2008, pp. 1412-1419.
- [13] Gatlin, B., Cuichhi, C., Hammersley, J., Olson, D., R.Reddy, and Burnside, G., "Computation of converging and diverging flow through an asymmetric tubular bifurcation," *ASME FEDSM97*, Vol. 3429, 1997, pp. 1–7.
- [14] Haller, G., "Distinguished material surfaces and coherent structures in threedimensional fluid flows," *Physica D*, Vol. 149, 2001, pp. 248–277.
- [15] Nowak, N., Kadake, P., and Annapragada, A., "Computational fluid dynamics simulation of airflow and aerosol deposition in human lungs," *Annals of Biomedical Engineering*, Vol. 31, 2003, pp. 374–390.
- [16] Ertbruggen, C., Hirsch, C., and Paiva, M., "Anatomically based three-dimensional model of airways to simulate flow and particle transport using computational fluid dynamics," *Journal of Applied Physiology*, Vol. 98, 2004, pp. 970–980.
- [17] Gemci, T., Ponyavin, V., Chen, Y., and Collins, R., "Computational model of airflow in upper 17 generations of human respiratory tract," *Journal of Biomechanics*, Vol. 41, 2008, pp. 2047–2054.
- [18] Schmidt, A., Zidowitz, S., A.Kriete, Denhard, T., Krass, S., and Pietgen, H.-O., "A digital reference model of the human bronchial tree," *Computational Medical Imaging and Graphics*, Vol. 28, 2004, pp. 719–723.
- [19] Walters, K., and Luke, W., "A Method for Three-Dimensional Navier-Stokes Simulations of Large-Scale Regions of the Human Lung Airway", *ASME Journal of Fluids Engineering*, Vol. 132, 2010, Paper No. 051101.

[20] Aref'ev, K., Fedotov, E., and Khrushchenko A., "Nonstationary Heat Exchange in the Trachea of Human Lungs", *Journal of Engineering Physics and Thermophysics*, Vol. 76, No. 4, 2003, pp. 892-898.

[21] Serikov, V., Fleming, N., Talalov, V., and Stawitcke, F., "Effects of the Ventilation Pattern and Pulmonary Blood Flow on Lung Heat Transfer", *European Journal of Applied Physiology*, Vol. 91, 2004, pp. 314-323.

[22] Zhang, Z., and Kleinstreuer, C., "Species Heat and Mass Transfer in a Human Upper Airway Model", *International Journal of Heat and Mass Transfer*, Vol. 46, 2003, pp. 4755-4768.

[23] Zhang, Z., Kleinstreuer, C., and Kim, C., "Water Vapor Transport and its Effects on the Deposition of Hygroscopic Droplets in a Human Upper Airway Model", *Aerosol Science and Technology*, Vol. 40, 2006, pp. 1-16.

[24] Weibel, E., "Morphometry of the human lung," 1963.

[25] Tu, S., and Aliabadi, S., "Development of a hybrid finite volume/element solver for incompressible flows", *International Journal of Numerical Methods in Fluids*. 2007; 55:177-203.

[26] Gridgen, Grid and Mesh generation for Computational Fluid Dynamics (CFD), Software Package, Ver. 15, Fort Worth, TX, 2010.

A Simple Nonadditive Model of Water

Gregory G. Wood¹

¹Department of Mathematics and Physics, California State University Channel Islands, Camarillo, CA, USA

Abstract—*Liquid water has rich thermodynamic behavior over a range of temperatures and pressures[1]. Models of water used in protein folding simulations must be fast and reflect the underlying hydrogen bond network accurately. Although greatly simplified, current models of water can account for more than 99% of the CPU time during numerical simulations[2]. Current models assume simple additivity of free energy which is incorrect over coupled degrees of freedom[3], [4] - and in the liquid state all of the water is coupled. A novel statistical mechanical model of water is presented encapsulating the essential nonadditive free energies without recourse to computationally expensive techniques.*

Keywords: model of water, thermodynamic properties, statistical mechanics, monte carlo simulation

1. Introduction

Liquid water is a system of great significance for the study of biomolecules[5], [6]. Although pure liquid water is a poor analog of the aqueous environments found in cells, in nature, it is nonetheless widely used in both experiment and numerical simulation. Even simple models of water are computationally expensive with orders of magnitude more CPU time spent on the water rather than the biomolecule itself, in typical simulations[2]. Typically one to a few layers of molecules are modeled around the biomolecule and beyond that bulk water is described with a continuum model. The water molecules are modeled as a set of points, with fixed distances between each, and interact with one another and with the biomolecule by a set of energy functions[7], [8]. Among other, simplified, models of water are treating water as two dimensional disks[9] and the Mercedes-Benz model[10]. Simple models, such as the Ising model in magnetism, assist fundamental understanding without employing complex or computationally expensive mathematics.

The first step in many models of liquid water for use with biomolecules is to model pure water and reproduce physical properties of liquid water[7], [8]. For the purposes of this work, the density and specific heat of liquid water over a broad range of temperatures and pressures is employed.

Instead of using continuous energy functions, a new model where the bond between adjacent molecules can be classified in one of a discrete set of states is presented. The character of each bond is given by an average length, energy and entropy. Further model parameters employed give rise to long range interactions: (A) a strain energy parameter which adds a

small energy term if two parallel bonds across a single unit cell are of dissimilar length and (B) an extra bond entropy term which adds extra entropy to atoms which more than one high-entropy bond adjoining them.

Specific heat at constant pressure, c_p , and density, ρ , are computed by standard statistical methods[11] from the partition function. An average value, \bar{x} , such as the average energy \bar{E} , required above, or the average bond length from which the density can be computed, is found via a sum over all states i as follows:

$$\bar{x} = \frac{\sum x_i \exp(-G_i/k_B T)}{\sum \exp(-G_i/k_B T)}, \quad (1)$$

where k_B is Boltzmann's constant and T is the temperature. The Gibbs free energy, G_i , is detailed below.

To compute the specific heat at constant pressure, the average energy of the system is found via Eq. 1, above, and the appropriate derivative is taken at fixed pressure as follows:

$$c_p = \left. \frac{\partial \bar{E}}{\partial T} \right|_p. \quad (2)$$

However, the complete sum over all states of the system is daunting for even modest size systems: the number of states is 3^N . Most of these states are highly unfavorable. Further, most are very similar to a very large number of other states. For small systems, the exact results of sums over all states is compared with approximations of summing over all types of states, as detailed below and checked for consistency.

Having three bond lengths implies the oxygen to oxygen distance distribution would be a collection of delta functions. However, these should be regarded as the centroids of gaussian-like distributions of possible bond lengths. To illustrate this, using the Heisenberg uncertainty relation with the equipartition of energy, it is possible to estimate a lower limit on the size of the width of such gaussians, in angstroms, as a function of temperature. At room temperature, 300 Kelvin, this is about 0.05 angstrom, which is about six times smaller than the experimental width of the nearest neighbor peak of the oxygen-oxygen distribution function[12], [13]. The details of this calculation are reproduced in appendix A, below.

2. Contributions to the Free Energy

The Gibbs Free Energy of some state i of the system is given by, $G_i = E_i + PV_i - TS_i$, where E is the energy, P

the pressure, T the temperature, V the volume, and S the entropy. The energy, volume and entropy of the states are adjustable model parameters of each type of bond modeled, thus for three states nine parameters are required. However the zero of both energy and entropy are arbitrary, reducing the number of parameters to seven.

The three states are referred to as short, medium and long in this article, reflecting the rank ordering of their bond length parameters. The short bond has the lowest entropy and energy and the long bond has the highest. Thus the low temperature anomalies of water[14] are not studied currently within this model, since that would require one or more extra states which violate these rules. The goal of this work is to show that most of the temperature and pressure phase space can be modeled well with a limited model.

In addition to straightforward local statistical mechanics, in which only energy and entropy are required, long range interactions are invoked to model strain energy and the localization of bonds by neighbors. These are termed nonadditive interactions as they are extra terms on top of the regular summation of energy and entropies. The two terms lead to quite different effects.

The first is the strain energy for parallel mismatched bonds. A small energy penalty is added to the state energy E_i , for each mismatched pair of bonds.

To give all bonds a physical location, all bonds are placed in an idealized hexagonal lattice and neighboring bonds, and the six nearest parallel bonds, are identified for each. This exercise was carried out by hand for 128 atoms in a 4x4x8 lattice and used to construct computer code which identifies neighboring bonds and parallel bonds for lattices of arbitrary size. Periodic boundary conditions are employed.

The second term is the extra bond entropy. A small entropy term is added when two or more non-ground state bonds meet at an oxygen atom. Six parameters are employed to account for two, three or four of either the medium or long bonds meeting at an oxygen. What about mixed states? A long bond can be treated as a medium bond if it will increase the extra bond entropy. For example, if one short, one medium and two long bonds meet at an oxygen, the larger of the extra bond entropy due to two long or three medium bonds will be used. In future work, a single parameter controlling the strength of these interactions will be employed from which all possible combinations of bond entropies will be derived. There is no unique way to determine such a parameter and many possibilities are being considered. Further, due to the tiny magnitudes of some of these parameters, they may be dropped altogether and only four, or perhaps three or four, higher entropy bonds meeting at an oxygen will warrant consideration for extra bond entropy. However, this is of great interest when increasing the number of states, or generalizing this model in any way. For the moment, only rank ordering of the six parameters is enforced such that: the extra bond entropy must increase

as the number of non-ground state bonds increases and (b) the extra bond entropy is greater for an equal number of long bonds over medium bonds. The total number of free parameters is 14, although as will be seen below, some of these parameters are quite tiny - four orders of magnitude smaller than the bond entropies.

3. Scaling the System

Even for the 128 atom system described above (256 bonds) the number of terms in the partition function is $3^{256} \approx 10^{122}$. Since this would take longer than the age of the universe to compute on all CPUs in existence, some kind of approximation must be made. First a very small system was created, a six atom oxygen ring with 18 bonds for which the partition function could be computed exactly and compared with various methods of sampling. The method arrived at is to sample each combination of numbers of short, medium and long bonds some large number of times, weighing each triplet of numbers (e.g. the number of short, medium, long bonds) by the appropriate multinomial coefficient. This is checked to ensure the two results are consistent and then the sampling algorithm is scaled to the larger systems. All possible triplets of number of types of bonds are sampled, although for large enough systems, sampling over various fractions of each could replace the numbers (1% short bonds, 3% medium bonds and 96% long bonds would be a sampled state instead of, say, 2 short 4 medium and 250 long bonds, for example).

4. Results

Three different pressures are considered, 0.013 MPa, 400 MPa and 1000 MPa. The lowest and highest pressures available from a thorough set of experiments[1] were employed and the middle number was chosen such that the density of the 400 MPa data should fall about half way between the two extremes. With only the three states considered and the fourteen adjustable parameters, excellent agreement to either density or specific heat and adequate fits to the other physical property are possible. In the data below, the closer fit is to density, as seen in Fig. 2 below. The excellent agreement with density across temperature and pressure is superior to the results of the TIP-5p model[7]. The TIP-5p model, along with the SPC/E[8] model are the most commonly employed models for molecular dynamics[7].

The effect of introducing the extra bond entropy and the strain is to improve the goodness of fit (chi-squared) by a factor of 3.5 - despite many of the extra bond entropy parameters being very small, see Table 1 below.

5. Conclusions

A simple three state model of liquid water yields excellent agreement with experimental density measurements, and adequate agreement with experimental specific heat data

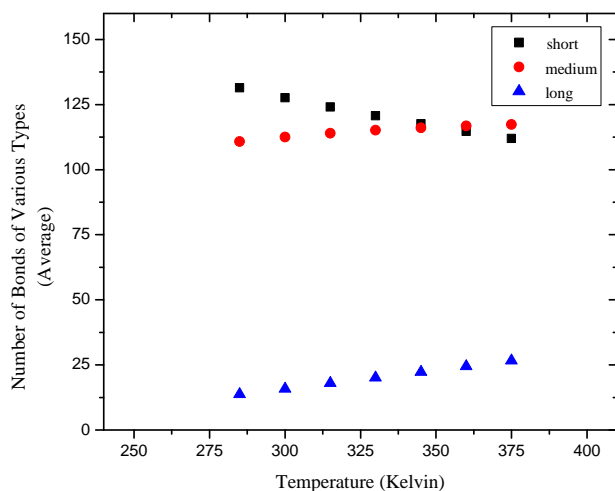


Fig. 1: Number of long, medium and short bonds at medium (400 MPa) pressure as a function of temperature. Note the crossover from a plurality of short to medium bonds at a temperature near 350 K. At higher pressure, the short bond state is more highly favored, with both the medium and long bonds suppressed. At low pressure, the longer bond is significantly more favored at the expense of the short bond, with about equal propensity for medium length bonds.

over a broad range of temperatures (273-373 kelvin) and pressures (0.013 MPa - 1000 MPa). The model employs two novel nonadditive terms: a (generally small) extra entropy term added at the intersection of multiple high-entropy bonds and a strain energy term for mismatched parallel bonds. Despite some terms being small, these nonadditive terms improve the goodness of fit by a factor of three and a half.

6. Future Work

A study of physical properties of small molecules in water shall be employed to determine similar parameters for a small set (as small as possible) between water and various atoms of biological relevance. Such a set of parameters would then be employed to model the water around and between proteins for the purposes of protein folding, drug design and docking. By replacing traditional means, either the run-time of simulations can be greatly reduced or the quantity of water modeled around proteins greatly increased.

7. Acknowledgements

The author would like to thank Brian W. Gilreath for assistance visualizing the structure of the tetragonal structure of ice and many useful discussions.

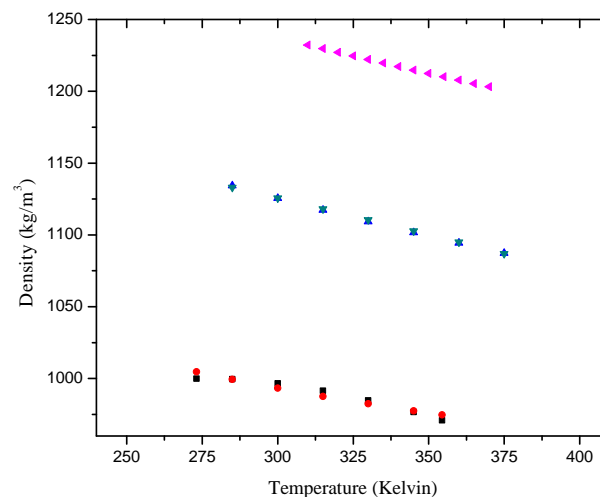


Fig. 2: Density in units of kilograms per cubic meter versus temperature in kelvin for liquid water at three pressures: low (0.013 MPa), medium (400 MPa) and high (1000 MPa). The higher pressure result in correspondingly higher densities. At each pressure, the density is nearly linear in temperature with a small negative slope. The calculated values are represented by squares (at low pressure), upright triangles (at medium pressure) and left pointing triangles (at high pressure). Agreement is within the size of the symbols at high and medium pressure and only slight differences exist at the lowest pressure, the largest of which is an 0.5% difference at the lowest temperature.

8. Appendix A

The Heisenberg uncertainty relation is that the product of the uncertainties in position and momentum must exceed half the rationalized Plank's constant. Applying this to the axis of the bond, for convenience labeled the x -axis, gives:

$$\Delta x \Delta p_x > \hbar/2. \quad (3)$$

The x -momentum is the product of mass times velocity (since any velocity here is far below the speed of light thus relativistic effects are negligible). All that is needed is a relation between momentum and position to give a lower bound on the uncertainty in position. There is a relation in energy, but to employ it, a specific form of potential (binding) energy is required. For this purpose, an approximation is introduced: that the bond acts like a single harmonic oscillator and thus the potential energy is given by $\frac{1}{2}kx^2$, where k is the spring constant and x is the displacement from equilibrium position. More complex relations such as the Lennard-Jones (6-12) potential can be considered, but these are well approximated by a harmonic oscillator when the energy is far below the dissociation

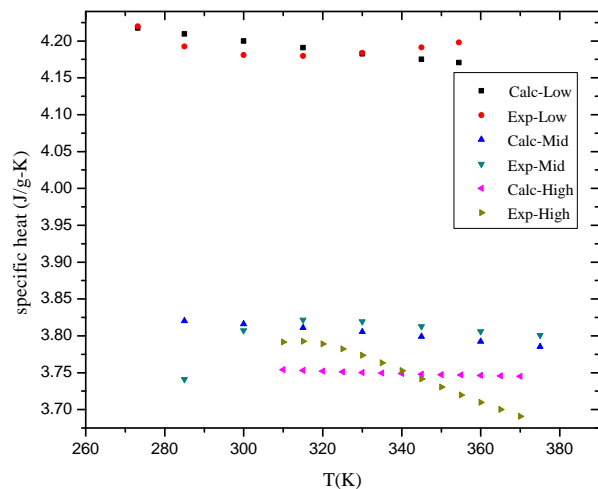


Fig. 3: Specific heat at constant pressure of liquid water in units of joules per gram-kelvin versus temperature in kelvin for three different pressures: low (0.013 MPa), medium (400 MPa) and high (1000 MPa). With only three states, the calculated data reflects the correct magnitudes in specific heat but is unable to account for the more abrupt changes at low temperature.

energy, which should be the case in question (liquid water at room temperature).

The equipartition of energy theorem requires each the degree of freedom to have the same energy as given below:

$$\frac{1}{2}k_B T = \frac{p_x^2}{2m} = \frac{1}{2}kx^2. \quad (4)$$

Combining these relations by setting $\Delta x = x$ and $\Delta p_x = p_x$, we find:

$$x > \frac{\hbar}{2(mk_B T)^{0.5}} \quad (5)$$

and

$$k^2 < k_B T / \hbar. \quad (6)$$

Plugging in known values at room temperature, the minimum value of x is about 0.05 angstroms, which is about six times smaller than the experimental width from the radial distribution function. It is worth noting that for a single, isolated harmonic oscillator, the wavefunctions can be found exactly along with the uncertainties in both position and momentum. In that case, equality in the Heisenberg relation holds, meaning the wavefunctions are the “tightest” possible. It is not surprising that a fluid cannot be modeled as a collection of isolated single harmonic oscillators.

Table 1: Model parameters for short, medium and long bonds of liquid water. All enthalpies, ΔH , have units of kcal/mol and all entropies are dimensionless “pure” entropies (to produce entropies in the proper units, these values need only be multiplied by Boltzmann’s constant in the appropriate units). The bond lengths, $\langle x \rangle$, are in angstroms and the extra bond entropies xbs , are also unitless. The energy, entropy of the short bond is set to zero without loss of generality[11]. The subscript of the extra bond entropy parameter denotes the number of such bonds meeting at a particular oxygen atom. The strain energy is 2.22×10^{-4} kcal/mol-K. Although fourteen free parameters are employed, three of the six extra bond entropy parameters are very small - three or four orders of magnitude less than the typical change in entropy from state to state of a single bond.

Parameter	long	medium	short
ΔH	3.25e-3	6.21e-4	0
ΔS	4.47e-3	2.89e-3	0
$\langle x \rangle$	3.37	3.01	2.71
xbs_2	7.32e-7	6.83e-7	0
xbs_3	3.84e-5	5.36e-6	0
xbs_4	1.51e-4	1.83e-5	0

References

- [1] Wagner and Pruβ, "The IAPWS Formulation 1995 for the Thermodynamic Properties of Ordinary Water Substance for General and Scientific Use," *J. Chem. Ref. Data*, vol. 31, pp. 387–535, 2002.
- [2] D. Qiu, P. S. Shenkin, F. P. Hollinger and W. C. Still, "The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii", *J. Phys. Chem. A*, vol. 101, pp. 3005-3014, 1997.
- [3] K. A. Dill "Additivity Principles in Biochemistry", *J. Biol. Chem.*, vol. 272, pp. 701–704, 1997.
- [4] A. E. Mark, and W. F. van Gunsteren, "Decomposition of the Free Energy of a System in Terms of Specific Interactions." *J. Mol. Biol.*, vol. 240, pp. 167–176, 1994.
- [5] K. A. Dill, T. M. Truskett, V. Vlachy, and B. Hrivar-Lee, "Modeling water, the hydrophobic effect, and ion solvation", *Annu. Rev. Biophys. Biomol. Struct.*, vol. 34, pp. 173–199, 2005.
- [6] M. J. Tait, and F. Franks, "Water in Biological Systems", *Nature*, vol. 230, pp. 91–94, 1971.
- [7] M. W. Mahoney, and W. L. Jorgensen, "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions", *J. Chem. Phys.*, vol. 112, pp. 8910–8922, May 2000.
- [8] H. J. C. Berendsen, J. R. Grigera, and T. P. Staatsma, "The missing term in effective pair potentials", *J. Phys. Chem.*, vol. 91, pp. 6269–6271, 1987.
- [9] T. M. Truskett, and K. A. Dill, "A Simple Statistical Mechanical Model of Water", *J. Phys. Chem.*, vol. 106, pp. 11829–11842, 2002.
- [10] T. Urbic, T. Vlachy, and K. A. Dill, "Confined Water: A Mercedes-Benz Model Study", *J. Phys. Chem. B.*, vol. 110, pp. 4963–4970, Feb. 2006.
- [11] F. Reif, *Fundamentals of Statistical and Thermal Physics*, 1st ed., New York, USA: McGraw Hill, 1965.
- [12] A. K. Soper, and M. G. Phillips, "A new determination of the structure of water at 25 °C", *Chem. Phys.*, vol. 107, pp. 47–60, Aug. 1986.
- [13] A. K. Soper, "On the determination of the pair correlation function from liquid structure factor measurements", *Chem. Phys.*, vol. 107, pp. 61–74, Aug. 1986.
- [14] C. H. Cho, S. Singh, and G. W. Robinson, "Water anomalies and the double-well Takahashi model", *Chem. Phys.*, vol. 232, pp. 329–341, June 1998.

MODELING AS A TOOL FOR CONTROLLING THE PRODUCTION OF BIOFUELS: ETHANOL FROM A BIOMASS

Aghareed M. Tayeb, N. A. Mostafa and H . A. Ashour

Faculty of Engineering, Minia University, Minia, Egypt

Abstract:

In recent years, the new trend is towards bio-fuel. One of these biofuels is ethanol (ethyl alcohol), which could be produced economically by the controlled fermentation of biomasses. The results of the alcoholic fermentation of beet sugar molasses and wheat milling residues (Akalona) were fed into a computer program. The kinetic parameters for these fermentation reactions were determined. These parameters were put into a kinetic model. Next, the model was tested, and the results obtained were compared with the experimental results of both beet molasses and Akalona. The deviation of the experimental results from the results obtained from the model was determined. An acceptable deviation of 1.2% for beet sugar molasses and 3.69% for Akalona was obtained. Thus, the present model could be a tool for chemical engineers working in fermentation processes both with respect to the control of the process and the design of the fermenter.

Keywords: Modeling, computation, biofuel, computer program, alcoholic fermentation.

Nomenclature:

n_s	= Substrate utilization coefficient
p	= Product concentration (kg/m ³)
P_0	= Initial product concentration (kg/m ³)
P_m	= Ethanol concentration above which cells do not grow (kg/m ³)
P^*_m	= Ethanol concentration above which cells do not produce ethanol (kg/m ³)
Δp	= Concentration driving force (kg/m ³)
q_p	= Specific ethanol production rate (g product/g cell/h)
r_s	= Reaction rate
S	= Substrate concentration (kg/m ³)
S_0	= Initial substrate concentration (kg/m ³)
S_m	= Substrate concentration calculated from the model (kg/m ³)
t	= Time (h)
V_0	= Maximum specific rate of ethanol production rate at zero ethanol concentration (g ethanol/g substrate/h)
X	= Biomass concentration (kg/m ³)
α	= Growth associated constant (g ethanol/g cell)

β	= Non-growth associated constant (g ethanol/g cell/h)
μ	= Specific growth rate of cells (h ⁻¹)
μ_0, μ_1	= Specific growth rate of cells in the presence of ethanol (h ⁻¹)
μ_{max}	= Maximum specific growth rate of cells (h ⁻¹)
\sim	= Kinetic parameter in Bovee model

Main nomenclature for the computer program:

FOPTIM	= Subroutine that defines the objective function
SMIN	= Subroutine that finds the minimum value for the objective function

1. Introduction:

Biofuels are a wide range of fuels which are in some way derived from biomass. The term covers solid biomass, liquid fuels and various biogases [1]. Biofuels are gaining increased public and scientific attention, driven by factors such as oil price spikes, the need for increased energy security, concern over greenhouse gas emissions from fossil fuels, and government subsidies.

Bioethanol is an alcohol made by fermenting the sugar components of plant materials and it is made mostly from sugar and starch crops. With advanced technology being developed, cellulosic biomass, such as trees and grasses, are also used as feedstock for ethanol production. Ethanol can be used as a fuel for vehicles in its pure form, but it is usually used as a gasoline additive to increase octane and improve vehicle emissions. Bioethanol is widely used in the USA and in Brazil.

Biofuels provided 1.8% of the world's transport fuel in 2008. Investment into biofuels production capacity exceeded \$4 billion worldwide in 2007 and is growing [2].

In the simulation of chemical and biochemical processes, the prediction of data has a dominant importance. The success or failure of this calculation depends on the use of a favorable mathematical model and upon reliable experimental data obtained in industry. Further, the optimal

automatic bioreactor control requires a mathematical model adapted to the potency of reliable sensors.

James F. Bartes, et al [3] provided a nonlinear predictive integrating temperature model for a fermentation process. The model specifies or represents relationships between attributes or variables related to the temperature of the fermentation process, including relationships between inputs to the fermentation process and resulting outputs of the fermentation process. The nonlinear predictive integrating temperature model may be based on heat balance of the fermentation process, including a balance between available cooling and current fermentation heat generation. The model variables may also include aspects or attributes of other processes or sub-processes that have bearing on or that influence operations of the fermentation process.

In biochemical processes, the mathematical model is a relationship describing the kinetic behavior which relates the biological rate of substrate consumption to substrate and product concentrations. The model has several parameters that can be estimated by fitting them to the experimental data.

The decrease in growth rate and the cessation of growth due to the depletion of substrate may be described by the relationship between μ and the residual growth limiting substrate, represented in the following equation [4]:

$$\mu = \frac{\mu_{\max} S}{K_s + S} \quad (1)$$

K_s is numerically equal to the substrate concentration when μ is one-half μ_{\max} and is a measure of the affinity of the organisms. The formation of a growth-linked product may be described by the equation:

$$dp/dt = q_p x \quad (2)$$

$y_{p/x}$ is the yield of product in terms of substrate consumed ($y_{p/x} = dp/dx$).

Combining the above two equations:

$$q_p = y_{p/x} \mu \quad (3)$$

The relationship between the specific ethanol production rate and the specific growth rate of cells can be represented by the following equation [5]:

$$q_p = (\alpha\mu) + \beta \quad (4)$$

The constants α and β are 2.2-2.9 g ethanol/g cell and 0.25-0.5 g ethanol/g cell/h, respectively. The data show that the overall good ethanol production rate was mainly contributed by the high specific growth rate.

Two other kinetic models were also proposed to describe the kinetic pattern of ethanol inhibition on the specific rates of growth and ethanol fermentation [6]:

$$\mu_1 / \mu_0 = 1 - (P/P_m)^\alpha \quad (\text{for growth}) \quad (5)$$

$$v_i / v_o = 1 - (p/P_m)^\beta \quad (\text{for ethanol production}). \quad (6)$$

The maximum allowable ethanol concentration above which cells do not grow was predicted to be 112 g/l. The ethanol-producing capability of the cells was completely inhibited at 115 g/l ethanol. On the other hand, there was a threshold concentration of ethanol (26 g/l) below which there was no inhibition.

At a high value of α ($\alpha > 3$), the inhibitory effect of ethanol was less pronounced, the ratio μ_1/μ_0 remained almost unchanged (close to unity) even though p/p_m increased from 0 to 0.3.

This kinetic model seemed to be useful for representing the kinetics of alcohol fermentation. The model parameters (α , β , p_m and p'_m) depend on the microbial species, the physiological conditions of the micro organism and the status of the culture medium.

Four types of dependence of μ_1 on the ethanol concentration p are as follows:

(1) **Linear relationship:**

$$\mu_1 = \mu_0 - k_1 p = \mu_0 (1 - p/p_m) \quad (7)$$

where k_1 is an empirical constant.

The above relationship was found to fit the kinetics of cellulose hydrolyzate to ethanol by *Saccharomyces cerevisiae*.

(2) **Exponential relationship:**

$$\mu_1 = \mu_0 \exp(-k_2 P) \quad (8)$$

where k_2 is an empirical constant which depends on the method of cultivation (batch or continuous) (dimension 1/g).

(3) **Hyperbolic relationship:**

$$\mu_1 = \mu_0 \frac{1}{1 + P/k_3} \quad (9)$$

where k_3 is a constant (g/l).

(4) **Parabolic relationship:**

$$\mu_1 = \mu_0 (1 - p/p_m)^{0.5} \quad (10)$$

or

$$\mu_1 = \mu_0 - (\alpha p / b - p) \quad (11)$$

At similar p ($b - p = b$), the relationship becomes linear.

A generalized non-linear equation is:

$$\mu_1 = \mu_0 (1 - p/p_m)^n \quad (12)$$

From the literature

$$P_m = 68 \text{ g/l}, p'm = 112 \text{ g/l}, \text{ or } P_m = 92.7 \text{ g/l}, p'm = 114.5 \text{ g/l}$$

The maximum specific growth rate (μ_{\max}) could be calculated using experimental data for the exponential growth phase according to the definition:

$$\mu = 1/t \ln[(X_i + 1)/X_i] \quad (\text{h}^{-1}) \quad (13)$$

The values of μ_{\max} were determined using linear regression analysis upon the experimental growth curves.

$$t_d = 0.693/\mu_{\max} \quad (14)$$

$$Y_{x/s} = dX / -ds \quad (15)$$

$Y_{x/s}$ = biomass yield coefficient from the sugar utilized.

$$Y_{p/s} = dp / -ds \quad (\text{g/g}) \quad (16)$$

$$y_{p/x} = Y_{p/s} / Y_{x/s} \quad (17)$$

$y_{p/x}$ = ethanol yield coefficient with respect to biomass formed.

The values of $Y_{x/s}$ and $Y_{p/s}$ were calculated from experimental data using linear regression analysis. The conversion yield Y (% of theoretical) was calculated from the relationship:

$$Y = Y_{p/s} / 0.538 \quad (18)$$

where 0.538 is the theoretical ethanol yield coefficient for the sucrose or glucose consumed.

The productivity of fermentation was calculated from:

$$P_r = \frac{P_{\max} - P_0}{\text{time to obtain } p} \quad (\text{g/l h}) \quad (19)$$

A particular test [7] was performed to determine the alcoholic inhibition constant in the reaction kinetic model. It was deduced that the alcohol concentration had no substantially different effect on the metabolic activity of the

immobilized cells as opposed to free ones. To evaluate the substrate utilization coefficient, n_s , experimental measurements of the amount of substrate consumed, ΔS , and ethanol produced, ΔP , in the reactor were carried out and substituted in the form:

$$-\Delta S = n_s \Delta P \quad (20)$$

2. Selection of the kinetic model:

A relationship describing the kinetic behavior of alcoholic fermentation was investigated by Bovee [8] in the form:

$$r_s = dS/dt = k S^\alpha p^\beta \quad (21)$$

Using the yield relation between product and substrate, it is possible to describe, in both batch and continuous cultures, the ethanol and sugar concentration versus time. This pattern has been successfully tested on several fermentations performed by yeasts, including *Saccharomyces Cerevisiae* used in the experimental part of the present work, and a bacterium.

This simple relationship is proposed as a tool for process control alcoholic fermentation. Parameters α and β were correlated to the activation or inhibition effects of the substrate and product. Parameter k increases with the initial sugar concentration.

The constraint of this model is:

$$p = -(S_0 - S) + P_0 \quad (22)$$

A flexible digital computer program, SUGAR, was developed in the present work, to fit the model's parameters to the experimental data, by minimizing the following objective function which was proposed by Bovee [8].

$$Q = I / N^2 \sum_{i=1}^N [(S_{i\text{exp}} - S_m)^2 + (P_{i\text{exp}} - P_m)^2] \quad (23)$$

where $S_{i\text{exp}}$ and $P_{i\text{exp}}$ are the experimental values of substrate and product, and S_m and P_m are the values calculated by the model. The parameters obtained can then be used for the calculations needed to design bioreactors.

3. Program "SUGAR" for kinetic calculations:

The program "SUGAR" is written in FORTRAN-77 code for the VAX II computer with a DEC version 4.5 operating system. "SUGAR" consists of the main program,

four subprograms and one minimization routine "SMIN". The flow diagram of SUGAR is shown in Fig. 1.

The input data consists of the experimental data of substrate and product concentrations, time and number of data sets. The parameters k , α and β are now calculated by minimizing the objective function. The substrate concentration is calculated by using the Runge - Kutta method. The input data to the program are the experimental results of N. A. Mostafa [9] and are given in Fig. 2 for one run. The output data of the program are the values of the computed parameters α , β and k and the calculated data and the deviation between experimental and calculated data. These outputs are given in Fig. 3.

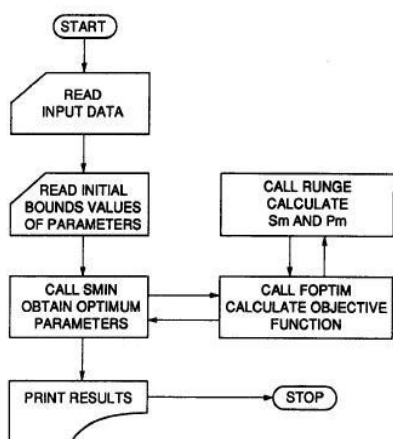


Fig. 1: Flow diagram of "SUGAR" program

9	129.166	000.350	0.2187
003.0	125.000	001.530	
024.0	100.000	007.040	
048.5	046.450	019.350	
096.0	012.220	022,880	
120.0	010.500	026.300	
123.0	007.600	023.900	
144.0	004.270	022.500	
147.0	004.650	025.000	
168.0	004.090	022.800	

Fig. 2: Input data to the program

4. Results and discussion:

The model so far reached is satisfactory enough when compared to other models [8]. The obtained results

from the model, as shown from Figs. 4 and 5, can be evaluated as follows:

(1) For beet sugar molasses:

The model satisfies the experimental results of beet sugar molasses with a value for the standard deviation (objective function) of 1.2. This value is to be compared with the value of the Bovee model [8] which showed the range of 0-1. This difference between the two values of the deviation ranges may be due to:

- (a) Bovee's work was based on pure glucose, an ideal substance for the kinetic study, whereas molasses, on which the present work is based, is a non-pure residue and is expected not to give as ideal results as given by pure glucose.
- (b) In Bovee's model, the effect of the yeast produced is not taken into consideration because it is assumed to be low. On the contrary, this is the condition of the present work where the experimental results indicated that the used *S. Cerevisiae* grows rapidly, giving a high cell density compared to other yeasts. Thus, it affects the results.

(2) For Akalona hydrolyzate:

Applying the model on the results of Akalona hydrolyzate gave a deviation value of 3.69 compared to 1.2 for beet sugar molasses. This may be explained as follows:

- (a) Molasses fermentation gives rise to mainly one sugar (sucrose) but Akalona hydrolyzate contains many sugars, as indicated by the analysis of Akalona hydrolyzate and by the literature [9]. This may be a reason for the deviation of the error range for Akalona hydrolyzate from its value for molasses.
- (b) Akalona hydrolyzate contains strange substances due to acid hydrolysis of the cellulosic content [10,11], which have an inhibitory effect on the yeast strain (*S. Cerevisiae*). The degree of substrate inhibition was found to be higher for bagasse hydrolyzate reported for ethanol fermentation of pure sugar. This, in turn, affects the value of the kinetic parameters, thus leading to a higher value for deviation.
- (c) As mentioned for beet sugar molasses, the relatively large amount of yeast produced affects the value of the standard error.

	ALFA = 0.000500000024			BETA = 0.004067549016			KAPPA = 1.5000000	
I	T	SE	S	DELS	PE	P	DEL P	
1	3.0	125.00000	125.79620	-0.79620	1.53000	1.08898	0.44302	
2	24.0	100.00000	98.56013	1.43987	7.04000	7.31240	-0.27240	
3	48.5	46.45000	46.79315	-0.34315	19.35000	18.67634	0.67366	
4	96.0	12.22000	17.97811	-5.75811	22.88000	25.57680	-2.69680	
5	120.0	10.50000	12.22000	-1.72000	26.30000	22.88000	3.42000	
6	123.0	7.60000	10.50000	-2.90000	23.90000	26.30000	-2.40000	
7	144.0	4.27000	7.60000	-3.33000	22.50000	23.90000	-1.40000	
8	147.0	4.65000	4.27000	0.38000	25.00000	22.50000	2.50000	
9	165.0	4.09000	4.65000	-0.56000	22.80000	25.00000	-2.20000	

Fig. 3: Output data of the program

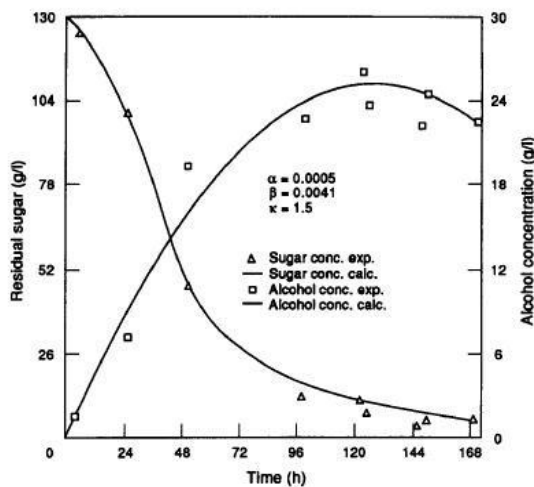


Fig. 4: Experimental and calculated results, from the model for fermentation of sugar beet molasses

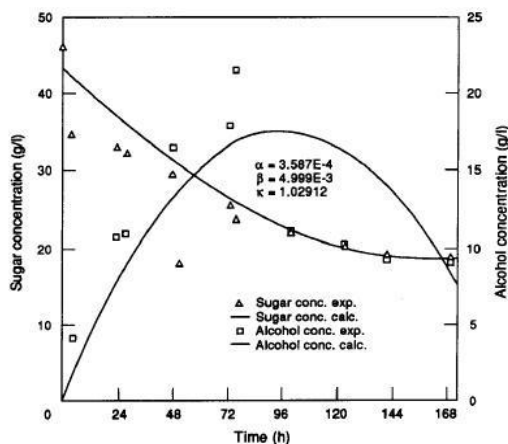


Fig. 5: Experimental and calculated results for fermentation of Akalona hydrolyzate

Conclusions:

The values of the kinetic parameters of the Bovee model [8] were determined from the experimental

results [9] of alcoholic fermentation of beet sugar molasses and Akalona. The computer simulation of the model showed a value of 1.2 as standard deviation for beet sugar molasses and 3.69 for Akalona. Thus, this model with its optimized values of α , β and k can be used as a tool for process control alcoholic fermentation of beet sugar molasses and Akalona.

Acknowledgements: The authors wish to express their deepest thanks and acknowledgement to Professor H. M. Asfour, for his participation in the work and his help.

References:

- [1] Demirbas, A. (2009), "Political, economic and environmental impacts of biofuels: A review". *Applied Energy* **86**: S108–S117. doi: [10.1016/j.apenergy.2009.04.036](https://doi.org/10.1016/j.apenergy.2009.04.036). edit
- [2] "Towards Sustainable Production and Use of Resources: Assessing Biofuels". *United Nations Environment Programme*. 2009-10-16. Retrieved 2009-10-24.
- [3] "Model predictive control of fermentation temperature in biofuel production", Patent 7831318 Issued on **November 9, 2010**. <http://www.patentstorm.us/patents/7831318/description.html>. **Inventors:** James F. Bartee, Maina A. Macharia, Patrick D. Noll, Michael E. Tay.
- [4] P. F. Stanbury and A. Whitaker, *Principles of Fermentation Technology*, pp. 11-14, Pergamon Press, Oxford (1984).
- [5] B. S. Fang; H. Y. Fang; C. S. Wu and C. T. Pan, *Biotechnol. Bioengng Syrup.*, 13, 464 (1983).
- [6] J. H. Luong, *Biotechnol. Bioengng*, 27, 280 (1985).
- [7] A. Gianetto; V. Speochia and G. Genon, *Chem. Engng Commun.*, 23, 215 (1983).

- [8] J. P. Bovee; P. Strehaiano; G. Goma and Y. Sevely, *Bioetchnol. Bioengng*, 26, 328 (1984).
- [9] N. A. Mostafa, "Studies on the kinetics of biochemical reactions", Ph. D. Thesis, Faculty of Engineering, Minia Univ., Minia, Egypt (1990).
- [10] H. E. Grethlein, *J. Appl. Chem. Biotechnol*, 28, 296 (1978).
- [11] T. K. Ghose and R. D. Tyagi, *Biotechnol. Bioengng*, 21, 1401 (1979).

Analytical and Numerical Simulation of Epidemic Models using Maple and Sage

Verónica Orjuela Contreras

Engineering Physics, Universidad EAFIT, Medellín, Antioquia, Colombia
Microengineering Group, Logic and Computation Group, vorjuela@eafit.edu.co

Abstract - The simulation of some epidemic models was made using computational software such as Maple and Sage, and the results were analytical or numerical solutions. The performance of each program was compared and these results could be applied to model the current spread of disease in the world. The obtained graphics show the behavior of the models in some hypothetic cases.

Keywords: Epidemic Models, Computational Software: Maple and Sage, SI model, SIR model, SEIR model, Vector Borne Model.

1 Introduction

Currently, the epidemics reach high level of spread of disease as a result of the globalization, and sometimes the behavior is unknown and the control can be very challenging. Due to this, epidemic models have been developed to predict the outbreak and the proliferation of the infection.

Nevertheless, these models are not linear and the solution cannot be obtained analytically, that is why only numerical solutions are gotten through the use of computational software such as Maple and Sage. In previous studies, the nonlinear models have been considered only from the point of view of computation of the so called basic reproductive number, using computer algebra[1,2,3,4,5].

The objective of this paper is to compare the performance of these two software: Maple[6] and Sage[7], modeling the SI model, SIR model, SEIR model and Vector Borne model, and contrast the results and the benefits of each one.

2 Problem

2.1 SI Model

In the SI Model the equations are:

$$\frac{d}{dt} x(t) = -\beta x(t) y(t) \quad (1) \quad \frac{d}{dt} y(t) = \beta x(t) y(t) \quad (2)$$

Where, β is the contact or infection rate of the disease, and $x(t)$ and $y(t)$ are susceptible and infected individuals respectively.

2.2 SIR Model

In the SIR Model the equations are:

$$\frac{d}{dt} x(t) = -\beta x(t) y(t) \quad (3)$$

$$\frac{d}{dt} y(t) = \beta x(t) y(t) - g y(t) \quad (4)$$

$$\frac{d}{dt} z(t) = g y(t) \quad (5)$$

Where, β is the contact or infection rate of the disease, g represents the mean recovery rate; $x(t)$, $y(t)$ and $z(t)$ are susceptible, infected and recovered individuals respectively.

2.3 SEIR Model

In the SEIR Model the equations are:

$$\frac{d}{dt} x(t) = -\beta x(t) z(t) \quad (6)$$

$$\frac{d}{dt} y(t) = \beta x(t) z(t) - \sigma y(t) \quad (7)$$

$$\frac{d}{dt} z(t) = \sigma y(t) - g z(t) \quad (8)$$

$$\frac{d}{dt} w(t) = g z(t) \quad (9)$$

Where, β is the contact or infection rate of the disease, σ is the transition rate of the exposed individuals to the infected one, g represents the mean recovery rate; $x(t)$, $y(t)$, $z(t)$ and $w(t)$ are susceptible, exposed, infected and recovered individuals respectively.

2.4 Vector Borne Model

In Vector Borne Model the equations are:

$$\frac{d}{dt} x_h(t) = -\beta_{m,h} x_h(t) y_m(t) \quad (10)$$

$$\frac{d}{dt} y_h(t) = \beta_{m,h} x_h(t) y_m(t) - g_h y_h(t) \quad (11)$$

$$\frac{d}{dt} z_h(t) = g_h y_h(t) \quad (12)$$

$$\frac{d}{dt} x_m(t) = -\beta_{h,m} x_m(t) y_h(t) \quad (13)$$

$$\frac{d}{dt} y_m(t) = \beta_{h,m} x_m(t) y_h(t) - g_m y_m(t) \quad (14)$$

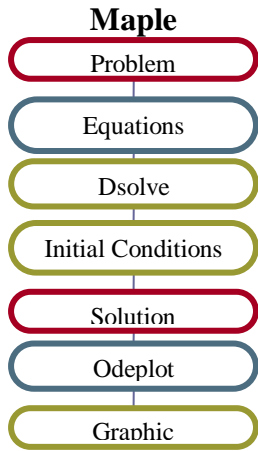
$$\frac{d}{dt} z_m(t) = g_m y_m(t) \quad (15)$$

Where, $\beta_{m,h}$ is the contact or infection rate of the disease due to mosquitoes over humans, $\beta_{h,m}$ is the contact or infection rate of the disease in humans when mosquitoes spread the infection, g_h represents the mean recovery rate in humans, and g_m the mean recovery rate in mosquitoes; $x_h(t)$, $y_h(t)$ and $z_h(t)$ are susceptible, infected and recovered individuals respectively, and $x_m(t)$, $y_m(t)$ and $z_m(t)$ are susceptible, infected and removed mosquitoes respectively.

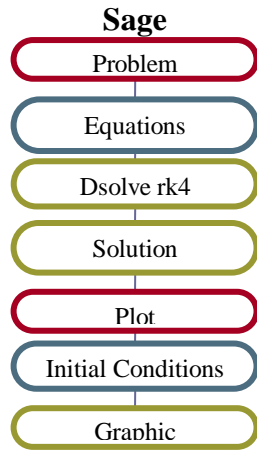
3 Method

In this section, the algorithms using Maple are presented in the first column, and in the second one, using Sage.

3.1 Algorithms using Maple



3.2 Algorithms using Sage



4 Results

With the previous algorithms it was possible to obtain the following results using Maple and Sage:

4.1 Results using Maple

4.1.1 SI Model

For the case of a closed population of constant size $N=n+a$, where n is the initial number of susceptible and a is the initial number of infected individuals, we have the following analytic solution

$$x(t) = \frac{e^{-t\beta(a+n)} n(a+n)}{a + e^{-t\beta(a+n)} n} \tag{16}$$

$$y(t) = \frac{(a+n)a}{a + e^{-t\beta(a+n)} n} \tag{17}$$

$$y(t) = - \frac{\sqrt{-2\mu\beta} a e^{\frac{1}{2}\beta t(2N+\mu t)}}{\beta\sqrt{\pi} e^{-\frac{1}{2}\frac{\beta N^2}{\mu}} \operatorname{erf}\left(\frac{\beta(N+\mu t)}{\sqrt{-2\mu\beta}}\right) a - \sqrt{-2\mu\beta} - \beta\sqrt{\pi} e^{-\frac{1}{2}\frac{\beta N^2}{\mu}} \operatorname{erf}\left(\frac{\beta N}{\sqrt{-2\mu\beta}}\right) a} \tag{20}$$

For the case of a closed population with variable size $N(t)=N+\mu \cdot t$, we have the analytic solution, n equation 18.

$$y(t) = - \frac{\sqrt{-2\mu\beta} a e^{\frac{1}{2}\beta t(2N+\mu t)}}{\beta\sqrt{\pi} e^{-\frac{1}{2}\frac{\beta N^2}{\mu}} \operatorname{erf}\left(\frac{\beta(N+\mu t)}{\sqrt{-2\mu\beta}}\right) a - \sqrt{-2\mu\beta} - \beta\sqrt{\pi} e^{-\frac{1}{2}\frac{\beta N^2}{\mu}} \operatorname{erf}\left(\frac{\beta N}{\sqrt{-2\mu\beta}}\right) a} \tag{18}$$

where a is the initial number of infected individuals, and erf represents the errors function as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{19}$$

A numerical illustration, for the case with closed population with constant size is made possible using Maple with the following commands

```

sys1:=diff(x(t),t)=-beta*(x(t))*(y(t)),diff(y(t),t)=beta*(x(t)*y(t));
fcns1:={x(t),y(t)};
p1:=dsolve({sys1,x(0)=45400,y(0)=2100},fcns1,type=numeric,method=classical);
odeplot(p1,[t,x(t)],[t,y(t)],0..300);
  
```

And the result is showed in the illustration 1

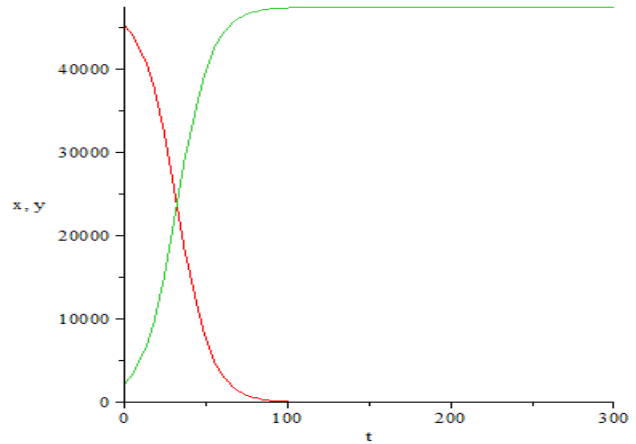


Illustration 1. Result SI Model, closed population with constant size

A numerical illustration for the case of population with variable size is as follows in equation 20:

The corresponding graphic is showed in illustration 2

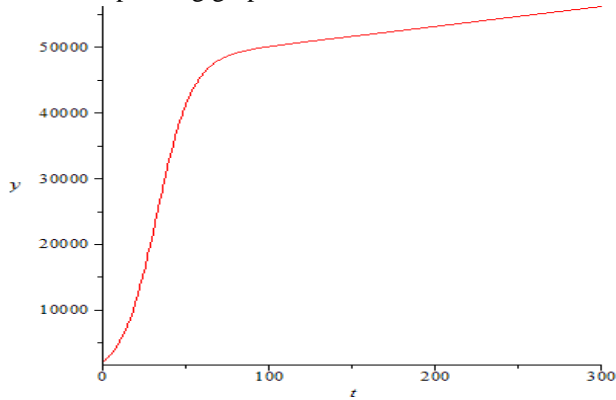


Illustration 2. Result SI Model population with variable size

4.1.2 SIR Model

A numerical illustration, for the case with closed population with constant size is illustrated as it follows

```
sys:=diff(x(t),t)=-beta*(x(t))*(y(t)),
diff(y(t),t)=beta*(x(t))*(y(t))-g*(y(t)),diff(z(t),t)=g*(y(t)):
fcns:={x(t),y(t),z(t)}:
p:=dsolve({sys,x(0)=45400,y(0)=2100,z=2500},fcns,type=
numeric,method=classical):
odeplot(p,[t,x(t)],0..300,numpoints=25);
```

And the results are showed in the illustration 3 and 4

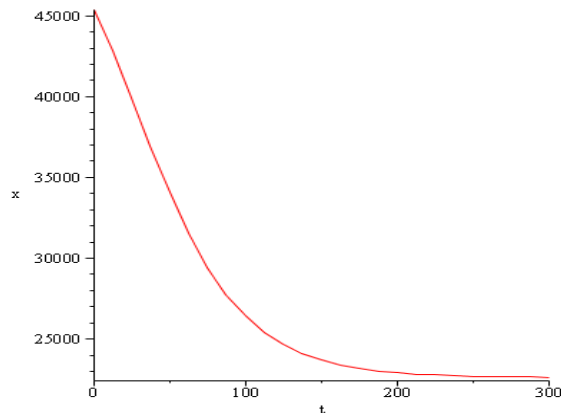


Illustration 3. Result SIR Model, susceptible individuals

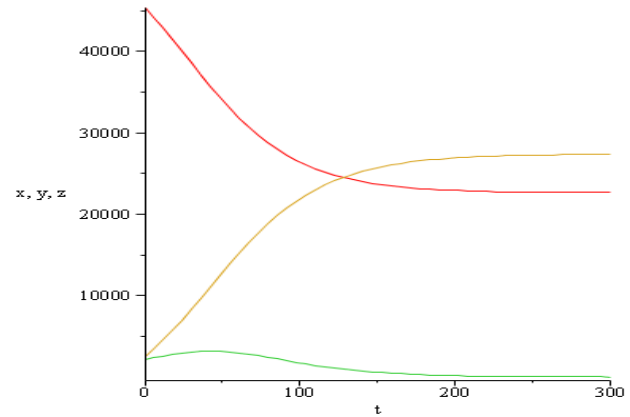


Illustration 4. Result SIR Model

4.1.3 SEIR Model

A numerical illustration, for the case with closed population with constant size is made using the next command

```
sys2:=diff(x(t),t)=-beta*(x(t))*(z(t)),diff(y(t),t)=beta*x(t)-
sigma*(y(t)),diff(z(t),t)=sigma*(y(t))-
g*(z(t)),diff(w(t),t)=g*(z(t)):
fcns2:={x(t),y(t),z(t),w(t)}:
p2:=dsolve({sys,x(0)=45400,y(0)=2100,z=2500,w(0)=1000
0},fcns2,type=numeric,method=classical):
odeplot(p2,[t,x(t)], [t,y(t)], [t,z(t)], [t,w(t)]0..300);
```

And the results are presented in illustration 5

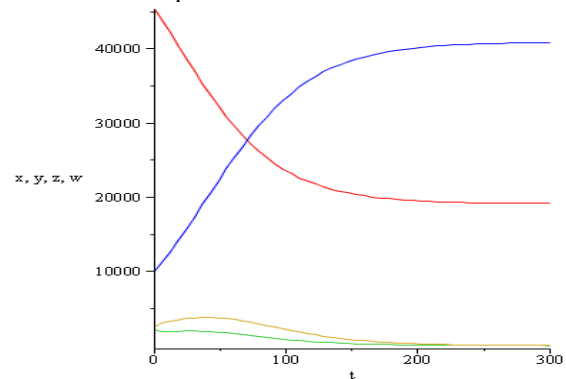


Illustration 5. Result SEIR Model

4.1.4 Vector Borne Model

A numerical result, for the case with closed population with constant size of individuals and mosquitoes is made through the following command

```
sys := diff(x[h](t), t) = -beta[h, m] * x[h](t) * y[m](t), diff(y[h](t), t) = beta[h, m] * x[h](t) * y[m](t) - g[h] * y[h](t), diff(z[h](t), t) = g[h] * y[h](t), diff(x[m](t), t) = -beta[h, m] *
x[m](t) * y[h](t), diff(y[m](t), t) = beta[h, m] * x[m](t) * y[h](t) - g[m] * y[m](t), diff(z[m](t), t) = g[m] * y[m](t) :
fcns := {x[h](t), y[h](t), z[h](t), x[m](t), y[m](t), z[m](t)} :
p := dsolve({sys, x[h](0) = 1000, y[h](0) = 330, z[h](0) = 170, x[m](0) = 500, y[m](0) = 112, z[m](0) = 50}, fcns, type = numeric, method = classical) :
odeplot(p, [t, x[h](t)], [t, y[h](t)], [t, z[h](t)], [t, x[m](t)], [t, y[m](t)], [t, z[m](t)], 0..150, color = [red, blue, green, black, yellow, brown]);
```

And the results can be observed in illustration 6

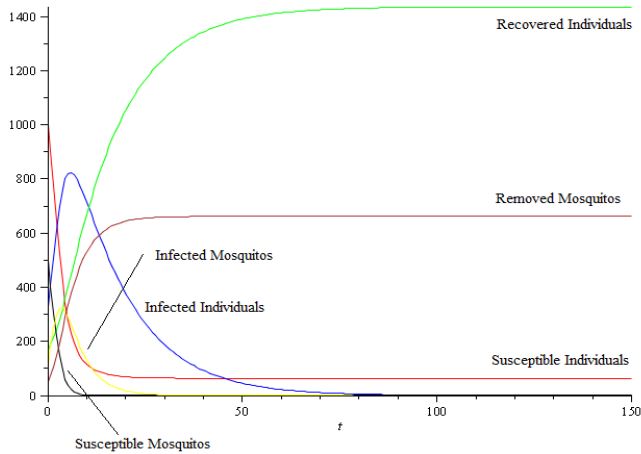


Illustration 6. Result Vector Borne Model

```
def si_ode(S_init, I_init,      # initial values
          endtime,           # time to model
          beta = 2/10^6,     # model parameters
          colors=['red','blue'], thickness=3 # plotting controls
          ):
    var('t S I')
    system = [-beta*S*I, beta*S*I]

    # P will be a set of 3-tuples of the form (t,S,I)
    P = desolve_system_rk4(system, [S,I,], ivar=t,
                          ics=[0, S_init, I_init],
                          step=1, end_points=endtime)

    plotS = plot(line([(time,s) for time,s,i in P],
                     linestyle = '--', rgbcolor=colors[0], thickness=thickness))
    plotI = plot(line([(time,i) for time,s,i in P],
                     linestyle = '-.', rgbcolor=colors[1], thickness=thickness))
    return plotS+plotI
```

And the initial conditions provides the illustration 7

Si_ode(45400, 2100, 300)

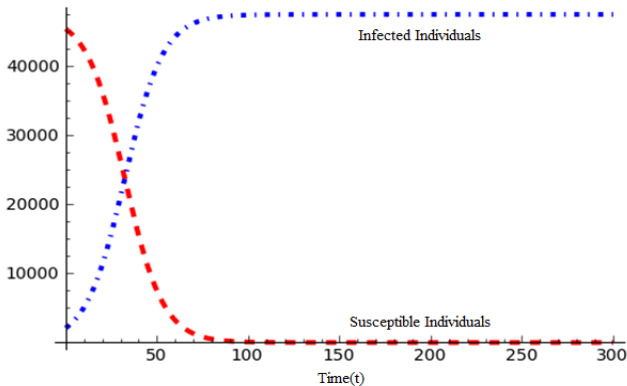


Illustration 7. Result SI Model

4.2 Results using Sage

4.2.1 SI Model

In Sage the algorithm for a numerical illustration, in the case with closed population with constant size, is the following one

4.2.2 SIR Model

The algorithm for a numerical illustration, in the case with closed population with constant size is

```
def sir_ode(S_init, I_init, R_init,      # initial values
            endtime,                  # time to model
            beta = 2/10^6, gamma=1/14, # model parameters
            colors=['black','red', 'blue'], thickness=3 # plotting controls
            ):
    var('t S I R')
    system = [-beta*S*I, beta*S*I- gamma * I, gamma *I]

    # P will be a set of 4-tuples of the form (t,S,I,R)
    P = desolve_system_rk4(system,[S,I,R], ivar=t,
                           ics=[0, S_init, I_init, R_init],
                           step=1, end_points=endtime)

    plots = plot(line([(time,s) for time,s,i,r in P],
                      linestyle = '-', rgbcolor=colors[0]), thickness=thickness)
    plotI = plot(line([(time,i) for time,s,i,r in P],
                      linestyle = '--', rgbcolor=colors[1]), thickness=thickness)
    plotR = plot(line([(time,r) for time,s,i,r in P],
                      linestyle = '-.', rgbcolor=colors[2]), thickness=thickness)
    return plots+plotI+plotR
```

The initial conditions are changed and the results are showed in illustration 8 and 9.
 sir_ode(45400, 2100, 2500, 300)

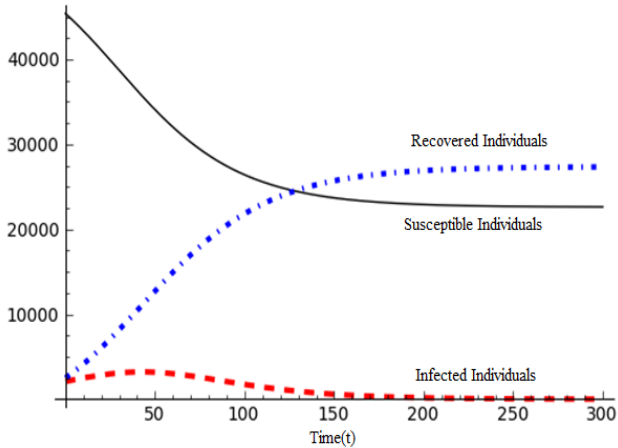


Illustration 8. Result SIR Model

```
sir_ode(45400,2100,2500,300,beta=1/50000)
```

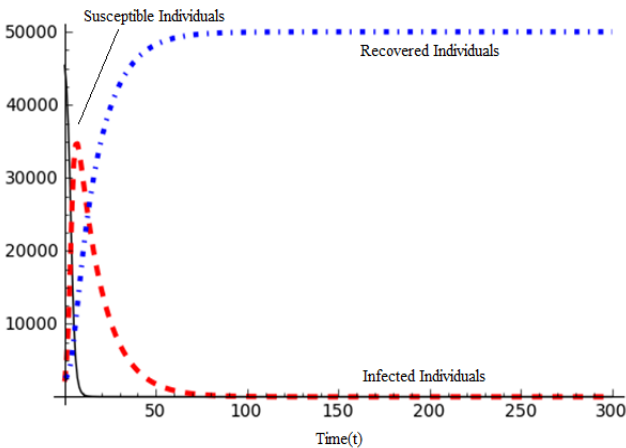


Illustration 9. Result SIR Model with different beta

```
plot1 = sir_ode(45400, 2100, 2500, 150)
plot2 = sir_ode(45400, 2100, 2500, 150,
               gamma=1/10^2, colors=['gray', 'purple', 'orange'])
show(plot1 + plot2)
```

The result is showed in the illustration 10.

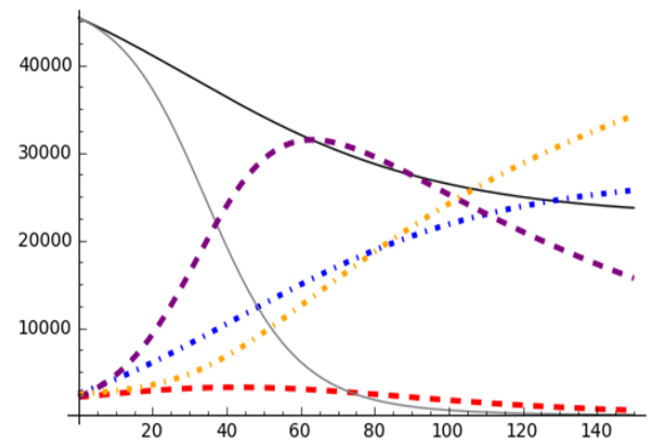


Illustration 10. Result SIR Model for both cases

4.2.3 SEIR Model

For the case of a closed population with constant size, it's possible to obtain a solution through the following algorithm

```

def seir_ode(S_init, E_init, I_init, R_init,          # initial values
            endtime,                                # time to model
            beta = 2/10^6, gamma=1/14, sigma=1/10, # model parameters
            colors=['black','red','blue','green'],  # plotting controls
            thickness=3):
    var('t S E I R')
    system = [-beta*S*I, beta*S*I- sigma*E, sigma*E- gamma*I, gamma*I]

    # P will be a set of 5-tuples of the form (t,S,E,I,R)
    P = desolve_system_rk4(system, [S,E,I,R], ivar=t,
                          ics=[0, S_init, E_init, I_init, R_init],
                          step=1, end_points=endtime)

    plotS = plot(line([(time,s) for time,s,e,i,r in P],
                      linestyle = '-', rgbcolor=colors[0]), thickness=thickness)
    plotE = plot(line([(time,e) for time,s,e,i,r in P],
                      linestyle = '--', rgbcolor=colors[1]), thickness=thickness)
    plotI = plot(line([(time,i) for time,s,e,i,r in P],
                      linestyle = '--', rgbcolor=colors[2]), thickness=thickness))
    plotR = plot(line([(time,r) for time,s,e,i,r in P],
                      linestyle = '-.', rgbcolor=colors[3]), thickness=thickness))
    return plotS+plotE+plotI+plotR

```

The initial conditions give the results showed in the illustration 11

seir_ode(45400, 9000, 3100, 2500, 300)

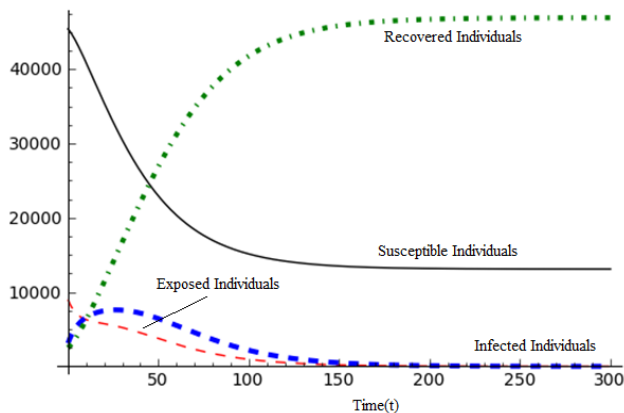


Illustration 11. Result Vector Borne Model

```

def vectbor_ode(Sh_init, Ih_init, Rh_init, Sm_init, Im_init, Rm_init, # initial values
               endtime,                                             # time to model
               betamh=2/10^4, gammah=1/14, betahm=1/10^4, gammam=1/7, # model parameters
               colors=['black','red','blue','green','yellow','gray'], # plotting controls
               thickness=3):
    var('t Sh Ih Rh Sm Im Rm')
    system = [-betamh*Sh*Im, betamh*Sh*Im- gammah*Ih, gammah*Ih, -betahm*Sm*Ih, betahm*Sm*Ih- gammam*Im,
             gammam*Im]

    # P will be a set of 7-tuples of the form (t,Sh,Ih,Rh,Sm,Im,Rm)
    P = desolve_system_rk4(system, [Sh, Ih, Rh, Sm, Im, Rm], ivar=t,
                          ics=[0, Sh_init, Ih_init, Rh_init, Sm_init, Im_init, Rm_init],
                          step=1, end_points=endtime)

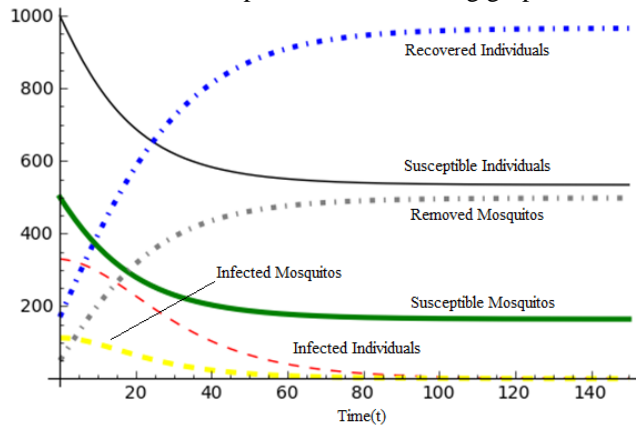
    plotSh = plot(line([(time,sh) for time,sh,ih,rh,sm,im,rm in P],
                      linestyle = '-', rgbcolor=colors[0]), thickness=thickness)
    plotIh = plot(line([(time,ih) for time,sh,ih,rh,sm,im,rm in P],
                      linestyle = '--', rgbcolor=colors[1]), thickness=thickness)
    plotRh = plot(line([(time,rh) for time,sh,ih,rh,sm,im,rm in P],
                      linestyle = '-.', rgbcolor=colors[2]), thickness=thickness))
    plotSm = plot(line([(time,sm) for time,sh,ih,rh,sm,im,rm in P],
                      linestyle = '-', rgbcolor=colors[3]), thickness=thickness))
    plotIm = plot(line([(time,im) for time,sh,ih,rh,sm,im,rm in P],
                      linestyle = '--', rgbcolor=colors[4]), thickness=thickness))
    plotRm = plot(line([(time,rm) for time,sh,ih,rh,sm,im,rm in P],
                      linestyle = '-.', rgbcolor=colors[5]), thickness=thickness))
    return plotSh+plotIh+plotRh+plotSm+plotIm+plotRm

```

4.2.4 VECTOR BORNE MODEL

A numerical result, for the case with closed population with constant size of individuals and mosquitoes is made through the following algorithm

The initial conditions produce the following graph



5 Conclusions

In this paper, several epidemic models were considered and the solutions were obtained through the application of two different software: Maple and Sage. The mathematical calculus and the speed, besides the computational language, were better in Maple. However the graphics of Sage gave a better view of the behavior of the models, even though Sage is very strict about its syntax. On the other side, Maple has a lot of commands to improve the graphics, therefore with a basic understanding, its performance is way better than Sage.

In addition, Maple is software which needs a license for its use, and Sage is free software that is available to everyone who wants it, only with an internet connection. Even so, Sage still has a lot of problems with the server, and not always is working good.

The epidemic models studied in this work could be useful to model some of the epidemics worldwide, due to easier incensement of the illness at the present time, and we hope that this model can predict the conduct and prevent the proliferation of the epidemics.

References

[1] Clarita Saldarriaga Vargas, Mathematical Model for Dengue Epidemics with Differential Susceptibility and Asymptomatic Patients Using Computer Algebra, Lecture Notes in Computer Science, Vol. 5743, pags 284-298, Springer 2009.

[2] Davinson Castaño Cano, Computer Algebra and Mechanized Reasoning in Mathematical Epidemiology, Proceedings of the World Congress on Engineering and Computer Science 2009 Vol I. WCECS 2009, October 20-22, 2009, San Francisco, USA

[3]Jeyver André Morales Taborda, Epidemic Thresholds via Computer Algebra, Proceedings of the 2008

International Conference on Modeling, Simulation & Visualization Methods, MSV 2008, Las Vegas, Nevada, USA, July 14-17, 2008, CSREA Press, 2008.

[4] Doracelly Hincapié palacio et al, The epidemic threshold theorem with social and contact heterogeneity, Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2008, Belur V. Dasarathy, Editors, 69730A.

[5]Juan Ospina et al, Epidemic Thresholds in SIR and SIIR Models Applying an Algorithmic Method, Biosurveillance and Biosecurity, International Workshop, BioSecure 2008 Vol 5354, Springer, NC, USA.

[6]Math Software for Engineers, Educators & Students | Maplesoft, www.maplesoft.com

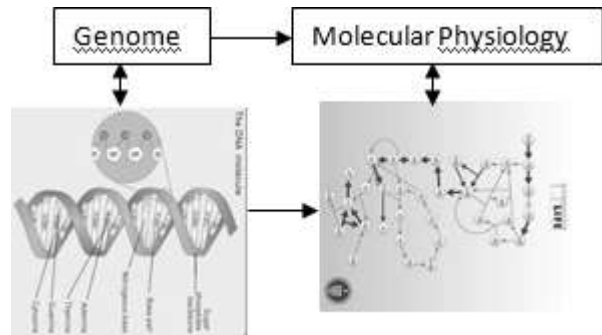
[7]Sage: Open Source Mathematics Software, www.sagemath.org

SIMULATING THE RECONSTRUCTION OF METABOLIC NETWORKS USING MAPLE

DIEGO IGNACIO VELEZ JARAMILLO
 COMPUTATIONAL AND THEORICAL PHYSICS GROUP
 LOGIG AND COMPUTACIONAL GROUP
 ENGENERING PHYSCIS PROGRAM
 EAFIT UNIVERSITY
 MEDELLIN, COLOMBIA

Abstract - In this article is simulated the reconstruction of metabolic networks by means of an algorithm in maple that represents the genome and its form of expression. The metabolic networks are represented by means of graphs, which are constructed from their matrices of adjacency (first squared that represents the connections between elements). Using the adjacency matrix also constructs a graph of density by means in which the algorithm for the reconstruction of the gene is applied.

Keywords: Genome, graphs, metabolic networks, molecular physiology, maple.



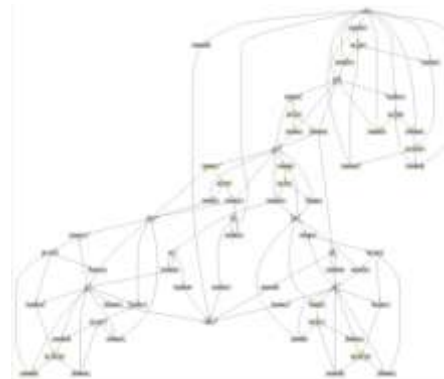
1 Introduction

A gene is an organized linear sequence of nucleotides that contains the necessary information for the synthesis of a macro-molecule with specific cellular function, normally proteins. The proteins occupy a place of maximum importance between constituent molecules of the alive beings, practically all the biological processes depend on the presence or the activity of this type of molecule. Due to this the importance of studying the form in which the genes manipulate this information and the form that the macromolecules express themselves. Also the expression form can be manipulated to take macro-molecules to a wished protein.

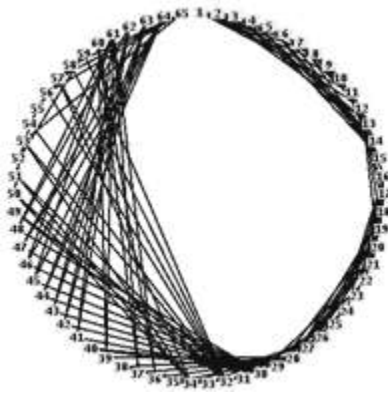
2 Problem

Show how from the genomes is reconstructed molecular physiology [1,2].

2.1 Molecular Physiology treat



```
g := Graph({{1, 2}, {1, 3}, {1, 4}, {1, 5}, {1, 6}, {1, 7}, {1, 8}, {1, 9},
{1, 10}, {1, 11}, {2, 12}, {2, 13}, {3, 13}, {3, 14}, {4, 14}, {4,
15}, {5, 14}, {5, 15}, {6, 12}, {6, 15}, {7, 12}, {7, 15}, {8, 12},
{8, 13}, {9, 14}, {9, 13}, {10, 16}, {11, 17}, {12, 18}, {12, 19},
{12, 20}, {16, 21}, {16, 22}, {16, 23}, {16, 24}, {16, 25}, {17,
23}, {17, 11}, {17, 26}, {18, 27}, {18, 28}, {19, 27}, {19, 28},
{19, 29}, {20, 28}, {20, 27}, {21, 30}, {22, 29}, {24, 31}, {25,
32}, {26, 28}, {26, 33}, {26, 34}, {28, 35}, {28, 36}, {28, 37},
{28, 38}, {28, 39}, {29, 40}, {40, 29}, {30, 40}, {30, 41}, {30,
42}, {30, 43}, {30, 44}, {30, 45}, {30, 46}, {30, 47}, {30, 48},
{31, 34}, {31, 49}, {49, 31}, {32, 49}, {32, 50}, {32, 51}, {32,
59}, {32, 53}, {32, 54}, {32, 57}, {32, 58}, {32, 56}, {33, 35},
{33, 38}, {34, 35}, {34, 38}, {34, 53}, {34, 57}, {34, 50}, {34,
51}, {36, 60}, {36, 61}, {37, 60}, {37, 61}, {39, 60}, {39, 61},
{41, 61}, {41, 62}, {42, 62}, {42, 63}, {43, 62}, {43, 63}, {44,
63}, {44, 64}, {45, 61}, {45, 64}, {46, 61}, {46, 64}, {47, 63},
{47, 64}, {48, 61}, {48, 62}, {50, 52}, {51, 52}, {52, 54}, {52,
59}, {53, 65}, {54, 55}, {55, 56}, {55, 59}, {55, 58}, {56, 65},
{58, 65}, {57, 65}});
```



3 Method

1. Represent metabolic networks as graphs, defined by its adjacency matrices
2. Represent the adjacency matrix as a graph of density
3. Represent the genome using the following algorithm for image reconstruction.

The following algorithm represents the genome and how it manipulates information to reconstruct the molecular physiology.

```
List_Density_Plot := proc(data)
  global i_row, j_column;
  local i, f, local_Am;
  f := x → -x + 1;
  local_Am := array(1..i_row, 1..j_column, [seq(map(f,
    [seq(data[i_row-i + 1, j], j = 1..j_column)]), i = 1..i_row)]);
  listdensityplot(transpose(local_Am));
end;
```

```
ns_sum := proc()
  global i_row, j_column, Am;
  local i, j, ns;
  for j from 1 to j_column do
    ns[j] := 0;
    for i from 1 to i_row do
      ns[j] := ns[j] + Am[i, j];
    od;
  od;
  RETURN(ns);
end;
```

```
we_sum := proc()
  global i_row, j_column, Am;
  local i, j, we;
  for i from 1 to i_row do
    we[i] := 0;
    for j from 1 to j_column do
      we[i] := we[i] + Am[i, j];
    od;
  od;
  RETURN(we);
end;
```

```
NwSe_diag_sum := proc()
  global i_row, j_column, Am;
  local i, j, we_diag;
  for i from 1 to i_row do
    we_diag[i, 1] := 0; we_diag[i, 2] := 0;
    for j from 1 to 100 while ((j ≤ j_column) and (i-j + 1 ≥ 1)) do
      we_diag[i, 1] := we_diag[i, 1] + Am[i-j + 1, j];
      we_diag[i, 2] := we_diag[i, 2] + 1;
    od;
  od;
  for j from 2 to j_column do
    we_diag[i_row + j - 1, 1] := 0; we_diag[i_row + j - 1, 2] := 0;
    for i from 1 to 100 while ((j + i - 1 ≤ j_column) and (i_row - i + 1 ≥ 1)) do
      we_diag[i_row + j - 1, 1] := we_diag[i_row + j - 1, 1]
      + Am[i_row - i + 1, j + i - 1];
      we_diag[i_row + j - 1, 2] := we_diag[i_row + j - 1, 2] + 1;
    od;
  od;
  RETURN(we_diag);
end;
```

```
NeSw_diag_sum := proc()
  global i_row, j_column, Am;
  local i, j, ew_diag;
  for i from 1 to i_row do
    ew_diag[i, 1] := 0; ew_diag[i, 2] := 0;
    for j from 1 to 100 while ((j_column - j + 1 ≥ 1) and (i - j + 1 ≥ 1)) do
      ew_diag[i, 1] := ew_diag[i, 1] + Am[i - j + 1, j_column - j + 1];
      ew_diag[i, 2] := ew_diag[i, 2] + 1;
    od;
  od;
  for j from 2 to j_column do
    ew_diag[i_row + j - 1, 1] := 0; ew_diag[i_row + j - 1, 2] := 0;
    for i from 1 to 100 while ((j_column - i - j + 2 ≥ 1) and (i_row - i + 1 ≥ 1)) do
      ew_diag[i_row + j - 1, 1] := ew_diag[i_row + j - 1, 1]
      + Am[i_row - i + 1, j_column - i - j + 2];
      ew_diag[i_row + j - 1, 2] := ew_diag[i_row + j - 1, 2] + 1;
    od;
  od;
  RETURN(ew_diag);
end;
```

```
vertical_distrib := proc(a)
  global i_row, j_column;
  local i, j, temp_array, temp_matrix;
  for j from 1 to j_column do
    for i from 1 to i_row do
      temp_array[i, j] := a[j]/i_row;
    od;
  od;
  temp_matrix := array(1..i_row, 1..j_column,
    [seq([seq(temp_array[i, j], j = 1..j_column)], i = 1..i_row)]);
  RETURN(temp_matrix);
end;
```

```
horizontal_distrib := proc(a)
  global i_row, j_column;
  local i, j, temp_array, temp_matrix;
  for i from 1 to i_row do
    for j from 1 to j_column do
      temp_array[i, j] := a[i]/j_column;
    od;
  od;
  temp_matrix := array(1..i_row, 1..j_column,
    [seq([seq(temp_array[i, j], j = 1..j_column)], i = 1..i_row)]);
  RETURN(temp_matrix);
end;
```



```

NwSe_diag_distrib := proc(a)
  global i_row, j_column;
  local i, j, temp_array, temp_matrix;
  for i from 1 to i_row do
    for j from 1 to 100 while ((j ≤ j_column) and (i-j + 1 ≥ 1)) do
      temp_array[i-j + 1, j] := a[i, 1]/a[i, 2];
    od;
  od;
  for j from 2 to j_column do
    for i from 1 to 100 while ((j + i-1 ≤ j_column) and (i_row-i + 1 ≥ 1)) do
      temp_array[i_row-i + 1, j + i-1] := a[i_row + j-1, 1]
      /a[i_row + j-1, 2];
    od;
  od;
  temp_matrix := array(1..i_row, 1..j_column,
    [seq([seq(temp_array[i, j], j = 1..j_column)], i = 1..i_row)]);
  RETURN(temp_matrix);
end;

```

```

NeSw_diag_distrib := proc(a)
  global i_row, j_column;
  local i, j, temp_array, temp_matrix;
  for i from 1 to i_row do
    for j from 1 to 100 while ((j_column-j + 1 ≥ 1) and (i-j + 1 ≥ 1)) do
      temp_array[i-j + 1, j_column-j + 1] := a[i, 1]/a[i, 2];
    od;
  od;
  for j from 2 to j_column do
    for i from 1 to 100 while ((j_column-i-j + 2 ≥ 1) and (i_row-i + 1 ≥ 1)) do
      temp_array[i_row-i + 1, j_column-i-j + 2] := a[i_row + j - 1, 1]/a[i_row + j - 1, 2];
    od;
  od;
  temp_matrix := array(1..i_row, 1..j_column,
    [seq([seq(temp_array[i, j], j = 1..j_column)], i = 1..i_row)]);
  RETURN(temp_matrix);
end;

```

4 Results

```
>G := SoccerBallGraph( );
```

Graph 1: an undirected unweighted graph with 60 vertices and 90 edge
(s)

```

i_row := 60 : j_column := 60 :
catscan := AdjacencyMatrix(G)

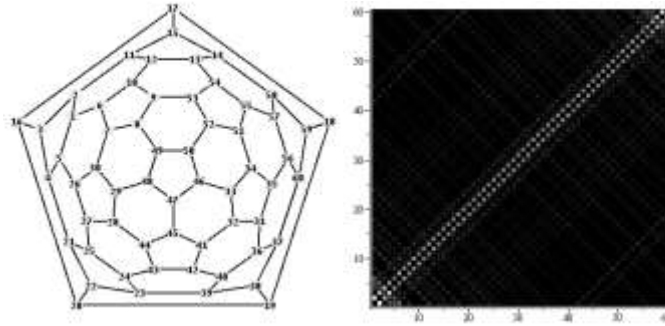
```

60 x 60 Matrix
Data Type: anything
Storage: triangular _{upper}
Order: C_order

```

we := we_sum( ) : ns := ns_sum( ) : NwSe_diag
:= NwSe_diag_sum( ) : NeSw_diag := NeSw_diag_sum( ) :
seq(we[i], i = 1..i_row) : seq(ns[j], j = 1..j_column) :
seq(NwSe_diag[i, 1], i = 1..i_row + j_column-1) :
seq(NeSw_diag[i, 1], i = 1..i_row + j_column-1) :
seq(NwSe_diag[i, 2], i = 1..i_row + j_column-1) : seq(NeSw_diag[i,
2], i = 1..i_row + j_column-1) :
img[1] := vertical_distrib(ns) : img[2] := horizontal_distrib(we) :
img[3] := NwSe_diag_distrib(NwSe_diag) : img[4]
:= NeSw_diag_distrib(NeSw_diag) :
# print([seq(ns[j], j=1..j_column)]): print(img[1]):
total_image := (evalm(sum(img[i], i = 1..4))) :
threshold := sort([seq(seq(total_image[i, j], j = 1..j_column), i = 1
..i_row)], '>')[sum(we[j], j = 1..i_row)] :
> listdensityplot(total_image);

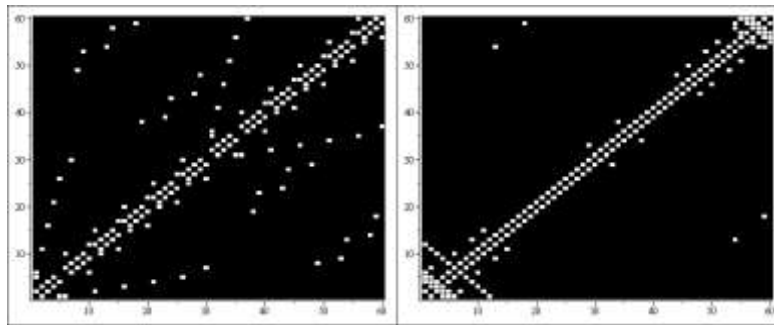
```



```
> g := x → if (x ≥ threshold) then 1 else 0 fi;
```

```
p1 := listdensityplot(catscan) : p2 := listdensityplot(map(g,
total_image)) :
```

```
> display(array([p1, p2]));
```



```
> H := FosterGraph( );
```

Graph 4: an undirected unweighted graph with 90 vertices and 135 edge(s)

```
> i_row := 90 : j_column := 90 :
```

```
Am := AdjacencyMatrix(H) :
```

```
we := we_sum( ) : ns := ns_sum( ) : NwSe_diag
```

```
:= NwSe_diag_sum( ) : NeSw_diag := NeSw_diag_sum( ) :
```

```
seq(we[i], i = 1 .. i_row) : seq(ns[j], j = 1 .. j_column) :
```

```
seq(NwSe_diag[i, 1], i = 1 .. i_row + j_column - 1) :
```

```
seq(NeSw_diag[i, 1], i = 1 .. i_row + j_column - 1) :
```

```
seq(NwSe_diag[i, 2], i = 1 .. i_row + j_column - 1) : seq(NeSw_diag[i,
2], i = 1 .. i_row + j_column - 1) :
```

```
img[1] := vertical_distrib(ns) : img[2] := horizontal_distrib(we) :
```

```
img[3] := NwSe_diag_distrib(NwSe_diag) : img[4]
```

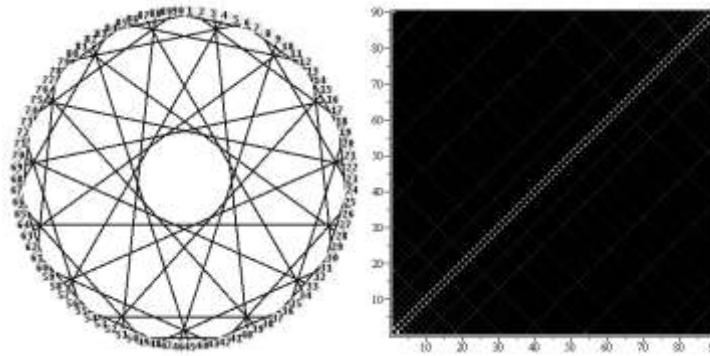
```
:= NeSw_diag_distrib(NeSw_diag) :
```

```
# print([seq(ns[j], j = 1 .. j_column)]: print(img[1]):
```

```
total_image := (evalm(sum(img[i], i = 1 .. 4))) :
```

```
threshold := sort([seq(seq(total_image[i, j], j = 1 .. j_column), i = 1
.. i_row)], '>')[sum(we[j], j = 1 .. i_row)] :
```

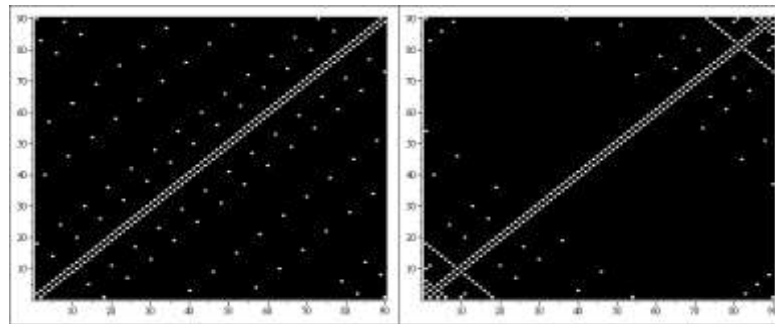
```
> listdensityplot(total_image); DrawGraph(H);
```



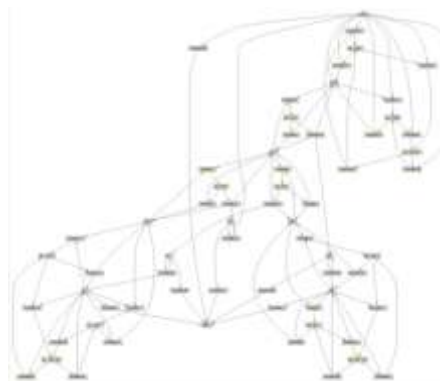
```
> g := x → if (x ≥ threshold) then 1 else 0 fi:
```

```
p1 := listdensityplot(catscan) : p2 := listdensityplot(map(g,
total_image)) :
```

```
> display(array([p1,p2]));
```



Metabolic network Taken for analysis



```
g := Graph({{1,2}, {1,3}, {1,4}, {1,5}, {1,6}, {1,7}, {1,8}, {1,9},
{1,10}, {1,11}, {2,12}, {2,13}, {3,13}, {3,14}, {4,14}, {4,
15}, {5,14}, {5,15}, {6,12}, {6,15}, {7,12}, {7,15}, {8,12},
{8,13}, {9,14}, {9,13}, {10,16}, {11,17}, {12,18}, {12,19},
{12,20}, {16,21}, {16,22}, {16,23}, {16,24}, {16,25}, {17,
23}, {17,11}, {17,26}, {18,27}, {18,28}, {19,27}, {19,28},
{19,29}, {20,28}, {20,27}, {21,30}, {22,29}, {24,31}, {25,
32}, {26,28}, {26,33}, {26,34}, {28,35}, {28,36}, {28,37},
{28,38}, {28,39}, {29,40}, {40,29}, {30,40}, {30,41}, {30,
42}, {30,43}, {30,44}, {30,45}, {30,46}, {30,47}, {30,48},
{31,34}, {31,49}, {49,31}, {32,49}, {32,50}, {32,51}, {32,
59}, {32,53}, {32,54}, {32,57}, {32,58}, {32,56}, {33,35},
{33,38}, {34,35}, {34,38}, {34,53}, {34,57}, {34,50}, {34,
51}, {36,60}, {36,61}, {37,60}, {37,61}, {39,60}, {39,61},
{41,61}, {41,62}, {42,62}, {42,63}, {43,62}, {43,63}, {44,
63}, {44,64}, {45,61}, {45,64}, {46,61}, {46,64}, {47,63},
{47,64}, {48,61}, {48,62}, {50,52}, {51,52}, {52,54}, {52,
59}, {53,65}, {54,55}, {55,56}, {55,59}, {55,58}, {56,65},
{58,65}, {57,65}});
```

```
> i_row := 65 : j_column := 65 :
```

```
Am := AdjacencyMatrix(g) :
```

```
we := we_sum( ) : ns := ns_sum( ) : NwSe_diag
:= NwSe_diag_sum( ) : NeSw_diag := NeSw_diag_sum( ) :
```

```
seq(we[i], i = 1 ..i_row) : seq(ns[j], j = 1 ..j_column) :
seq(NwSe_diag[i, 1], i = 1 ..i_row + j_column-1) :
seq(NeSw_diag[i, 1], i = 1 ..i_row + j_column-1) :
```

```
seq(NwSe_diag[i, 2], i = 1 ..i_row + j_column-1) : seq(NeSw_diag[i,
2], i = 1 ..i_row + j_column-1) :
```

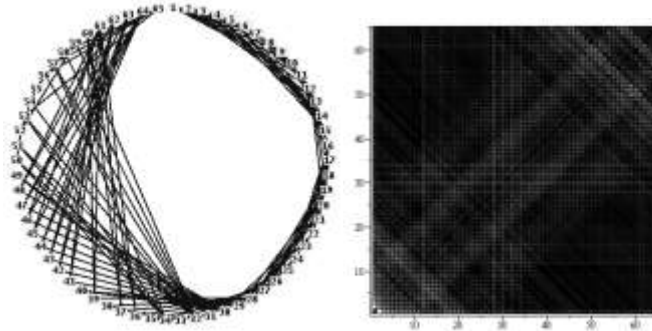
```

img[1] := vertical_distrib(ns) : img[2] := horizontal_distrib(we) :
img[3] := NwSe_diag_distrib(NwSe_diag) : img[4]
:= NeSw_diag_distrib(NeSw_diag) :

# print([seq(ns[j],j=1..j_column)]): print(img[1]):
total_image := (evalm(sum(img[i], i = 1..4))) :
threshold := sort([seq(seq(total_image[i,j],j = 1..j_column), i = 1
..i_row)], '>')[sum(we[j],j = 1..i_row)] :

> listdensityplot(total_image); DrawGraph(g);

```



```

> g := x → if (x ≥ threshold) then 1 else 0 fi:

```

```

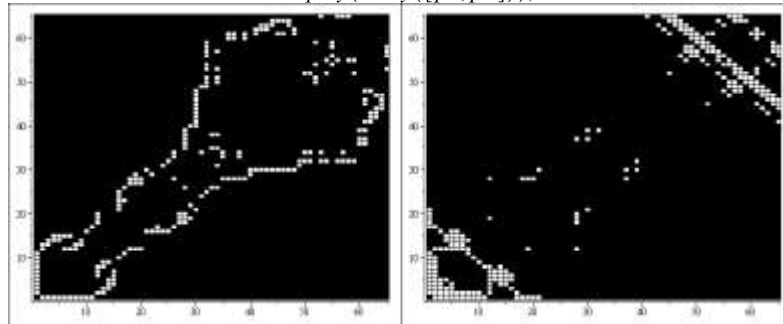
p1 := listdensityplot(Am) : p2 := listdensityplot(map(g,
total_image)) :

```

```

> display(array([p1,p2]));

```



5 Conclusions

In this article I simulate the reconstruction of metabolic networks by means of an algorithm that represents the information codified within a gene. By means of this information it is possible to know the form that a macromolecule is processed and representing it by means of a density graph. In the development process many reconstructions were realized beginning by graphs of low complexity in which was observed that the realized reconstructions were exact, as the complexity of the interactions are increased and these behave in an asymmetric form, it's observed that the reconstructions move away a little from their original structure. These discrepancies are acceptable considering that conserve the central structures. This can be interpreted like the variation between phenotypes from an original genotype.

6 References

1. [HTTP://WWW.MATHWORKS.COM/PRODUCTS/BIOINFO/DEMOS.HTML?FILE=/PRODUCTS/DEMOS/SHIPPING/BIOINFO/GRAPHTHEORYDEMO.HTML](http://www.mathworks.com/products/bioinfo/demos.html?file=/products/demos/shipping/bioinfo/graphtheorydemo.html)
2. [HTTP://EN.WIKIPEDIA.ORG/WIKI/METABOLIC_NETWORK_MODELING](http://en.wikipedia.org/wiki/Metabolic_network_modeling)
3. [HTTP://WWW.MAPLESOFT.COM/APPLICATIONS/VIEW.ASPX?SID=4273&VIEW=HTML](http://www.maplesoft.com/applications/view.aspx?sid=4273&view=html)
4. GRAPH THEORY MAPLE SOFTWARE.

Development of a computing model for resistance screening of *Citrus limon* cultivars infected by the causal agent of “Mal secco” *Phoma tracheiphila*

K. Khanchouch^{1,2}, M.R. Hajlaoui², and H. Kutucu³

¹Department of techniques, ISAJC University of Tunis, Tunis, Tunisia

²Laboratory of plant protection, National research agronomic institute, Tunis, Tunisia

³Department of Mathematics, Izmir institute of technology, Ural-Izmir, Turkey

Abstract - The mathematical survey of the different studied cultivars, using a polynomial model, permits to describe the resistance state of the infected plants. The polynomial interpolation at 5 degrees appears to be the most adequate for this mathematical model. Comparison of R^2 values showed that the polynomial regression at the 5 degree gives the best results. The statistical analysis confirm those obtained by the polynomial model. This polynomial model have the advantage to give a strict evaluation of the state of the resistance of the cultivar tested and not a relative estimation as its in the case of the different mathematical and statistical others classic tools usually used to evaluate the state of the plant resistance. The computing model is able to distinguish between the three resistances levels of *Citrus limon* cultivars tested.

Keywords: Biomath, Modelling, *Citrus limon*, *Phoma tracheiphila* and Bioinformatic

1 Introduction

Plant diseases are responsible of 14.1% of the world crop loses which represent \$220 billion of dollars. These phytopathological damages imply several others problems in different sectors concerning human health, the environment and some social and economic aspects of our life [1]. In order to have an efficient solution to control the causal agents of these diseases it is very important to understand the mechanism of these illnesses very well [2, 3].

In phytopathology, the Mathematical tools used offer models describing the process of the infection [4]. These mathematical models allow to describe the pathological processes and therefore to foresee the most efficient control methods. The mainly mathematical tools used to model the plant diseases are: Disease progress curves, Linked Differential Equation (LDE) and Area Under disease Progress Curve (AUDPC). Statistical analyses are also employed in the studies of epidemiology of plant diseases. Each tool is utilized for an acute appropriate purpose to model some aspects of the disease development.

The specificity of the host-parasite relationships determine the variables and the adequate mathematical model to be used. On the basis of these chosen mathematical tools the most Known model developed in the phytopathological studies are: Monomolecular, Exponential, Gompertz and Logistic models. The logistic model which was proposed firstly by Veshulst in 1838 to represent human population growth was after developed by Van der Plank (1963) [5], to being more appropriate for most polycyclic diseases. This growth model is the most widely used for describing epidemics of plant disease [3,6].

Using the logistic model alone or combined with others tools many plant diseases were described. In the case of the *Citrus* disease “Mal secco”, there is no reports referring to the development of a model allowing to test the resistance degree of the susceptible infected host plants. The causal agent of the “Mal secco”, *Phoma tracheiphila* (Petri) [Kanc et Ghik.], is responsible of many important losses in the *Citrus* crop orchards and it's the most destructive fungal disease of lemon plantation worldwide [2]. As fungicides treatments showed non efficient results to control this pathogen, the research of resisting cultivars remains the most efficient solution to decrease the losses inflicted by the pathogen [7].

2 Material and methods

2.1 Biological Material

Plants belong to *Citrus limon* cultivars and a highly virulent pathogenic isolate of the causal agent of Mal secco were used. The green house inoculation method used is its described by Hajloui et al (2000) [8]. 120 inoculation points in total are assessed per plant. A scale of six classes is used to evaluate the reaction of tested plants. The classes are numbered from 0 to 5.

2.2 The Mathematical Model

2.2.1 Polynomial interpolation

The cumulative percentage frequency of each class is calculated for all the tested plants. The calculation of the cumulative frequency is determined as described below:

$$Y_i = \left[\sum_0^i \text{frequency of } x_i \right] / 120 * 100$$

Y_i = The cumulative frequency at the respective class, X_i .

X_i = class 'i' varying from "0" to "5"

120, it's the number of the inoculation points tested.

The representative curve of the cumulative percentage frequency for the different tested plants it's a polynomial. The degree of the polynomial is fixed referring to the theorem of Lagrange. This is approved using the data of the inoculation test and the calculation of the coefficient of determination " R^2 ".

$$f(x_i) = a x_i^5 + b x_i^4 + c x_i^3 + d x_i^2 + e x_i^1 + f$$

To determine the coefficients of the polynomial function a linear system of six equations is used:

$$\begin{cases} Y_0 = a x_0^5 + b x_0^4 + c x_0^3 + d x_0^2 + e x_0^1 + f \\ Y_1 = a x_1^5 + b x_1^4 + c x_1^3 + d x_1^2 + e x_1^1 + f \\ Y_2 = a x_2^5 + b x_2^4 + c x_2^3 + d x_2^2 + e x_2^1 + f \\ Y_3 = a x_3^5 + b x_3^4 + c x_3^3 + d x_3^2 + e x_3^1 + f \\ Y_4 = a x_4^5 + b x_4^4 + c x_4^3 + d x_4^2 + e x_4^1 + f \\ Y_5 = a x_5^5 + b x_5^4 + c x_5^3 + d x_5^2 + e x_5^1 + f \end{cases}$$

To calculate the coefficients a, b, c, d, e and f, we use Gaussian elimination method [9]. After the coefficients are determined, the polynomial is used to calculate the area under the curve of the infected plant by integrating $f(x)$ from 0 to 5:

$$\int_0^5 f(x) = \left[a \frac{1}{6} x^6 + b \frac{1}{5} x^5 + c \frac{1}{4} x^4 + d \frac{1}{3} x^3 + e \frac{1}{2} x^2 + f x^1 \right]$$

2.2.2 Statistical analysis

For data analysis of the artificial inoculated plants, the biosoftware Statistica version 5.1 was used. Newman and keul-Keuls test at $p=0,05$ of Anova order 1 was performed.

2.2.3 Computing the plant resistance level

Mathematical description of the resistance level of the infected plants is based on the characteristics of their polynomial regression curve. Three types of polynomial curve can be described:

Type A: with an upper concave convection

Type B: with a mixed convection curve

Type C: with a lower concave convection

The parametric analysis of each polynomial curve by calculating its derivative near the convection points, allows distinguishing between the three types.

$$f'(x_i) = 5a x_i^4 + 4b x_i^3 + 3c x_i^2 + 2d x_i + e$$

The derivative calculation for each polynomial regression curve is performed from the point $x_i = 1$ to the point $x_i = 4$. A linear regression curve to fit the calculated values of the derivative is determined using these formulations:

$$y_i' = a x_i + b$$

$$a = \frac{\sum_1^i (x_i - \bar{x})(y_i' - \bar{y}')}{\sum_1^i (x_i - \bar{x})^2}$$

$$b = \bar{y}' - a \bar{x}$$

with :

\bar{x} = means of x_i

\bar{y}' = means of y_i'

The adjustment of the fitted linear derivative regression curve is appreciated using its R^2 value. Higher value of R^2 characterizes both the type A and C while the type B is determined by its R^2 as equal or less than 0,5. The coefficient "a" is positive for type A and negative for type C.

2.3 Algorithm building

In order to determine the coefficients a, b, c, d, e and f, we use Gaussian elimination method:

INPUT : A(n, n), b(n)

OUTPUT: x(n) as the solution

for k=1 to n-1

 for i = k+1 to n

 factor = A(i, k)/A(k, k)

 for j=k+1 to n

 A(i, j) = A(i, j) - factor*A(k, j)

 end

 b(i)=b(i)-factor*b(k)

 end

end

x(n)=b(n)/A(n,n)

```

for i=n-1 to 1
  sum=0
  for j=i+1 to n
    sum=sum+a(i,j)*x(j)
  end
  x(i)=(b(i)-sum)/a(i,i)
end

```

After the coefficients are determined, the polynomial is used to calculate the area under the curve of the infected plant by integrating $f(x)$, from 0 to 5. We use Simpson's method to compute the area under the regression polynomial curve:

INPUT : $p(x)$, a (lower limit of integration), b (upper limit of integration), n (the number of subintervals to divide interval [a,b], n must be divisible by 3)
 OUTPUT: Integral of $p(x)$ from a to b

```

SECTIONS = n/3
h = (b-a) / n
APPROX = 0
for i=1 to SECTIONS:
  x0 = a + 3 * (i-1) * h
  x1 = x0 + h
  x2 = x1 + h
  x3 = x2 + h
  APPROX = APPROX + p(x0) + 3*p(x1) + 3*p(x2) + p(x3)
end
INTEGRAL = 3 * h/8 * APPROX

```

To discriminate between the three types of the regression polynomial curves determined by the mathematical model we elaborated this Algorithm:

INPUT : $p(x) = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$
 OUTPUT: Fitted Line ($y = a'x + b'$), R^2 and Decision

Step 1: for $i=1$ to 4
 $x_i = i$
 end

Step 3: Find the derivative of $p(x)$

$$p'(x) = 5ax^4 + 4bx^3 + 3cx^2 + 2dx + e$$

Step 4: for $i=1$ to 4
 $y_i = p'(i)$
 end

Step 5: Find \bar{x} (the average of x_i)
 for $i=1$ to 4
 $sum = sum + x_i$
 end
 $\bar{x} = sum/4$

Step 6: Find \bar{y} (the average of y_i)

```

for i=1 to 4
  sum=sum+yi
end
 $\bar{y} = sum/4$ 

```

Step 7: Find a' (the coefficient of x for the fitted line)
 for $i=1$ to 4

```

  sum1=sum1+(xi -  $\bar{x}$ )(yi -  $\bar{y}$ )
  sum2=sum2+(xi -  $\bar{x}$ )2
end
a' = sum1/sum2

```

Step 8: Find b' (the constant term of the fitted line)

$$b' = \bar{y} - a' \bar{x}$$

Step 9: Find \hat{y}_i for all $i=1, \dots, 4$

```

for i=1 to 4
   $\hat{y}_i = a' i + b'$ 
end

```

Step 10: Find R^2
 for $i=1$ to 4

```

  sum1=sum1+(yi -  $\hat{y}_i$ )2
  sum2=sum2+(yi -  $\bar{y}$ )2
end
R2 = sum1/sum2

```

Step 11: If ($R^2 \leq 0.5$) Then
 the tested plant is classified as tolerant

Else If ($a > 0$) Then
 the tested plant is classified as sensible

Else If ($a < 0$) Then
 the tested plant is classified as resistance

End If

2.4 Results and discussion

The greenhouse artificial inoculation of the tested lemon cultivars was surveyed on the basis of severity of the disease symptoms following foliar inoculation. Three Citrus cultivars; Eureka, Interdonato and Monachello were used in this test.

All the tested plants developed the pathological symptoms caused by the parasite after the incubation period. Cultivars Eureka and Monachello expressed, relatively, the highest and the reduced degree of the disease index. However,

Interdonato, showed an intermediate index of the disease severity. Statistical analysis indicate a significant differences between the three inoculated cultivars at $p=0,05$. According to this analysis, the tested plants were ranked in three different groups of resistance (Table 1).

Table 1: Infection severity rating of lemon cultivars.

Cultivars	Severity of the disease	Ranked group
Eureka	4,216	I
Interdonato	2,225	II
Monachello	0,8	III

I= sensible, II= Tolerant and III= Resistant

Calculation of the coefficients of the polynomial regression curves by the biomathematical model for the three resistance groups is resumed in the table 2.

Table 2: Parameters of the fitted polynomial curves

Cultivars	Regression curve	area
Eureka	$y = 0,187x^5 - 1,736x^4 + 6,701x^3 - 12,43x^2 + 13,944x + 2,5$	119,17
Interdonato	$y = 0,263x^5 - 2,534x^4 + 7,638x^3 - 5,798x^2 + 0,430x + 47,5$	297,85
Monachello	$y = 0,208x^5 - 3,159x^4 + 18,611x^3 - 54,340x^2 + 82,847x + 41,666$	455,43

The polynomial's coefficients determined by the biomathematical model were in concordance with its obtained by Mathematica uses Hermite interpolation technique to find fitting curves to a given sets of data [10].

Elevated value of R^2 , determined by the model for all the groups of resistance, indicates the strength of fit between the polynomial regression curve and the percentage of the cumulative frequency. The area under the curve (AUC) calculated for the group I was the lowest. The group II and III were characterized but their respective increased value of the AUC.

Drawing the representative fitted curve, it was also found that the type A of the polynomial is correlated with the group I of resistance. While the type B was attributed to the group II of resistance. The type C is associated with the group III of resistant plant (Figure 1).

The algorithm build recognize the different studied groups and attribute to the tested plant its level of resistance. The tests performed on *citrus limon* cultivars using the biomathematical model are in perfect concordance with those obtained by the usually statistical tools.

The mathematical model is able to evaluate the resistance of the infected plant without using comparative methods. The model describes the repartition of the different classes of the

disease to evaluate more precisely the reaction of the host infected by the parasite.

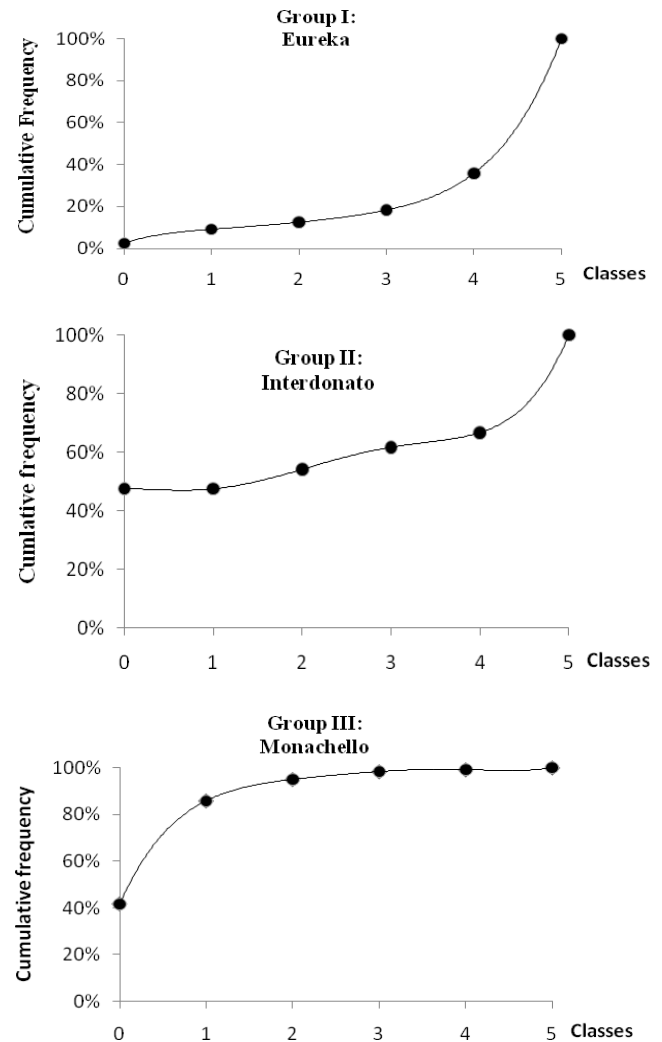


Figure 1: The polynomial regression fitted curves

The response of the studied cultivars, as it's recorded and analyzed by the mathematical model reflects the natural behavior of the tested plant in the orchards.

Taking into account the obtained results, the model proved to be an efficient new method for the resistance screening in the host parasite biological interaction system of *citrus limon* cultivars and their pathogenic fungi *Phoma tracheiphila*.

3 Conclusions

The factors most likely to influence the results of the phytopathological tests are mainly the techniques of the inoculation and the methods of the analysis used. As the procedures vary between the laboratories the results also

differ even for the same tested biological sample. In order to overpass these difficulties and standardize the protocols of the disease assessment, the biomathematical model offer an appropriate solution.

Extending the results of this study to others groups of resistance in Citrus plants its needed to cover the whole specter of the host plants of the parasite, from the very susceptible one to the highly resistance.

4 References

- [1] Agrios GN (2005). Plant Pathology. Fifth Edition, Elsevier Academic Press, London, UK. Royle DJ, Ostry ME (1995). Disease and pest control in the bioenergy crops poplar and willow. *Biomass Bioenergy*, 9: 69-79.
- [2] Gulsen O., Uzun A., Pala, E H.. Canihos and Kafa G.. 2007. Development of seedless and Mal Secco tolerant mutant lemons through bud wood irradiation. *Scientia Horticulturae* Volume 112, Issue 2, Pages 184-190.
- [3] Jeger MJ (2004). Analysis of disease progress as a basis for evaluating disease management practices. *Annu. Rev. Phytopatol.* 42: 61-82.
- [4] Van Maanen A, Xu XM (2003). Modelling plant disease epidemics. *European J. Plant Pathol.* 109: 669-682.
- [5] Van der Plank JE (1963). Plant disease: epidemics and control. Academic Press, NY., USA. eslie Lamport. "LaTeX: A Document Preparation System". Addison-Wesley Publishing Company, 1986.
- [6] Segarra J, Jeger MJ, Van den Bosch F (2001). Epidemic dynamics and patterns of plant diseases. *Phytopathology*, 91: 1001-1010.
- [7] Slolel, Z., and Salerno, M. 1988. Mal secco. Pages 18-20 in: compendium of citrus Diseases. J.O. Whiteside, S. M. Garnsey, and L.W. Timmer, eds. American Phytopathological Society, St. Paul, MN.
- [8] Hajlaoui M.R., Kanchouch K., Guermech A. et Cherif M, 2000 a. Recherche sur l'agent du mal secco des Citrus, *Phoma tracheiphila* : I- Etude de quelques caractéristiques biologiques du parasite. *Annales de l'Institut Nationale de la Recherche Agronomique de Tunisie*. Vol 73: 163-182.
- [9] Gareth Williams, *Linear Algebra with Applications*, Seventh Edition, Jones & Bartlett Publishers, 2009, 554 pages.
- [10] Stephen Wolfram, *The MATHEMATICA ® Book*, Version 4, Cambridge University Press, 1999, 1469 pages.

SESSION

**ALGEBRAIC BIOLOGY AND BIOINFORMATICS,
ABB**

Chair(s)

Prof. Matthew He

The Genetic Code, 8-Dimensional Hypercomplex Numbers and Dyadic Shifts

Sergey V. Petoukhov

Department of Biomechanics, Mechanical Engineering Research Institute of RAS, Moscow, Russia

Abstract - The article is devoted to algebraic features of structural phenomena of molecular ensembles of the genetic code. Matrix forms of presentations of the genetic code allow showing deep relations of the genetic code with dyadic shifts and algebras of 8-dimensional hypercomplex numbers. Hadamard matrices and orthogonal systems of Rademacher and Walsh functions, which are well-known formalisms from discrete signal processing, participate in this discovery of hidden structural features of the genetic code. The described results are useful to understand a non-casual character of the genetic code systems, which has a deep algebraic nature. The results lead to new theoretical approaches in the field of algebraic biology.

Keywords: Code, Hypercomplex Numbers, Dyadic Shifts

1 Introduction

A biological meaning of genetic informatics is reflected in the brief statement: "life is a partnership between genes and mathematics" [22]. We are trying to find math which is a partner of the genetic code. One of the possible directions of search is to use matrix forms of presentation and analysis of ensembles of molecular elements of the genetic code. Matrix representations and methods are widely and successfully used in the theory of error-correcting coding and processing of information, theoretical physics, computer science, the theory of hypercomplex numbers, etc. In this regard, a scientific field called "Matrix genetics" exists, which studies the matrix presentation of the genetic code, including through borrowing matrix methods from the field of digital signal processing [10, 11, 14, 15, 17]. Our results are a part of "algebraic biology", which gave rise to thematic conferences and international societies; the journal "Bulletin of Mathematical Biology" identifies this area as a separate category.

This article is devoted to author's results on algebraic features of structural phenomena of molecular ensembles of the genetic code. More precisely it shows relations of the genetic code with dyadic shifts, algebras of 8-dimensional hypercomplex numbers, Hadamard matrices, orthogonal systems of Rademacher and Walsh functions and the sequency theory by Harmuth [6-9].

BIOCAMP. Manuscript received March 9, 2011. This work was supported in part by the Russian Federal Agency of Science and Innovations (the contract № 02.740.11.0100) and by the Russian Federal Agency on Education (the contract № P377).

2 Genetic matrices, dyadic shifts, Rademacher functions and 8-dimensional hypercomplex numbers

The four letters of the genetic alphabet A (adenine), C (cytosine), G (guanine), U/T (uracil in RNA or thymine in DNA) represent specific poly-atomic constructions. The set of these four constructions bears the substantial symmetric system of distinctive-uniting attributes (or, more precisely, pairs of "attribute-antiattribute"). The system of such attributes divides the genetic four-letter alphabet into the following three pairs of letters, which are equivalent from a viewpoint of one of these attributes or its absence: 1) C = U & A = G (according to the binary-opposite attributes: "pyrimidine" or "non-pyrimidine", that is purine); 2) A = C & G = U (according to the attributes "keto" or "amino"); 3) C = G & A = U (according to the attributes: three or two hydrogen bonds are materialized in these complementary pairs). The possibility of such division of the genetic alphabet into three binary sub-alphabets is known from the work [12]. We utilize these known sub-alphabets in the field of matrix genetics which studies matrix forms of presentation of the genetic code. Let us mark these three kinds of binary-opposite attributes by numbers N = 1, 2, 3 and ascribe to each of the four genetic letters the symbol "0_N" (the symbol "1_N") in a case of presence (of absence correspondingly) of the attribute under number "N" in this letter. As a result we obtain the representation of the genetic four-letter alphabet in the system of its three "binary sub-alphabets corresponding to attributes" (Fig. 1).

	Symbols of genetic letters from a viewpoint of binary-opposite attributes	C	A	G	U/T
№1	0 ₁ – pyrimidine (one molecular ring); 1 ₁ – purine (two rings in a molecule)	0 ₁	1 ₁	1 ₁	0 ₁
№2	0 ₂ – amino; 1 ₂ – keto	0 ₂	0 ₂	1 ₂	1 ₂
№3	0 ₃ – a letter with three hydrogen bonds; 1 ₃ – a letter with two hydrogen bonds	0 ₃	1 ₃	0 ₃	1 ₃

Fig. 1. Three binary sub-alphabets according to three kinds of binary-opposite attributes in a set of nitrogenous bases C, A, G, U.

On the basis of the idea about a possible analogy between discrete signals processing in computers and in a genetic code system, one can present the genetic 4-letter alphabet in the following matrix form [C A; U G] (Fig. 2). Then the Kronecker family of matrices with such alphabetical kernel can be considered: [C A; U G]⁽ⁿ⁾, where (n) means the integer Kronecker (or tensor) power [11, 14, 15, 17]. The matrix [C A; U G]⁽³⁾ contains 64 triplets in a strict order (Fig. 2).

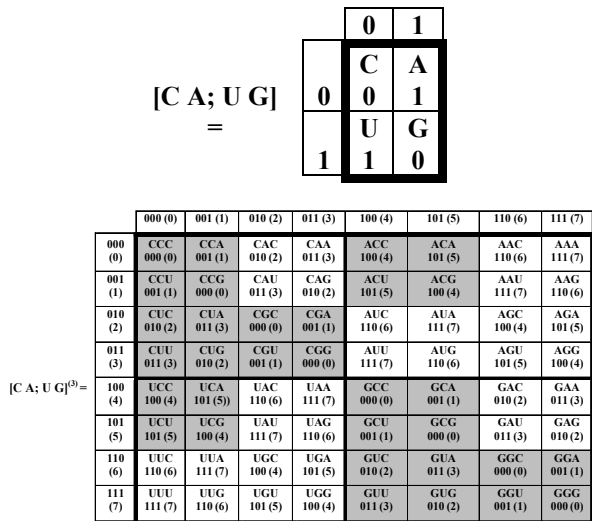


Fig. 2. Genetic matrices [C A; U G] and [C A; U G]⁽³⁾ with binary numerations of their columns and rows on the base of the binary sub-alphabets № 1 and № 2 from Fig. 1. Matrix cells contain a symbol of a multiplet, a dyadic-shift numeration of this multiplet and its expression in decimal notation. Decimal numerations of columns, rows and multiplets are written in brackets. Black and white cells contain triplets with strong and weak roots correspondingly (see the text).

All the columns and rows of the matrices on Fig. 2 are binary numerated and disposed in a monotonic order by the following algorithm which uses biochemical features of the genetic nitrogenous bases and which can be used in bio-computers of any organism really. Numerations of columns and rows are formed automatically if one interprets multiplets of each column from the viewpoint of the first binary sub-alphabet (Fig. 1) and if one interprets multiplets of each row from the viewpoint of the second binary sub-alphabet. For example, the column 010 contains all the triplets of the form "pyrimidine-purine-pyrimidine"; the row 010 contains all the triplets of the form "amino-keto-amino". Each of the triplets in the matrix [C A; U G]⁽³⁾ receives its dyadic-shift numeration by means of modulo-2 addition of binary numerations of its column and row. Here one should explain that this kind of addition is one of the main operations in digital signal processing; by definition the modulo-2 addition of two numbers written in binary notation is made in a bitwise manner in accordance with the rules:

$$0 + 0 = 0, 0 + 1 = 1, 1 + 0 = 1, 1 + 1 = 0 \quad (1)$$

For example, the triplet CAG receives its dyadic-shift numeration 010 (or 2 in decimal notation) because it belongs to the column 011 and the row 001. The series of binary numbers

$$000, 001, 010, 011, 100, 101, 110, 111 \quad (2)$$

forms a diadic group, in which modulo-2 addition serves as the group operation [9]. The distance in this symmetry group is known as the Hamming distance. Since the Hamming distance satisfies the conditions of a metric group, the diadic group is a metric group. The modulo-2 addition of any two binary numbers from (2) always results in a new number

from the same series. The number 000 serves as the unit element of this group. The reverse element for any number in this group is the number itself. Changes in the initial binary sequence (2), produced by modulo-2 addition of its members with any binary numbers (2), are termed diadic shifts [1, 9]. If any system of elements demonstrates its connection with diadic shifts, it indicates that the structural organization of its system is related to the logic of modulo-2 addition. This article gives some evidences that the genetic code is related to the logic of modulo-2 addition.

Black and white cells in the genomatrix [C A; U G]⁽³⁾ reflect the following peculiarities of the genetic code. A combination of letters on the two first positions of each triplet is termed a "root" of this triplet; a letter on its third position is termed a "suffix". The set of 64 triplets contains 16 possible variants of such roots. Taking into account properties of triplets, the set of 16 possible roots is divided into two subsets with 8 roots in each. The first of such octets contains roots CC, CU, CG, AC, UC, GC, GU, GG. These roots are termed "strong roots" [13] because each of them defines four triplets with this root, coding values of which are independent on their suffix. For example, four triplets CGC, CGA, CGU, CGG, which have the strong root CG, encode the same amino acid Arg, although they have different suffixes (Fig. 3). The second octet contains roots CA, AA, AU, AG, UA, UU, UG, GA. These roots are termed "weak roots" because each of them defines four triplets with this root, coding values of which depend on their suffix. An example of such a subfamily in Fig. 3 is represented by four triplets CAC, CAA, CAU and CAG, two of which (CAC, CAU) encode the amino acid His and the other two of which (CAA, CAG) encode the amino acid Gln.

THE STANDARD CODE	
8 subfamilies of triplets with strong roots ("black triplets") and the amino acids, which are encoded by them	8 subfamilies of triplets with weak roots ("white triplets") and the amino acids, which are encoded by them
CCC, CCU, CCA, CCG → Pro	CAC, CAU, CAA, CAG → His, His, Gln, Gln
CUC, CUU, CUA, CUG → Leu	AAC, AAU, AAA, AAG → Asn, Asn, Lys, Lys
CGC, CGU, CGA, CGG → Arg	AUC, AUU, AUA, AUG → Ile, Ile, Ile, Met
ACC, ACU, ACA, ACG → Thr	AGC, AGU, AGA, AGG → Ser, Ser, Arg, Arg
UCC, UCU, UCA, UCG → Ser	UAC, UAU, UAA, UAG → Tyr, Tyr, Stop, Stop
GCC, GCU, GCA, GCG → Ala	UUC, UUU, UUA, UUG → Phe, Phe, Leu, Leu
GUC, GUU, GUA, GUG → Val	UGC, UGU, UGA, UGG → Cys, Cys, Trp, Trp
GGC, GGU, GGA, GGG → Gly	GAC, GAU, GAA, GAG → Asp, Asp, Glu, Glu
THE VERTEBRATE MITOCHONDRIAL CODE	
CCC, CCU, CCA, CCG → Pro	CAC, CAU, CAA, CAG → His, His, Gln, Gln
CUC, CUU, CUA, CUG → Leu	AAC, AAU, AAA, AAG → Asn, Asn, Lys, Lys
CGC, CGU, CGA, CGG → Arg	AUC, AUU, AUA, AUG → Ile, Ile, Met, Met
ACC, ACU, ACA, ACG → Thr	AGC, AGU, AGA, AGG → Ser, Ser, Stop, Stop
UCC, UCU, UCA, UCG → Ser	UAC, UAU, UAA, UAG → Tyr, Tyr, Stop, Stop
GCC, GCU, GCA, GCG → Ala	UUC, UUU, UUA, UUG → Phe, Phe, Leu, Leu
GUC, GUU, GUA, GUG → Val	UGC, UGU, UGA, UGG → Cys, Cys, Trp, Trp
GGC, GGU, GGA, GGG → Gly	GAC, GAU, GAA, GAG → Asp, Asp, Glu, Glu

Fig. 3. The Standard Code and the Vertebrate Mitochondrial Code possess the basic scheme of the genetic code degeneracy with 32 triplets of strong roots and 32 triplets of weak roots (Initial data from <http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>).

How these two subsets of triplets with strong and weak roots are disposed in the genomatrix [C A; U G]⁽³⁾ (Fig. 2) which was constructed formally on the base of the genetic alphabet and Kronecher multiplications without any mention about the degeneracy of the genetic code and about amino acids? Can one anticipate any symmetry in their disposition? It should be noted that the huge quantity 64! ≈ 10⁸⁹ of variants exists for dispositions of 64 triplets in the (8x8)-

matrix. One can note for comparison, that the modern physics estimates time of existence of the Universe in 10^{17} seconds. In such a situation an accidental disposition of the 20 amino acids and the corresponding triplets in a (8x8)-matrix will give almost never any symmetry in their disposition in matrix halves, quadrants and rows.

But it is phenomenological fact that the disposition of the 32 triplets with strong roots ("black triplets" in Fig. 2) and the 32 triplets with weak roots ("white triplets") has a symmetric character unexpectedly (see Fig. 2). For example the left and right halves of the matrix mosaic are mirror-anti-symmetric to each other in its colors: any pair of cells, disposed by mirror-symmetrical manner in these halves, possesses the opposite colors. One can say that each row of this mosaic matrix corresponds to an odd function. In addition each row of the mosaic matrix [C A; U G]⁽³⁾ has a meander-line character (the term "meander-line" means here that lengths of black and white fragments are equal to each other along each row). But the theory of discrete signal processing uses such odd meander functions for a long time under the name "Rademacher functions". Rademacher functions contain elements "+1" and "-1" only. Each of the matrix rows presents one of the Rademacher functions if each black (white) cell is interpreted such that it contains the number +1 (-1). Fig. 4 shows a transformation of the mosaic matrix [C A; U G]⁽³⁾ (Fig. 2) into a numeric matrix in the result of such replacements of black and white triplets by means of numbers "+1" and "-1" correspondingly.

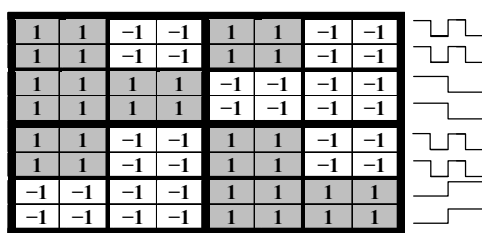


Fig. 4. Rademacher form R of presentation of the genomatrix [C A; U G]⁽³⁾ from Fig. 2. A relevant system of Rademacher functions is shown at the right side.

The Rademacher form R of the genomatrix [C A; U G]⁽³⁾ (Fig. 4) can be decomposed into sum of 8 sparse matrices $r_0, r_1, r_2, r_3, r_4, r_5, r_6, r_7$ (Fig. 5) in accordance with the principle of dyadic-shifts numerations of cells and triplets from Fig. 2. More precisely any sparse matrix r_k ($k=0, 1, \dots, 7$) contains entries "+1" or "-1" from the matrix R on Fig. 4 in those cells which correspond to cells with the same dyadic-shift numeration "k" of triplets on Fig. 2; all the other cells of the matrix r_k contain zero.

The author has revealed that this set of 8 matrices r_0, r_1, \dots, r_7 (where r_0 is identity matrix) is closed relative to multiplication and it satisfies the table of multiplication on Fig. 6.

The multiplication table on Fig. 6 is asymmetrical relative to the main diagonal and corresponds to the non-commutative associative algebra of 8-dimensional hypercomplex numbers. This matrix algebra is non-division algebra because it has zero divisors. It means that such non-zero hypercomplex numbers exist whose product is equal to

zero. These genetic 8-dimensional hypercomplex numbers are different from Cayley's octonions (<http://en.wikipedia.org/wiki/Octonion>). The algebra of Cayley's octonions is non-associative algebra and correspondingly it does not possess a matrix form of its presentation (each of matrix algebras is an associative algebra). The known term "octonions" is not appropriate for the case of the multiplication table on Fig. 6 because this term is usually used for members of normed division non-associative algebra (<http://en.wikipedia.org/wiki/Octonion>).

$$R = r_0 + r_1 + r_2 + r_3 + r_4 + r_5 + r_6 + r_7 =$$

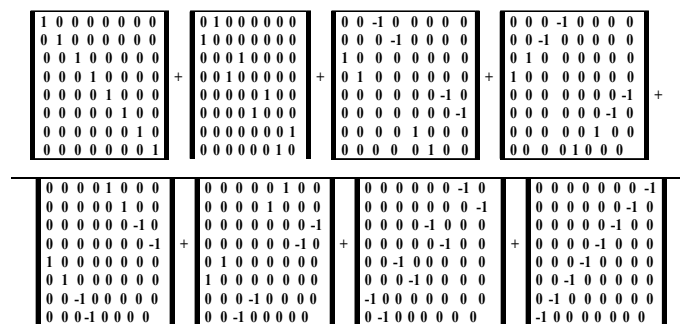


Fig. 5. The dyadic-shift decomposition of the Rademacher form R (Fig. 4) of the genomatrix [C A; U G]⁽³⁾ into sum of 8 sparse matrices r_0, r_1, \dots, r_7 .

	1	r ₁	r ₂	r ₃	r ₄	r ₅	r ₆	r ₇
1	1	r ₁	r ₂	r ₃	r ₄	r ₅	r ₆	r ₇
r ₁	r ₁	1	r ₃	r ₂	r ₅	r ₄	r ₇	r ₆
r ₂	r ₂	r ₃	-1	-r ₁	-r ₆	-r ₇	r ₄	r ₅
r ₃	r ₃	r ₂	-r ₁	-1	-r ₇	-r ₆	r ₅	r ₄
r ₄	r ₄	r ₅	r ₆	r ₇	1	r ₁	r ₂	r ₃
r ₅	r ₅	r ₄	r ₇	r ₆	r ₁	1	r ₃	r ₂
r ₆	r ₆	r ₇	-r ₄	-r ₅	-r ₂	-r ₃	1	r ₁
r ₇	r ₇	r ₆	-r ₅	-r ₄	-r ₃	-r ₂	r ₁	1

Fig. 6. The multiplication table of basic matrices r_0, r_1, \dots, r_7 (where r_0 is identity matrix) which corresponds to the 8-dimensional algebra over the field of real numbers. It defines the 8-dimensional numeric system of genetic R_{123} -octetons.

For this reason we term these hypercomplex numbers, which are revealed in matrix genetics, as "dyadic-shift genetic octetons" (or briefly "octetons"). In addition we term such kinds of matrix algebras, which are connected with dyadic-shift decompositions, as dyadic-shift algebras (or briefly DS-algebras). The author supposes that DS-algebras are important for genetic systems. All the basic matrices r_0, r_1, \dots, r_7 are disposed in the multiplication table (Fig. 6) in accordance with dyadic-shift numerations of cells on Fig. 2.

Below we will describe another variant of genetic octetons which is connected with Hadamard genomatrices. For this reason we term the first type of geno-octetons (Fig. 4-6) as R_{123} -octetons (here R is the first letter of the name Rademacher; the index 123 means the order 1-2-3 of positions in triplets).

A general form of R_{123} -octetons (Fig. 5) is the following:

$$R_{123} = x_0 * \mathbf{1} + x_1 * \mathbf{r}_1 + x_2 * \mathbf{r}_2 + x_3 * \mathbf{r}_3 + x_4 * \mathbf{r}_4 + x_5 * \mathbf{r}_5 + x_6 * \mathbf{r}_6 + x_7 * \mathbf{r}_7 \quad (4)$$

where coefficients x_0, x_1, \dots, x_7 are real numbers. Here the first component x_0 is a scalar. Other 7 components $x_1 \cdot \mathbf{r}_1, x_2 \cdot \mathbf{r}_2, x_3 \cdot \mathbf{r}_3, x_4 \cdot \mathbf{r}_4, x_5 \cdot \mathbf{r}_5, x_6 \cdot \mathbf{r}_6, x_7 \cdot \mathbf{r}_7$ are non-scalar units but imaginary units. Some properties of these octetons lead to the idea that for a system of genetic coding the main significance belong not to the entire set of possible real values of coordinates of 8-dimensional hypercomplex numbers but only to the subset of numbers $2^0, 2^1, 2^2, \dots, 2^n, \dots$ [16]. It seems that for genetic systems DS-algebras are algebras of dichotomous biological processes and systems.

3 Permutations and the DS-algebra

The theory of discrete signal processing pays a special attention to permutations of information elements. This paragraph shows that all the possible permutations of positions inside all the triplets lead to new mosaic genomatrices whose Rademacher forms of presentation are connected with the same DS-algebra (Fig. 6).

A simultaneous permutation of positions in triplets transforms the most of the triplets in cells of the initial genomatrix $[C A; U G]^{(3)}$. For example, in the case of the cyclic transformation of the order 1-2-3 of positions into the order 2-3-1, the triplet CAG is transformed into the triplet AGC, etc. Because each of the triplets is connected with the binary numeration of its column and row, these binary numerations are also transformed correspondingly; for example, the binary numeration 011 is transformed into 110. The six variants of the order of positions inside triplets are possible: 1-2-3, 2-3-1, 3-1-2, 3-2-1, 2-1-3, 1-3-2. The initial genomatrix $[C A; U G]_{123}^{(3)}$ is related with the first of these orders (Fig. 4). Other five genomatrices $[C A; U G]_{231}^{(3)}, [C A; U G]_{312}^{(3)}, [C A; U G]_{321}^{(3)}, [C A; U G]_{213}^{(3)}, [C A; U G]_{132}^{(3)}$, which correspond to other five orders, are shown on Fig. 7 (subscripts indicate the order of positions in triplets).

In these genomatrices on Fig. 7 black-and-white mosaics of each row corresponds again to one of Rademacher functions. The replacement of all the triplets with strong and weak roots by entries “+1” and “-1” correspondingly transforms these genomatrices into their Rademacher forms $R_{231}, R_{312}, R_{321}, R_{213}, R_{132}$. Each of the Rademacher forms $R_{231}, R_{312}, R_{321}, R_{213}, R_{132}$ can be again decomposed into sum of 8 sparse matrices $r_0, r_1, r_2, r_3, r_4, r_5, r_6, r_7$ in accordance with dyadic-shift numerations of its cells (see details in [16]). Each of the 6 sets with eight sparse matrices $r_0, r_1, r_2, r_3, r_4, r_5, r_6, r_7$ is unique and different from other sets (r_0 is identity matrix in all the sets).

Unexpected facts are that, firstly, each of these sets is closed relative multiplication and, secondly, each of these sets corresponds to the same multiplication table from Fig. 6.

It means that this genetic DS-algebra of 8-dimensional hypercomplex numbers possesses at least 5 additional matrix forms of its presentation. Our results demonstrate that this DS-algebra of genetic R-octetons possesses a wonderful invariance relative not only to all the variants of positional permutations in triplets but also to some other permutations which are connected with Gray code and dyadic-shift transformations [16]. All the properties of R_{123} -octetons hold true in the cases of different matrix forms of presentation of R-octetons with the same multiplication table (Fig. 6).

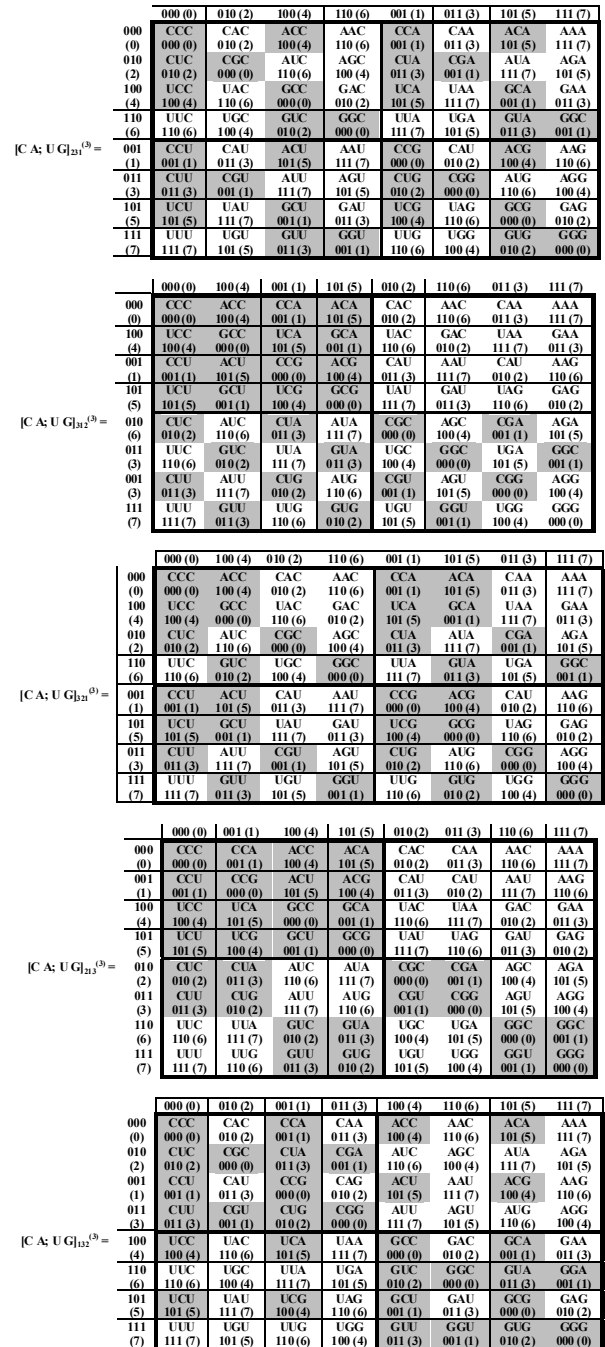


Fig. 7. Five genomatrices $[C A; U G]_{231}^{(3)}, [C A; U G]_{312}^{(3)}, [C A; U G]_{321}^{(3)}, [C A; U G]_{213}^{(3)}, [C A; U G]_{132}^{(3)}$, which correspond to orders of positions in triplets 2-3-1, 3-1-2, 3-2-1, 2-1-3, 1-3-2 relative to the genomatrix $[C A; U G]_{123}^{(3)}$ on Fig. 2. Black and white cells contain triplets with strong and weak roots correspondingly. Binary numerations of columns and rows are shown.

The analysis of evolution of variants (or dialects) of the genetic code from the viewpoint of the DS-algebra of the R-octetons has allowed revealing two phenomenological rules [16]:

Rule #1. In all the organisms with sexual reproduction only those triplets can be involved in the evolutionary changing their correspondence to amino acids or to stop-signals, which possess dyadic-shift numerations 4, 5, 6, 7 in the genomatrix $[C A; U G]^{(3)}$ (Fig. 2); in other words, only

those triplets can be involved which are connected with the basic matrices r_4, r_5, r_6, r_7 (Fig. 5) of genetic R-octetons.

Rule #2. In all the dialects of the genetic code only triplets with dyadic-shift numerations 2, 6, 7 can be start-codons. In other words, only those triplets can be start-codons, which are connected with the basic matrices r_2, r_6, r_7 (Fig. 5) of genetic R-octetons.

4 Hadamard genomatrices and another DS-algebra

By definition a Hadamard matrix of dimension “n” is the $(n*n)$ -matrix $H(n)$ with elements “+1” and “-1”. It satisfies the condition $H(n)*H(n)^T = n*I_n$, where $H(n)^T$ is the transposed matrix and I_n is the identity $(n*n)$ -matrix. Rows of Hadamard matrices are termed Walsh functions. Hadamard matrices are widely used in error-correcting codes such as the Reed-Muller code and Hadamard codes; in the theory of compression of signals and images; in spectral analysis and multi-channel spectrometers with Hadamard transformations; in quantum computers with Hadamard gates; in a realization of Boolean functions by means of spectral methods; in the theory of planning of multiple-factor experiments and in many other branches of science and technology. The works [10, 14, 15] have revealed that Kronecker families of genetic matrices are related with some kinds of Hadamard matrices (“Hadamard genomatrices”) by means of so termed U-algorithm. This paragraph describes that the dyadic-shift decompositions of Hadamard genomatrices lead to special 8-dimensional hypercomplex numbers. For the U-algorithm, phenomenological facts are essential that the letter U in RNA (and correspondingly the letter T in DNA) is a unique letter in the genetic alphabet in the two following senses:

- Each of three nitrogenous bases A, C, G has one amino-group NH_2 , but the fourth basis U/T has not it. From the viewpoint of existence of the amino-group (which is very important for genetic functions) the letters A, C, G are identical to each other and the letter U is opposite to them;
- The letter U is a single letter in RNA, which is replaced in DNA by another letter T.

This uniqueness of the letter U can be utilized in genetic computers of organisms. Taking into account this unique status of the letter U, the author has revealed the existence of the following formal “U-algorithm”, which demonstrates the close connection between Hadamard matrices and the matrix mosaic of the genetic code [10, 14, 15, 17].

By definition the U-algorithm contains two steps: 1) on the first step, each of the triplets in the black-and-white genomatrix (for example, in the genomatrix $[C A; U G]^{(3)}$ on Fig. 2) should change its own color into opposite color each time when the letter U stands in an odd position (in the first or in the third position) inside the triplet; 2) on the second step, black triplets and white triples are interpreted as entries “+1” and “-1” correspondingly. For example, the white triplet UUA (see Fig. 2) should become the black triplet (and its matrix cell should be marked by black color) because of the letter U in its first position; for this reason the triplet UUA is interpreted finally as “+1”. Or the white triplet UUU

should not change its color because of the letter U in its first and third positions (the color of this triplet is changed twice according to the described algorithm); for this reason the triplet UUU is interpreted finally as “-1”.

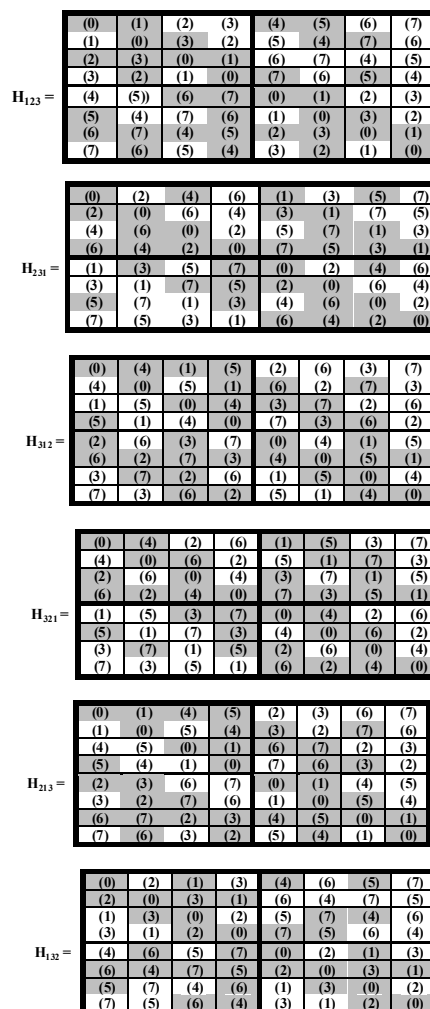


Fig. 8. The Hadamard genomatrices $H_{123}, H_{231}, H_{312}, H_{321}, H_{213}, H_{132}$ which are received from the genomatrices $[C A; U G]_{123}^{(3)}, [C A; U G]_{231}^{(3)}, [C A; U G]_{312}^{(3)}, [C A; U G]_{321}^{(3)}, [C A; U G]_{213}^{(3)}, [C A; U G]_{132}^{(3)}$ (Fig. 2 and 7) by means of the U-algorithm. Brackets contain dyadic-shift numerations of cells in decimal notation by analogy with matrices on Fig. 2 and 8. Black color and white color of cells mean entries “+1” and “-1” in these cells correspondingly.

By means of the U-algorithm, all the genomatrices $[C A; U G]_{123}^{(3)}, [C A; U G]_{231}^{(3)}, [C A; U G]_{312}^{(3)}, [C A; U G]_{321}^{(3)}, [C A; U G]_{213}^{(3)}, [C A; U G]_{132}^{(3)}$ (Fig. 2 and 7) are transformed into relevant numeric genomatrices $H_{123}, H_{231}, H_{312}, H_{321}, H_{213}, H_{132}$ on Fig. 8.

One can make the dyadic-shift decomposition of each of these six Hadamard genomatrices $H_{123}, H_{231}, H_{312}, H_{321}, H_{213}, H_{132}$ (Fig. 8) by analogy with the described decompositions of the genomatrices $R_{123}, R_{231}, R_{312}, R_{321}, R_{213}, R_{132}$. In the result six new different sets of 8 sparse matrices $h_0, h_1, h_2, h_3, h_4, h_5, h_6, h_7$ arise (where h_0 is identity matrix). It is unexpectedly but each of these six sets for Hadamard genomatrices is closed relative to multiplication. Moreover each of these sets $h_0, h_1, h_2, h_3, h_4, h_5, h_6, h_7$ corresponds to

the same multiplication table on Fig. 9 [16].

	1	h ₁	h ₂	h ₃	h ₄	h ₅	h ₆	h ₇
1	1	h ₁	h ₂	h ₃	h ₄	h ₅	h ₆	h ₇
h ₁	h ₁	-1	h ₃	-h ₂	h ₅	-h ₄	h ₇	-h ₆
h ₂	h ₂	h ₃	-1	-h ₁	-h ₆	-h ₇	h ₄	h ₅
h ₃	h ₃	-h ₂	-h ₁	1	-h ₇	h ₆	h ₅	-h ₄
h ₄	h ₄	h ₅	h ₆	h ₇	-1	-h ₁	-h ₂	-h ₃
h ₅	h ₅	-h ₄	h ₇	-h ₆	-h ₁	1	-h ₃	h ₂
h ₆	h ₆	h ₇	-h ₄	-h ₅	h ₂	h ₃	-1	-h ₁
h ₇	h ₇	-h ₆	-h ₅	h ₄	h ₃	-h ₂	-h ₁	1

Fig. 9. The multiplication table for the dyadic-shift decompositions of Hadamard genomatrices H₁₂₃, H₂₃₁, H₃₁₂, H₃₂₁, H₂₁₃, H₁₃₂ (Fig. 8).

The existence of the multiplication table (Fig. 9) means that a new 8-dimensional DS-algebra or a new system of 8-dimensional hypercomplex numbers exists on the base of these Hadamard genomatrices which are connected with six different matrix forms of presentation of this hypercomplex system. We term these new 8-dimensional hypercomplex numbers as H-octetons (here “H” is the first letter in the name Hadamard) because they differ from R-octetons (Fig. 6) and Cayley’s octonions. The six Hadamard genomatrices H₁₂₃, H₂₃₁, H₃₁₂, H₃₂₁, H₂₁₃, H₁₃₂ are different matrix forms of presentation of the same H-octeton whose coordinates are equal to 1 (x₀=x₁=...=x₇=1).

The DS-algebra of H-octetons (Fig. 9) is the non-commutative associative non-division algebra. It has zero divisors: for example (h₃+h₄) and (h₂-h₅) are non-zero H-octetons, but their product is equal to zero. The quantity and the disposition of signs “+” and “-“ in the multiplication table on Fig. 9 are identical to their quantity and disposition in a Hadamard matrix. In addition, indexes of basic matrices are again disposed in the multiplication table (Fig. 9) in accordance with the dyadic-shift numeration on Fig. 2.

It should be noted that Hadamard matrices play important roles in many tasks of discrete signal processing; they are devoted to tens of thousands of publications (see a review in [19]). Only a few symmetrical Hadamard matrices are usually used in the field of discrete signal processing. But dyadic-shift decompositions of these “engineering” Hadamard matrices do not lead to any 8-dimensional hypercomplex numbers in contrast to the asymmetrical Hadamard genomatrices described in our article. Moreover the author knows no publications about the facts that Hadamard matrices can be the base for matrix forms of presentation of 8-dimensional hypercomplex numbers. It seems that the genetic code has led the author to discovering the new interesting fact in the field of the theory of Hadamard matrices about the unexpected relation of some Hadamard matrices with multidimensional DS-algebras and their systems of hypercomplex numbers. This fact can be useful for many applications of Hadamard genomatrices for simulating of bioinformation phenomena, for technology of discrete signal processing, etc. A great number of Hadamard (8x8)-matrices exists (according to some experts, their number is equal to approximately 5 billion). Perhaps, only the genetic Hadamard matrices, which represent a small

subset of a great set of all the Hadamard matrices, are related with multidimensional DS-algebras but it is an open question now.

Why living nature uses just such the genetic code that is associated with Hadamard genomatrices? We suppose that its reason is related with solving in biological organisms the same information tasks which lead to a wide using of Hadamard matrices in digital signal processing and in physics.

5 Discussion

The author has revealed a close relation of the genetic code with 8-dimensional hypercomplex numbers (first of all, R-octetons and H-octetons) and with dyadic shifts and Hadamard matrices. This relation is interesting in many aspects. Some of them are the following.

Numeric presentations of genetic sequences are useful to study hidden genetic regularities [3, 4, 44, 17, etc.]. On the base of the described results, new approaches of numeric presentations of genetic sequences can be proposed for such aims taking into account additionally known applications of hypercomplex numbers to analysis of genetic sequences [2, 5, 20, 21, 23, etc.]. It seems appropriate to interpret genetic sequences as sequences of 8-dimensional vectors where genetic elements are replaced by their special numeric presentations which are connected with the described DS-algebras. Then Hadamard spectrums, dyadic distances and some other characteristics of these vector sequences can be studied. If the quantity of vector elements in a genetic sequence is not divisible by 8, the remaining short vector can be extended to an 8-dimensional vector by adding to its end of the required number of zeros by analogy with methods of digital signal processing.

Walsh functions play the main role in the fruitful sequency theory by Harmuth for signal processing [6-9]. Rows of Hadamard genomatrices correspond to special kinds of Walsh functions which define special variants of sequency analysis. The author believes that this “genetic” sequency analysis can be a key to understand important features not only of genetic informatics but also of many other inherited physiological systems (morphogenetic, sensori-motor, etc.). In comparison with spectral analysis by means of sine waves, which is applicable to linear time-invariant systems, the sequency analysis is based on non-sinusoidal waves and it is used to study systems which are changed in time (biological systems belong to such systems) [7, 9]. Genetic DS-algebras can also be useful in a realization of the famous idea by Boole on algebraic theory of laws of thinking. The author believes that mechanisms of biological morphogenesis are closely associated with spatial and temporal filters from the field of sequency analysis for genetic systems. Taking into account the sequency theory by Harmuth together with our data about Hadamard genomatrices and genetic H-octetons, one can assume that biological evolution can be interpreted largely like the evolution of physiological spatial and temporal filters of the sequency theory.

The notion “number” is the main notion of mathematics. In modern theoretical physics, systems of 8-dimensional

hypercomplex numbers (mainly, Cayley's octonions and split-octonions) are one of important objects. The discovery of the relation of the genetic code with special types of 8-dimensional hypercomplex numbers allows generating of heuristic associations between theoretical physics and mathematical biology. The described DS-algebras can be useful for development of algebraic biology [16].

Bioinformatics should solve many problems about inherited properties of biological bodies:

- Noise-immunity property of genetic coding;
- Management and synchronization of a huge number of inherited cyclic processes;
- Doubling of bio-information (mitosis, etc);
- Compression of inherited biological data;
- Spatial and temporal filtering of genetic information;
- Primary structure of proteins;
- Multi-channel informatics;
- Hidden rules of structural interrelations among parts of genetic systems;
- Laws of evolution of dialects of the genetic code, etc.

The principle of dyadic shifts and DS-algebras of genetic octetons can be useful for many of these problems.

In addition, one can mention about known facts of analogies between the genetic code and the symbolic system of ancient Chinese book "I Ching" (see a review in [17]). This symbolic system is a base of many branches of Oriental medicine including acupuncture, Tibetan pulse diagnostics, etc. which use ancient ideas of "I Ching" about inherited physiological systems. Using dyadic shifts for studying not only the genetic code but also the mysterious tables of "I Ching" reveals the hidden regularities and symmetrical patterns in this ancient system [16]. Results of matrix genetics give new approaches for better understanding the "I Ching".

6 References

- [1] N. U. Ahmed, K. R. Rao, *Orthogonal transforms for digital signal processing*, Springer-Verlag New York, Inc., 1975.
- [2] T. Bulow, "Non-commutative Hypercomplex Fourier Transforms", in *Geometric Computing with Clifford Algebras*, T. Bulow, M. Felsberg, G. Sommer, Ed. Berlin: Springer-Verlag, 2001, pp. 187-207.
- [3] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals", *J. Cell Mol. Med.*, vol. 6, no. 2, 2002, pp. 279-303.
- [4] P. D. Cristea, "Symmetries in genomics", *Symmetry: Culture and Science*, vol. 21 no.1-3, 2010, pp. 71-86.
- [5] M. Felberg, "Commutative Hypercomplex Fourier Transforms of Multidimensional Signals", in *Geometric computing with Clifford algebras*, T.Bulow, M.Felsberg, G.Sommer, Ed. Berlin: Springer-Verlag, 2001, pp. 209-229.
- [6] H. F. Harmuth, *Transmission of information by orthogonal functions*. Berlin: Springer, 1970.
- [7] H. F. Harmuth, *Sequency theory. Foundations and applications*. N.-Y.: Academic Press, 1977.
- [8] H. F. Harmuth, *Nonsinusoidal waves for radar and radio communication*. N.-Y.: Academic Press, 1981.
- [9] H. F. Harmuth, *Information theory applied to space-time physics*. Washington: The Catholic University of America, DC, 1989.
- [10] M. He, S.V. Petoukhov, "The genetic code, Hadamard matrices and algebraic biology", *Journal of Biological Systems*, vol. 18, 2010, Spec01, pp. 159-175.
- [11] M. He, S. V. Petoukhov, *Mathematics of bioinformatics: theory, practice, and applications*. USA: John Wiley & Sons, Inc., 2011.
- [12] S. Karlin, F. Ost, B. E. Blaisdell, "Patterns in DNA and amino acid sequences and their statistical significance", in *Mathematical methods for DNA sequences*, M. S. Waterman, Ed. Florida: CRC Press, 1999.
- [13] B. G. Konopel'chenko, U. B. Rumer, "The classification of codons in the genetic code", *Doklady Akademii Nauk of the USSR*, vol. 223, no. 2, 1975, pp. 471-474 (in Russian).
- [14] S. V. Petoukhov, "Hadamard matrices and quint matrices in matrix presentations of molecular genetic systems", *Symmetry: Culture and Science*, vol. 16, no. 3, 2005, pp. 247-266.
- [15] S. V. Petoukhov, *Matrix genetics, algebras of the genetic code, noise-immunity*, Moscow: Regular and Chaotic Dynamics, 2008 (in Russian).
- [16] S. V. Petoukhov, "The genetic code, 8-dimensional hypercomplex numbers and dyadic shifts", February, 2011. Available: <http://arxiv.org/abs/1102.3596>
- [17] S. V. Petoukhov, M. He Symmetrical analysis techniques for genetic systems and bioinformatics: Advanced patterns and applications. Hershey, USA: IGI Global, 2010.
- [18] E. Schrodinger, *What is life? The physical aspect of the living cell*. Cambridge: University Press, 1955.
- [19] J. Seberry, B. J. Wysocki, T. A. Wysocki, "On some applications of Hadamard matrices", *Metrica*, vol. 62, 2005, pp. 221-239.
- [20] J. J. Shu, Y. Li, "Hypercomplex Cross-correlation of DNA Sequences", *Journal of Biological Systems*, vol. 18, no. 4, 2010, pp. 711-725.
- [21] J. J. Shu, L. S. Ouw, "Pairwise alignment of the DNA sequence of the DNA sequence using hypercomplex number representation", *Bull. Math. Biol.*, vol. 66, no. 5, 2004, pp.1423-1438.
- [22] I. Stewart, *Life's other secret: The new mathematics of the living world*. New-York: Penguin, 1999.
- [23] H. Toyoshima, "Computationally efficient implementation of hypercomplex digital filters" in *IEICE Trans. Fundamentals*, H. Toyoshima, Ed. B85-A., Aug 2002, pp. 1870-1876.

ALIGNMENT-FREE PHYLOGENETIC OUTLINE OF A RANDOM-SEQUENCE LIBRARY OF NON-BIOLOGICAL PROTEINS

Miguel A. Jiménez-Montaña¹ and Matthew He²

¹Facultad de Física e Inteligencia Artificial
Universidad Veracruzana, Xalapa, 91000 Veracruz, México
Email: ajimenez@uv.mx

²Division of Math, Science, and Technology
Nova Southeastern University, Ft. Lauderdale, USA
Email: hem@nova.edu

“It seems as though biologists are extraordinarily fond of randomness. A population is defined as one, randomly mating, interbreeding unit, although truly random mating would hardly be practicable in a reasonably large population. Similarly, spontaneous mutations are viewed as randomly sustained base substitutions, in spite of our knowledge of mutational hot spots. I suspect that this extraordinarily strong belief in randomness stems from our too strong faith in the power of natural selection.”

S. Ohno, [24]

Abstract - To assess the degree of randomness and complexity of randomly generated sequences, in an *in vitro* selection experiment by Keefe and Szostack [1], we calculated the Kolmogorov complexity, the algorithmic redundancy, and the Shannon entropy of the sequences. We built an alignment-free phylogenetic tree, employing the algorithmic information distance between each pair of sequences to construct the distance-matrix. The tree represents the history of the set of molecular sequences, and allows us to follow in more detail how chemical function improves with respect to the original sequence. We remark the fact that in directed evolution, the highly predominant changes are between neighboring codons. Thus, the amino acid changes in the protein are not arbitrary, but dictated by the amino acid assignments in the code.

Keywords: Kolmogorov complexity, Shannon entropy of the sequences, algorithmic redundancy, phylogenetic tree, non-biological proteins.

1. Introduction

The frequency of occurrence of functional proteins in collections of randomly generated sequences is an important constraint on models of the evolution of biological proteins. Therefore, the experimental determination of this frequency, by isolating proteins with a specific function from a large random-sequence library of known size, is a relevant endeavor in this field. In an effort to substantiate the hypothesis that primordial functional proteins originated from random sequences, Keefe and Szostak [1] used *in vitro* selection of messenger RNA displayed proteins to sample a large population of distinct randomly generated sequences.

Starting from a library of 6×10^{12} polypeptides, each containing 81 contiguous *randomly chosen* amino acids, they selected functional proteins by enriching for those that bind to ATP. As a result, following eight rounds of selection, they obtained four new ATP-binding protein families, designated A, B, C, D (Fig. 3a of their paper), that appear to be unrelated to each other or to anything found in

the current databases of biological proteins. One of these proteins (Family B) was optimized by directed evolution for improved binding affinity. DNA sequencing of the output from this selection revealed a distant clone (clone 18-19) that differed from the consensus sequence at 15 out of 80 positions, and bound ATP with far greater affinity and specificity than all other clones from that round of selection. From this experiment, Keefe and Szostak [1] estimate that roughly 1 in 10^{11} of all random-sequence proteins have ATP-binding activity.

The X-ray crystal structure of the nucleotide binding domain for protein 18-19 was originally solved by Lo Surdo et al. [2] and found to adopt a novel zinc-nucleated a/b-fold not yet observed by nature. As described in detail in [3], the structural comparison of protein 18-19 with the databank of biological protein folds revealed that the *de novo* evolved protein shared certain structural features with some proteins found in nature. However, unlike many naturally occurring proteins, protein 18-19 requires high concentrations of free ligand in order to remain stably folded and soluble.

In two recent publications, Szostak's group examined the extent to which a *de novo* evolved protein, originally selected on the basis of ligand binding affinity, could be evolved to remain stably folded in the absence of exogenous ligand [3]. These authors designed an *in vitro* selection experiment using mRNA display to isolate variants of protein 18-19 that remained bound to an ATP agarose affinity resin in the presence of increasing concentrations of chemical denaturant. In the second publication [4], they used structural and functional studies to investigate the *in vitro* evolutionary processes in greater detail. We refer the reader to the original papers for further details.

Since proteins acquire functionality (meaning) throughout evolution, to complement the mentioned works, we consider the construction of a phylogenetic tree (Fig. 1) for the evolved proteins in the earliest experiment [1]. The tree represents the history of the set of molecular sequences, and allows us to follow in more detail how chemical function improves with respect to the original sequence. It is commonly believed that to infer such a tree one must first arrange the sequences relative to each other in a way that presents the best available hypothesis of homology at each and every

position in those molecules; i.e., an optimal multiple sequence alignment (MSA). There are nonetheless alternative approaches to molecular phylogenetic inference that do not involve prior MSA (reviewed in [5]). These involve two steps: the calculation of a matrix of pairwise distances among unaligned molecular sequences, followed by generation of a tree using a distance-based method such as neighbor joining [6]. The fundamental difference from alignment-based methods lies mainly in the first step; i.e., how pairwise distances in the underlying distance matrix are defined. The majority of alignment-independent approaches involve information theory and the Kullback-Liebler discrepancy or relative entropy; they are based on the statistical properties of *n*-grams. Or in compression methods, employing the algorithmic information (also called Kolmogorov complexity) shared by two sequences (see Discussion). A notable example of this last approach is the paper by Li et al. [7], who employed the *algorithmic information distance* between a pair of sequences [8,9], to construct a distance-matrix for building a whole mitochondria genome phylogeny without first aligning the sequences. Our approach is closely related to theirs, differing mainly in the software employed to estimate the algorithmic distance.

The simplest way to describe our methodology is in the context of the following linguistically motivated question: Is it possible to identify the subject treated in a text in a way that permits its automatic classification among many other texts in a given corpus? As shown by Benedetto et al. [10] among others, the answer is positive. For DNA sequences, a solution to this kind of problem was delineated by Loewenstern et al. [11] as follows:

"If we took a corpus of DNA sequences, we could gain insight into the degree of similarity between a test sequence and the corpus by compressing the corpus with the test sequence appended, and subtracting the size of this compressed file from the size of the compressed corpus alone. We could classify a test sequence by following the above procedure with two different sample populations of text, assigning the test sequence to the label of the population with which it compressed best"

Here, we follow this idea to classify pseudorandom amino acid sequences.

2. Materials and Methods

Alignment-free Sequence Comparison Algorithms:

In a former publication [12] we introduced the WinGramm Suite [13]. It consists of a set of programs aimed to calculate informational and algorithmic quantities, such as n-gram entropies, context-free grammatical complexity, and algorithmic distance, as well as surrogate statistics, in order to reveal the information content, the complexity or the redundancy embodied in symbol sequences [14, 15, 16, 17, 18].

Here, we have employed the WinGramm Suite to obtain the phylogenetic classification of non-biological amino acid sequences. For this end, we applied our programs to:

- 1) Calculate the context-free grammatical complexity, algorithmic distance and redundancy, Shannon entropy and surrogate statistics of the protein sequences.
- 2) Build a phylogenetic tree to classify these sequences, taken from different clones in the directed evolution experiment.

3. Results

Classification of Pseudorandom Proteins:

Globular proteins have amino acid sequences which are highly complex, indistinguishable from pseudorandom sequences [19]. In that paper the authors estimated the Shannon entropy and applied two compression algorithms (one of them is included in the WinGramm Suite) to estimate the algorithmic complexity of a large, non-redundant, set of protein sequences finding that proteins are fairly close to pseudorandom sequences. They found an entropy reduction due to correlations of about 1 %, corroborated with compression algorithms, which indicates that proteins have approximately 99 % of the complexity of random polypeptides with the same amino acid composition. These results give support to the conclusion of Pande et al. [20], White and Jacobs [21], and others that

protein sequences are “slightly edited random sequences”.

To set up our problem, we consider a sample of 17 sequences from the set generated by Keefe and Szostack [1], in their original *in vitro* selection experiment (appearing in the supplementary information file of the paper). All of the sequences have the following structure:

MDYKDDDDKKT
(Random)₈₁WSASCHHHHHHMGMSG.

From each of these sequences, we dropped the short invariant segments encoding affinity tags for purification, at the beginning and end, retaining the 81 amino acid random segment. The first 13 sequences were obtained from round 8, belonging to families A, B, and C, which have 4 sequences each. The thirteenth sequence constitutes the single representative of family D. The last 4 sequences were acquired from round 18 (Table 1). With the help of the WinGramm Suite [13] we calculated the algorithmic distance between each sequence pair, and obtained the distance matrix (supplementary information Table 2). From this matrix we built the phylogenetic tree (Fig. 1). Comparing this tree with the information in Fig. 3a of [1], we noticed a mistake in their figure: Family A should read Family C and *vice versa*. Professor Szostak acknowledged the misprint (personal communication). The tree displays the right assignment of sequences to families and, correctly, allocates the sequences of generation 18th with family B (see above).

To assess the degree of randomness and the complexity of the experimentally evolved sequences, we calculated the grammatical complexity, the corresponding S-measure (also called Z-score), the algorithmic redundancy and the Shannon entropy of the random segments (Table 1). From the S-measure of the complexity, $S(K)$, defined by the difference between the original value of K and its mean surrogate value, divided by the SD of the standard surrogate values:

$$S = \frac{|K_{orig} - \langle K_{surr} \rangle|}{\sigma_{surr}}$$

it is clear that the evolved sequences are as random as their surrogates. $S_{aver} = 1.6191$ SD. For the families with more than one member (A,

B and C), we concatenated the strings in each group and compared the resulting string with a sequence, of the same length, constructed from concatenated random surrogates. For example, for Family A, we constructed the sequence F_A concatenating the strings in the family: (08-05), (08-07), (08-09), (08-48) (Table 1). We compared the grammatical complexity of F_A , $K(F_A)$, with the complexity of the string S_A , $K(S_A)$, which was constructed from the concatenation of standard-random surrogates of each sequence in the family. Although, both F_A and S_A were built from pseudorandom sequences, the complexity of F_A is much lower than the complexity of S_A because the sub words of F_A are very similar among themselves, and the sub words of S_A are independent pseudorandom sequences. Thus, the complexity of F_A is a good deal lower than the average complexity of its surrogates (Table 1). The sequence F_A can be considered to be the "corpus" of family A. Thus, an unknown sequence may be identified as belonging or not to family A, after compressing it with this "corpus". While the average algorithmic redundancy of the 17 sequences is very low, 1.4 %, the same quantity of the concatenated sequences is high: 42.2 %, 45.6 %, and 44.4 % for F_A , F_B , and F_C , respectively (Table 1). However, the average Shannon entropy (H_{aver}) of the evolved sequences and of the concatenated sequences is almost the same (Table 1). H_{aver} differs from its maximum value, H_{max} , only in 0.18398 bits. This is due to the fact that, contrary to the algorithmic quantities, H depends only on the composition of the sequence, except for finite size effects [22, 23], and not on the order of the symbols.

As we mentioned above, the experiment shows that starting from random amino acid sequences, after a few rounds of Darwinian evolution *in vitro*, it is possible to select a functional protein. Nonetheless, the final protein which carries a biochemical function (in a suitable environment), not only *looks as random as the starting polypeptide* without function,

from which it was generated, but has informational parameters that confirm this fact (Table 1).

4. Discussion and Conclusions

Biological sequences encode information, and the occurrence of evolutionary events separating two sequences sharing a common ancestor will result in the loss of the shared information. Sequences which do not share common ancestor will not share more information than would be expected at random. Therefore, we consider that the appropriate distance matrix was the one defined by the *algorithmic information distance* between a pair of sequences. Because this distance is based on Kolmogorov complexity (estimated by the grammar complexity), that was designed to measure the information content of individual objects. Here, we made a new application of this concept, since concatenating the sequences of a family we measure the information content of the family. Then we compute the shared information between the new sequence and the family.

The further optimization of sequence 18-19 described in [4] consist of twelve single-base mutations, seven of which are transitions. Therefore, the increased stabilization and solubility of the protein is highly influenced by the structure of the genetic code. In the vicinity of a functional protein, in protein space, it is not very difficult to get improvements by fine-tuning it. This is so because, although DNA base mutations are random, each codon does not have the same probability to mutate to any of the other 61 sense codons. In short-term natural evolution and in directed evolution, the highly predominant changes are between neighboring codons. Thus, the amino acid changes in the protein are not arbitrary, but dictated by the amino acid assignments in the code.

Sequence	K	$\langle K \rangle_{sd-surr}$	$\langle K \rangle_{pair-Surr}$	S (K)	R %	H bits
A8-05	81	79.9	81	0.9047	0.74600	4.13920994
A8-07	81	80.1	80.5	1.2853	1.12359	4.14729973
A8-09	81	79.6	80.7	0.7905	1.25000	4.13801925
A8-48	81	79.9	80	0.8162	0.99751	4.11634521
B8-01	78	80	81	2.5276	2.62172	4.1396914
B8-04	78	80.7	80	3.9185	3.22580	4.15270775
B8-08	78	80.3	80	3.5941	2.86426	4.13992263
B8-10	79	80.4	81	2.0314	1.61893	4.14920872
C8-06	80	80	81	0.6031	0.49751	4.13510411
C8-11	80	80.3	81	0.0000	0.0000	4.12701735
C8-17	80	80.6	81	0.3331	0.24937	4.13212155
C8-19	81	80.6	80.5	0.8160	0.49627	4.12905168
D8-20	81	80.6	79	1.0938	0.87172	4.12407239
18-01	78	80.3	81	3.6191	2.98507	4.14924243
18-02	81	80.7	80	0.9047	0.74626	4.17014976
18-03	78	80.4	79	2.2447	2.74313	4.12228268
18-19	79	80.5	81	2.0417	1.37328	4.13367787
Average	79.70	80.28	80.45	1.6191	1.43591	4.13794837
SQR	1.2727	0.3141	0.6643	1.1743	1.01817	0.01276136
F _A	171	296.1	299	31.38664	42.2490	4.13989713
F _B	165	299.5	301.7	28.4291	45.67007	4.1497239
F _C	168	300.8	301	37.8086	44.426	4.13400159

Table 1 Grammatical Complexity, S-values, algorithmic redundancy and entropy for pseudorandom protein sequences^a

^a The labels of the first 17 sequences are the same as in the additional information from Keefe and Szostack [1]. The last three sequences were obtained by concatenating the sequences of the corresponding family, as explained in the text.

In the first column, K is the grammar complexity; the 2nd and 3rd columns are average values of K, for standard and pair-conserving surrogates [12, 13]. S (K) is the S-measure of K, R is the algorithmic redundancy in % and H is the entropy in bits.

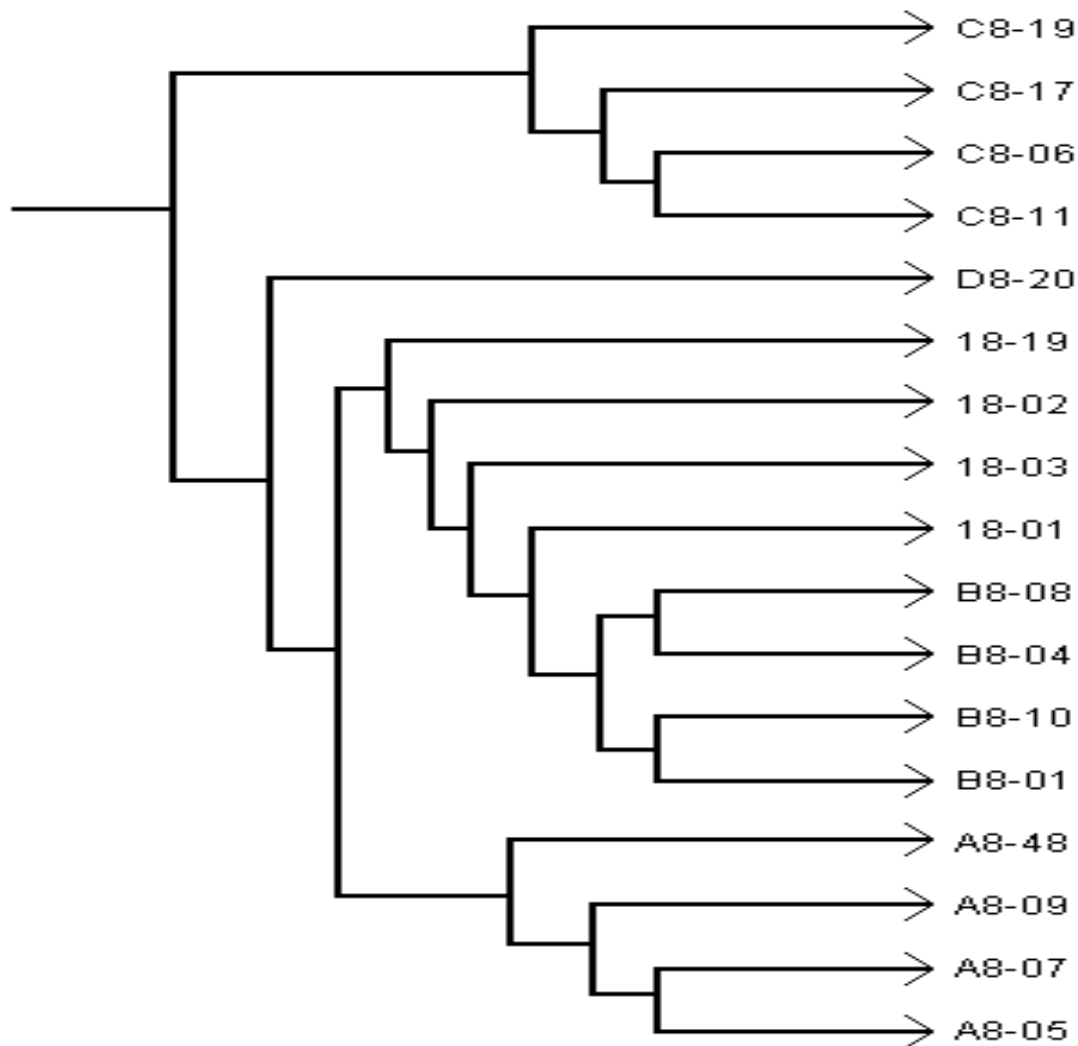


Fig. 1 Phylogenetic tree for the non-biological protein sequences from the experiment performed by Keefe and Szostak (2001).

Acknowledgements

The first author of this paper would like to thank CONACYT, MEXICO Project: 81484; Sistema Nacional de Investigadores; and PROMEP, Project: UV-CA-197, for partial support.

References

1. Keefe, A.D., Szostak, J.W.: Functional proteins from a random-sequence library. *Nature* 410, 715–718 (2001)
2. Lo Surdo, P., Walsh, M.A., Sollazzo, M.: A novel ADP-and zinc-binding fold from function-directed in vitro evolution. *Nat Struct Molec Biol* 11, 382–383 (2004)
3. Smith, M.D., Rosenow, M.A., Wang, M., Allen, J.P., Szostak, J.W., Chaput, J.C.: Structural insights into the evolution of a non-biological protein: Importance of surface residues in protein fold optimization. *PLoS ONE* 2, e467. doi:10.1371/journal.pone.0000467. (2007)
4. Mansy, S.S., Zhang, J.L., Kummerle, R., Nilsson, M., Chou, J.J., Szostak, J.W.,

- Chaput, J.C.: Structure and evolutionary analysis of a non-biological ATP-binding protein. *J. Mol. Biol.* 371, 501–513 (2007)
5. Vinga, S., Almeida, J.: Alignment-free sequence comparison—a review. *Bioinformatics* 19, 513–523 (2003)
 6. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425 (1987)
 7. Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H.: An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17 (2), 149–154 (2001)
 8. Zurek, H.: Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature* 341, 119–124 (1989)
 9. Li, M., Vitányi, P.: *An introduction to Kolmogorov complexity and its applications*. Berlin: Springer. (1997)
 10. Benedetto, D., Caglioti, E., Loreto, V.: Language Trees and Zipping. *Physical Review Letters* 88 (4), 048702 (2002)
 11. Lowenstern, D., Hirsh, H., Yianilos, P., Noordewier, M.: DNA sequence classification using compression-based induction. DIMACS Technical Report 95-04, 1–12 (1995)
 12. Jiménez-Montaña, M.A., Feistel, R., Diez-Martínez, O.: Information Hidden in Signals and Macromolecules I. Symbolic Time-series Analysis. *Nonlinear Dynamics, Psychology & Life Sciences* 8 (4), 445–478 (2004)
 13. Jiménez-Montaña, M. A., Feistel, R.: WinGramm: *Grammatical Complexity Analysis of Sequences*. Internal Report. Faculty of Physics & Artificial Intelligence, University of Veracruz, Mexico (2003). The suite of programs and user manual may be downloaded from: <http://www.io-warnemuende.de/~homepages/rfeistel/>
 14. Gatlin, L.L.: *Information Theory and the Living System*. New York: Columbia University Press (1972)
 15. Ebeling, W., Jiménez-Montaña, M.A.: On grammars, complexity, and information measures of biological macromolecules. *Mathematical Biosciences* 52, 53–71 (1980)
 16. Ebeling, W., Feistel, R.: *Physics of self-organization and evolution* (in German). Berlin: Akademie-Verlag (1982)
 17. Jiménez-Montaña, M.A.: On the syntactic structure of protein sequences and the concept of grammar complexity. *Bull. Math. Biol.* 46(4), 641–659 (1984)
 18. Milosavljevic, A.: Discovering patterns in DNA sequences by the algorithmic significance method. In J.T.L. Wang, B.A. Shapiro, & D. Shasha (Eds.), *Pattern discovery in biomolecular data* (pp. 3–23). Oxford: Oxford University Press (1999)
 19. Weiss, O., Jiménez-Montaña, M.A., Herzel, H.: Information content of protein sequences. *J. Theor. Biol.* 206, 379–386 (2000)
 20. Pande, S.V., Grosberg, A.Y., Tanaka, T.: Nonrandomness in protein sequences: Evidence for a physically driven stage of evolution? *Proc. Natl. Acad. Sci. U.S.A.* 91, 12972–12975 (1994)
 21. White, S.H., Jacobs, R.E.: The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J. Mol. Evol.* 36, 79–95 (1993)
 22. Schmitt, O., Herzel, H., Ebeling, W.: A New Method to Calculate Higher-Order Entropies from Finite Samples. *Europhysics Letters* 23, 303 (1993)
 23. Herzel, H.: Complexity of symbol sequences. *Systems Analysis Modelling Simulation* 5, 435–444 (1988)
 24. Ohno, S.: Modern Coding Sequences Are in the Periodic-to-Chaotic Transition. *Hämatol. Bluttransf.* Vol 32 pp 512–519 (1989)

AMINO ACIDS, EUCLIDEAN DISTANCE AND SYMMETRIC MATRIX

Matthew X. He¹, Miguel A. Jiménez-Montaño², Paolo E. Ricci³

¹Division of Math, Science and Technology
Nova Southeastern University
Ft. Lauderdale, FL 33314, USA
Email: hem@nova.edu

²Facultad de Física e Inteligencia Artificial
Universidad Veracruzana, Xalapa, 91000 Veracruz, México
Email: ajimenez@uv.mx

³Dipartimento Di Matematica
Università di Roma "La Sapienza", Roma 00185, Italia
Email: riccip@uniroma1.it

Abstract: In this paper we introduce the general notion of matrix associated with basic building blocks of protein amino acid and discuss the fundamental properties of these matrices. We further apply general amino acid matrix to a special amino acid Euclidean distance matrix introduced by Graham [1] and study the basic properties of this matrix and provide statistical discription to amino acid distances.

Keywords: Amino acid, Euclidean distance, genetic code, codons, symmmatric matrix.

1. Introduction

It is well known that the genetic code is encoded in combinations of the four nucleotides found in DNA and then RNA. DNA contains the complete genetic information that defines the structure and function of an organism. Proteins are formed using the genetic code of the DNA. Three different processes are responsible for the inheritance of genetic information and for its conversion from one form to another:

- **Replication:** a double stranded nucleic acid is duplicated to give

identical copies. This process perpetuates the genetic information.

- **Transcription:** a DNA segment that constitutes a gene is read and transcribed into a single stranded sequence of RNA. The RNA moves from the nucleus into the cytoplasm.
- **Translation:** the RNA sequence is translated into a sequence of amino acids as the protein is formed. During translation, the ribosome reads three bases (a codon) at a time from the RNA and translates them into one amino acid

These processes are called the Central Dogma of Molecular Biology. The genetic code in messenger ribonucleic acid (mRNA) is composed of A, C, G and U (U for uracil). A mathematical view of genetic code is a map

$$g: C \rightarrow A,$$

where $C = \{(x_1x_2x_3): x_i \in \mathbf{R} = \{A, C, G, U\}\}$ = the set of codons and $\mathbf{A} = \{\text{Ala, Arg, Asp, ..., Val, UAA, UAG, UGA}\}$ = the set of amino acids and termination codons. A

codon is three bases in a DNA or RNA sequence which specify a single amino acid.

One noticeable feature of the genetic code is that some amino acids are encoded by several different but related base codons or triplets. There are 64 triplets or codons. Three triplets (UAA, UAG, and UGA) are stop codons-no amino acids corresponds to their code. The remaining 61 codons represent 20 different amino acids. These genetic code triplets of three bases in mRNA that encode for specific acids during the translation process have some interesting and mathematical logic in their organization. An examination of this logical organization may allow us to better understand the logical assembly of the genetic code and life.

In next section, we introduce the general notion of matrix associated with amino acids and discuss the fundamental properties of these matrices. In section 3, we further apply general amino acid matrix to a special amino acid matrix with Euclidean distances introduced by Graham [1] and study the

basic properties of this matrix and frequency distributions of the amino acid distances.

2. Symmetric Matrix Associated with Amino Acids

The 20 standard amino acids in the genetic code display a much higher structural diversity than the four nucleobases within 64 codons. Although the occurrence of 20 coded amino acids and their contribution to the origin and evolution of the genetic code have been subjected to a wide range of excellent investigations, it has been unclear what principle governs the selection of the 20 amino acids into the genetic code [2, 3]. It was shown in [7] that amino acids distribution within the genetic code is symmetric along the two possible evolutionary axes through the framework of Quasi-28-gon model.

In this section, we arrange the 20 amino acids in a 20x20 square matrix. Abbreviations of the 20 amino acids are represented by the notations summarized in table below.

Table 1. Amino Acid Abbreviations

3-letter notation	1-letter notation
Tyr	Y
His	H
Gln	Q
Arg	R
Thr	T
Asn	N
Lys	K
Asp	D
Glu	E
Gly	G
Phe	F
Leu	L
Ala	A
Ser	S
Pro	P
Ile	I
Met	M
Val	V
Cys	C
Trp	W

Table 2. Amino Acid Matrix

	Y	H	Q	R	T	N	K	D	E	G	F	L	A	S	P	I	M	V	C	W
Y	YY	YH	YQ	YR	YT	YN	YK	YD	YE	YG	YF	YL	YA	YS	YP	YI	YM	YV	YC	YW
H	HY	HH	HQ	HR	HT	HN	HK	HD	HE	HG	HF	HL	HA	HS	HP	HI	HM	HV	HC	HW
Q	QY	QH	QQ	QR	QT	QN	QK	QD	QE	QG	QF	QL	QA	QS	QP	QI	QM	QV	QC	QW
R	RY	RH	RQ	RR	RT	RN	RK	RD	RE	RG	RF	RL	RA	RS	RP	RI	RM	RV	RC	RW
T	TY	TH	TQ	TR	TT	TN	TK	TD	TE	TG	TF	TL	TA	TS	TP	TI	TM	TV	TC	TW
N	NY	NH	NQ	NR	NT	NN	NK	ND	NE	NG	NF	NL	NA	NS	NP	NI	NM	NV	NC	NW
K	KY	KH	KQ	KR	KT	KN	KK	KD	KE	KG	KF	KL	KA	KS	KP	KI	KM	KV	KC	KW
D	DY	DH	DQ	DR	DT	DN	DK	DD	DE	DG	DF	DL	DA	DS	DP	DI	DM	DV	DC	DW
E	EY	EH	EQ	ER	ET	EN	EK	ED	EE	EG	EF	EL	EA	ES	EP	EI	EM	EV	EC	EW
G	GY	GH	GQ	GR	GT	GN	GK	GD	GE	GG	GF	GL	GA	GS	GP	GI	GM	GV	GC	GW
F	FY	FH	FQ	FR	FT	FN	FK	FD	FE	FG	FF	FL	FA	FS	FP	FI	FM	FV	FC	FW
L	LY	LH	LQ	LR	LT	LN	LK	LD	LE	LG	LF	LL	LA	LS	LP	LI	LM	LV	LC	LW
A	AY	AH	AQ	AR	AT	AN	AK	AD	AE	AG	AF	AL	AA	AS	AP	AI	AM	AV	AC	AW
S	SY	SH	SQ	SR	ST	SN	SK	SD	SE	SG	SF	SL	SA	SS	SP	SI	SM	SV	SC	SW
P	PY	PH	PQ	PR	PT	PN	PK	PD	PE	PG	PF	PL	PA	PS	PP	PI	PM	PV	PC	PW
I	IY	IH	IQ	IR	IT	IN	IK	ID	IE	IG	IF	IL	IA	IS	IP	II	IM	IV	IC	IW
M	MY	MH	MQ	MR	MT	MN	MK	MD	ME	MG	MF	ML	MA	MS	MP	MI	MM	MV	MC	MW
V	VY	VH	VQ	VR	VT	VN	VK	VD	VE	VG	VF	VL	VA	VS	VP	VI	VM	VV	VC	VW
C	CY	CH	CQ	CR	CT	CN	CK	CD	CE	CG	CF	CL	CA	CS	CP	CI	CM	CV	CC	CW
W	WY	WH	WQ	WR	WT	WN	WK	WD	WE	WG	WF	WL	WA	WS	WP	WI	WM	VV	WC	WW

It's easy to see that this matrix A is a 20x20 square matrix and A is symmetric since the matrix A is the same as its transpose A^T . The symmetric matrix has a number of properties [5] that we only list a few main results here.

- If A is a square symmetric matrix, then the eigenvalues of A are all real.
- If A is a square symmetric matrix, then the power of matrix A is also symmetric.

3. Amino Acid Distance Matrix

In this section, we consider a square matrix. The entries of this matrix are given by the amino distances. The amino acid distance (physicochemical) was introduced by Granham [1] as follows:

$$D_{ij} = [\alpha (c_i - c_j)^2 + \beta (p_i - p_j)^2 + \gamma (v_i - v_j)^2]^{1/2}$$

where c = composition, p = polarity, and v = molecular volume. In a Euclidean space having these properties as axes, D_{ij} would be the distance between amino acids. The properties are not assumed to be mutually independent; the axes are made orthogonal to facilitate distance calculations. Each property is weighted by dividing the mean distance found with it along in the formula. The constants α , β , γ are squares of the inverses of the D's as indicated in [1]. The similarity between any two amino acid may be measured by this distance. It was observed that if the distance between a pair of amino acids is large, the similarity of the two is small and then the corresponding mutational deterioration will be serious. On the contrary, the small distance for a pair of amino acids suggests a weak deterioration in their mutual mutations. Evidently it leads $D_{ij} = 0$ if $i = j$. All other distances were determined in [1]. We arrange all the distances in a matrix A as follows:

Table 3. Euclidian Distance of Amino Acids

	Y	H	Q	R	T	N	K	D	E	G	F	L	A	S	P	I	M	V	C	W
Y	0	83	99	77	92	143	85	160	122	147	22	36	112	144	110	33	36	55	194	37
H	83	0	24	29	47	68	32	81	40	98	100	99	86	89	77	94	87	84	174	115
Q	99	24	0	43	42	46	53	61	29	87	116	113	91	68	76	109	101	96	154	130
R	77	29	43	0	71	86	26	96	54	125	97	102	112	110	103	97	91	96	180	102
T	92	47	42	71	0	65	78	85	65	59	103	92	58	58	38	89	81	69	149	128
N	143	68	46	86	65	0	94	23	42	80	158	153	111	46	91	149	142	133	139	174
K	85	32	53	26	78	94	0	101	56	127	102	107	106	121	103	102	95	97	202	110
D	160	81	61	96	85	23	101	0	45	94	177	172	126	65	108	168	160	152	154	181
E	122	40	29	54	65	42	56	45	0	98	140	138	107	80	93	134	126	121	170	152
G	147	98	87	125	59	80	127	94	98	0	153	138	60	56	42	135	127	109	159	184
F	22	100	116	97	103	158	102	177	140	153	0	22	113	155	114	21	28	50	205	40
L	36	99	113	102	92	153	107	172	138	138	22	0	96	145	98	5	15	32	198	61
A	112	86	91	112	58	111	106	126	107	60	113	96	0	99	27	94	84	64	195	148
S	144	89	68	110	58	46	121	65	80	56	155	145	99	0	74	142	135	124	112	177
P	110	77	76	103	38	91	103	108	93	42	114	98	27	74	0	95	87	68	169	147
I	33	94	109	97	89	149	102	168	134	135	21	5	94	142	95	0	10	29	198	61
M	36	87	101	91	81	142	95	160	126	127	28	15	84	135	87	10	0	21	196	67
V	55	84	96	96	69	133	97	152	121	109	50	32	64	124	68	29	21	0	192	88
C	194	174	154	180	149	139	202	154	170	159	205	198	195	112	169	198	196	192	0	215
W	37	115	130	102	128	174	110	181	152	184	40	61	148	177	147	61	67	88	215	0

Since D is defined as a Euclidean distance, we have following properties:

- $D(a, a) = 0$, reflective property
- $D(a, b) = D(b, a)$, symmetric property
- $D(a,c) \leq D(a, b) + D(b, c)$, triangle inequality

for any amino acids a, b, and c. The equality may hold true for special amino acids. It's trivial to note that 20 distances hold 0 due to reflective property. Due to symmetric

property, we have 190 distances of 20 amino acids. It's also easy to see that the minimum distance occurs at

$$D(\text{Ile, Leu}) = 5 \text{ and}$$

maximum distance occurs at

$$D(\text{Trp, Cys}) = 215.$$

All other distances are in between 5 and 215. The frequency of distance distribution is illustrated below. It shows that 190 distances are approximately divided into two parts between 5 to 121 and then 121 to 215.

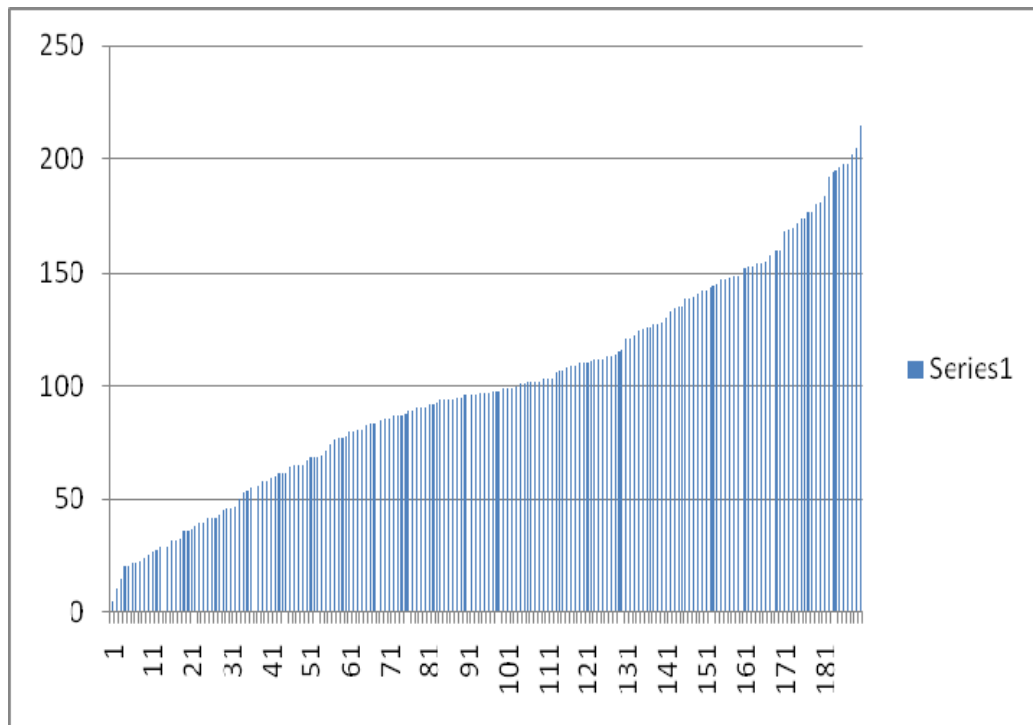


Figure 1. Frequency of Euclidean Distance of Amino Acids

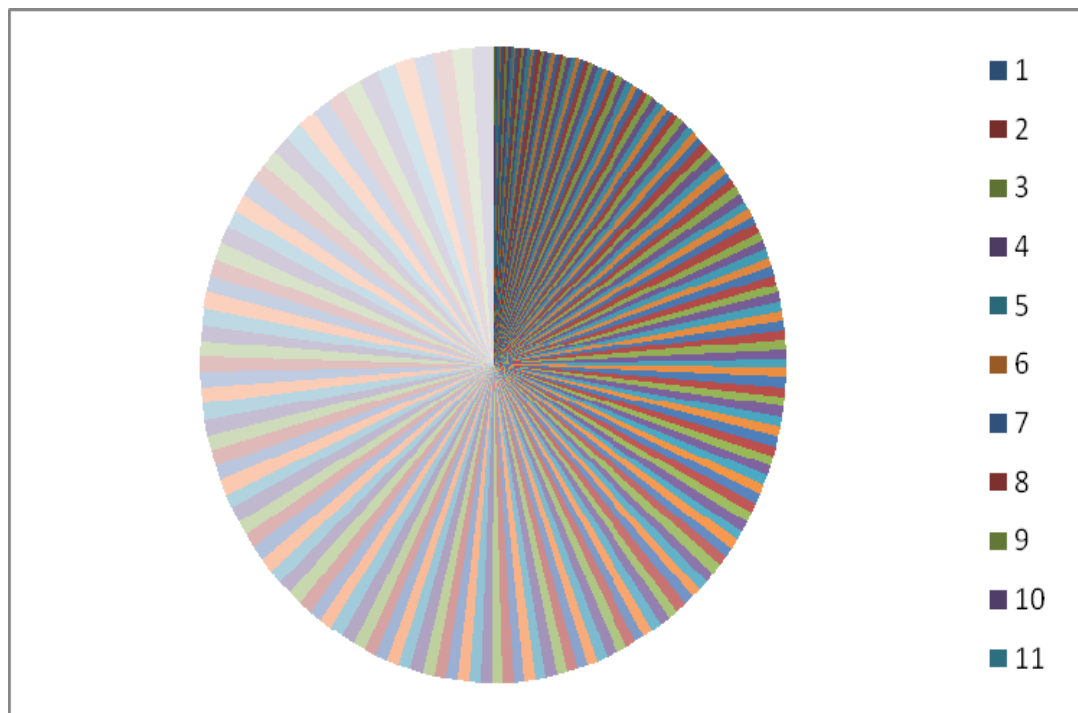


Figure 2. Circular Chart of Euclidean Distance Amino Acids

Furthermore we found that distances of amino acids that the equality of triangle inequality occur at

$$D(S, K) = D(S, H) + D(H, K), \\ (121 = 89 + 32)$$

$$D(S, K) = D(S, Q) + D(Q, K), \\ (121 = 68 + 53)$$

or

$$D(\text{Ser, Lys}) = D(\text{Ser, His}) + D(\text{His, Lys}), \\ (121 = 89 + 32)$$

$$D(\text{Ser, Lys}) = D(\text{Ser, Gln}) + D(\text{Gln, Lys}), \\ (121 = 68 + 53)$$

with an equal sum of 121.

All other distances of three amino acids do not form equality of triangle inequality.

The amino acid distance 121 between **Ser** and **Lys** is located at the centroid position (20/3, 20/3) of lower triangle (0, 0), (20, 0), (0, 20) of amino acid matrix. The centroid of a rigid triangular object is its center of mass: the object can be balanced on its centroid in a uniform gravitational field. The centroid cuts every median in the ratio 2:1, i.e. the distance between a vertex and the centroid is twice the distance between the centroid and the midpoint of the opposite side.

It appears that three amino acids Ser-His-Lys and Ser-Gln-Lys form a pair of interesting tripeptides SHK and SQK.

Our study showed a close relation between amino acid distance and symmetric matrix through a Euclidean distance. It is hoped that these relationships will help us further explore the protein evolution. Life is based on a repertoire of structured and interrelated molecular building blocks that are shared and passed around. The same and related molecular structures and mechanisms show up repeatedly in the genome of a single species and a cross a very wide spectrum of

divergent species. The matrices are storages of digital data. The matrices appear in various dimensions with different shapes. Many literatures on mathematics and biological systems have merged in recent years [5, 6] to further advance our understanding of life and its evolutions. Mathematical rules, physics laws, chemical properties, biological structures and functionalities and environmental impact are the govern bodies of living and nonliving worlds.

Reference

1. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution, *Science*, 1974, 185; 862.
2. Giolio M D., Capobianco M, Medugno M. On the Optimization of the Physicochemical Distances Between Amino Acids in the Evolution of the Genetic Code. *J. Theor. Biol*, 1994, 168: 43.
3. Davydov, O. V. Amino Acid Contribution to the Genetic Code Structure: End-atom chemical rules of doublet composition, *J. Theor Biology*, 1998, 193: 679-690.
4. Bapat, R.B., Raghavan, T.E.S., *Nonnegative Matrices and Applications*, Cambridge University Press, 1997.
5. Percus, J., *Mathematics of Genome Analysis*, Cambridge U. Press, New York, 2002.
6. Pevzner, P. *Computational Molecular Biology*, MIT Press, Cambridge, 2000.
7. Yang, C.M. Chemistry and the 28-Gon Polyhedral Symmetry of the Genetic Code, *J. Of Symmetrion*, Vol. 12, No. 3-4, 2001, 331-347

SESSION

**PROTEIN STRUCTURE PROCESSING +
MACHINE LEARNING + CLASSIFICATION + HPC
+ MODELING + GENOMICS + CLUSTERING +
GENE REGULATORY NETWORKS + HEALTH
INFORMATICS + MICROARRAY +
BIOINFORMATICS + ASSISTIVE TECHNOLOGY +
BIO-INSPIRED SYSTEMS + DATA ANALYSIS**

Chair(s)

Prof. Hamid R. Arabnia

SCOPE: An Open-Source, C++ Implementation for Calculation of Protein Energetics from First Principles

Timothy Matthew Fawcett¹, Stephanie Irausquin¹, Mikhail Simin¹, and Hodayoun Valfar¹

¹Computer Science and Engineering, University of South Carolina, Columbia, SC, USA

Abstract - *SCOPE (Semi Classical Open Source Protein Energy) is an open-source program that has been implemented in the Object-Oriented C++ language, capable of computing non-bonded energies for protein structures from first principles. SCOPE is also capable of manipulating protein structures within the Rotamer space instead of the typical Cartesian space. This approach simplifies calculation of the transitional force field through elimination of unnecessary terms such as bond lengths, bond angles, and other peptide geometrical constraints. Elimination of unnecessary force calculation is beneficial in improving computational performance while the OO approach results in better program maintenance and customization for other projects. Finally, the calculation of forces has been compared and confirmed with respect to other commonly used programs such as CHARMM and Xplor-NIH. Further development of SCOPE can be very beneficial in refinement of computationally modeled structures, or potentially Ab-Initio calculation of structures from first principles without any reliance on homology modeling.*

Keywords: protein structure generation, non-bonded energy, protein folding, protein structure refinement

1 Introduction

Proteins play a critical role in maintaining the homeostatic functions of a biological cell and are often referred to as the working molecules of a cell [1]. Although proteins are prevalently recognized for their enzymatic activities, they are also involved in structural or mechanical functions, as well as regulatory functions [1]. Given that a protein must be folded into its native structure in order to carry out its particular function, it is of no surprise that misfolded proteins are linked with disease [2]. Certain cancers, cystic fibrosis and amyloid diseases such as Alzheimer's, Parkinson's, and Type II Diabetes are such examples [2-5]. Understanding the mechanisms involved in protein folding and protein structure prediction has never been so important. Collaborations between experimental and computational fields have the potential to aid in a number of different applications that will not only accelerate treatments and therapies for a number of diseases, but will also replace the use of costly and time consuming approaches with faster, cheaper computer simulations [6-9].

A protein's structure often dictates its function and therefore investigation of structure of biologically active proteins has intrigued scientists for several decades [7]. Within the last 25 years, the combination of both novel and powerful experimental and theoretical techniques, have contributed to a number of important advances in elucidating protein folding mechanisms; yet there are still challenges that need to be overcome in order to obtain a complete solution [7]. Currently, the "protein folding problem" is often described as 3 different problems: (1) the folding code – what thermodynamic balance of inter-atomic forces dictates protein structure; (2) protein structure prediction – how to predict a protein's native structure given its amino acid sequence; and (3) the folding process - the kinetics associated with how proteins fold quickly [6].

The concept of an energy landscape is fundamental to the mechanism of protein folding [10]. The thermodynamic hypothesis of protein folding states that a protein will fold to a certain form because it is the most favorable [11]. Here an open source software program, SCOPE (Semi Classical Open Source Protein Energy), is presented which allows the user to recreate structures and explore the calculated non-bonded energy potentials associated with those structures using only the initial structure and its dihedral angles as input. Furthermore, due to formulation of protein structure in the rotamer space, several of the traditional force-terms are no longer required. The simplified force field can result in a smoother and more manageable energy landscape.

2 Methods

2.1 Program Details

SCOPE utilizes an object oriented approach and is written in C++. The class structure starts from the fundamental *Atom* class and through compositional inheritance constructs the *AminoAcid*, and finally the *PolyPeptide* objects. The *AminoAcid* class contains an array of *Atoms* to represent an amino acid, while the *PolyPeptide* contains an array of amino acids which constitute a protein. The *AminoAcid* class is a factory class which constructs all 20 amino acids; it contains the attributes of the backbone atoms, as well as the ϕ , ψ , and ω angles, which are part of the REDCRAFT engine [12]. Because backbone atoms are the same for all amino acids except proline, there is only one array of atoms that contains the backbone atoms. The proline amino acid differs from all other amino acids in that it has no

amide hydrogen and its sidechain is linked back to the backbone atoms. Therefore, in the case that a proline amino acid is created, the array of atoms containing backbone atoms for all other amino acids is modified by converting each hydrogen backbone atom into a C δ ; the coordinates of the backbone atoms are then updated accordingly. When the protein is created by the use of ϕ and ψ angles, there is an assumption of a perfect geometry, this translates to perfect bond lengths and favorable bond angles for all atoms of each residue.

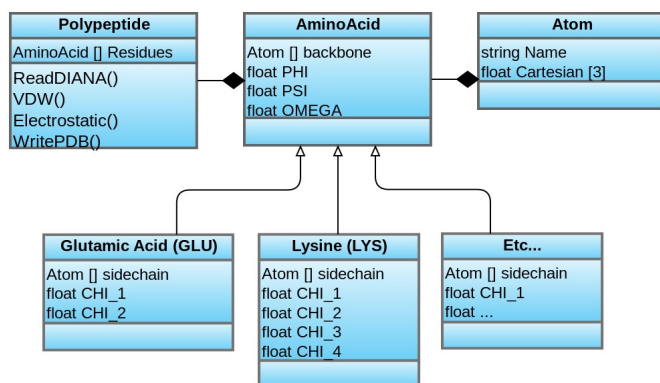


Figure 1: A UML diagram of the class structure for amino acids

The 20 different amino acids inherit the factory amino acid class (Fig. 1). Each class contains the appropriate side chain atoms as well as the χ angles for that particular amino acid. For example, Glycine has a side chain with a single hydrogen atom and no χ angle. Each amino acid class contains the same functions that will rotate their χ angles and update the positions of the side-chain atoms. The side-chain atoms of each amino acid records their own coordinates to a .pdb file.

2.2 Implementation

SCOPE expects two input files from the user. The first input file is a DIANA[13] file (.ang) whose format contains the dihedral angles of each residue; if available, the angles are listed in the following order: ϕ , ω , χ , and finally the ψ angle (Fig. 2). A protein is then generated one amino acid at a time by reading in each residue and rotating its angles so that they correspond to the values of the coinciding DIANA file.

# Structure of 3LAY001 from MOLMOL					
44	LEU	CHI1	-87.498	CHI2	155.773 PSI -41.124
45	THR	PHI	-58.676	CHI1	83.557 PSI 159.867
46	THR	PHI	-60.703	CHI1	-61.376 PSI -62.493
47	GLU	PHI	-57.493	CHI1	47.340 CHI2 -84.354 CHI3 158.889
		PSI	-21.774		
48	GLN	PHI	-77.998	CHI1	-77.310 CHI2 -179.757 CHI3 -19.924
		PSI	-44.705		

Figure 2: Example for a DIANA formatted file

The second input file is a protein structure file (.psf) which contains information related to the topology of the molecule. This topology file provides a rich set of information such as

which 3 atoms make a bond angle, and which 4 atoms make a dihedral angle. Both CHARMM [14] and Xplor-NIH [15] create a .psf file compatible with SCOPE's requirements. SCOPE utilizes the topology information to calculate Van der Waals energy and electrostatic energy of the protein. These energies are then output to the command line along with a .pdb file of the recreated protein.

Because of our previous assumption of perfect geometry during protein construction, SCOPE refrains from calculating the energies associated with bonded terms. As mentioned previously, SCOPE calculates the non-bonded Van der Waals and electrostatic energy terms seen in CHARMM and Xplor-NIH simultaneously. This is accomplished through a series of loops that compares each atom with every other atom. The algorithm begins by comparing the first atom to all other atoms, one at a time and computing a potential energy for each comparison. Similarly, the second atom is compared to all other atoms, except the first atom, one at a time and an energy term is computed for each comparison. These comparisons and energy calculations continue for all remaining atoms so that no duplicate calculations are made, thereby alleviating unnecessary calculations that would needlessly increase computational demands.

The Van der Waals term is used to measure the attraction and repulsion of two atoms. The 12 - 6 Lennard-Jones Potential is used to calculate its value(1). In this equation, σ_{ij} represents the sum of the Van der Waals radii of the two atoms ($\sigma_{ij} = \sigma_i + \sigma_j$); ϵ_{ij} signifies the well depth of the graph calculated as $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$; and r_{ij} denotes the distance between the two atoms(1). The sigma and epsilon values for each atom are the same value in the CHARMM program. The Van der Waals potential can also be calculated using different number of bond exclusions. The default value is the 1-4 atom exclusion, which means 4 atoms with three bonds separating them are excluded from the calculation. A flag can be set when the program is executed to exclude nothing (every atom to atom calculation), 1-2 atom exclusions (2 atoms with a single bond), or 1-3 atom exclusions (3 atoms with 2 bonds are excluded). These exclusions are cumulative so a 1-4 atom exclusions includes the exclusion of 1-2 atoms and 1-3 atoms.

$$LJ = \epsilon_{ij} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} \quad (1)$$

The electrostatic term is used to determine the electrical charge between two atoms. The electrostatic potential is found using Coulomb's Law (2). The charge of each atom is denoted by q_{ij} ; ϵ_0 symbolizes the permittivity of vacuum; just as in the Lennard-Jones equation, the distance between the two atoms is represented by r_{ij} .

$$E = \frac{q_i q_j}{4 \pi \epsilon_0 r_{ij}} \quad (2)$$

Both the Van der Waals and electrostatic energy terms include a distance constraint (r_{ij}). This is to account for instances where the distance between two atoms may be extremely large; in such a case, non-bonded energies are not calculated but set to zero instead.

2.3 Testing Strategy

Initially, SCOPE's ability to generate structures was tested. This was accomplished by creating a peptide of 5 residues in MolMol [16], which is referred to as 5RES, with ϕ and ψ angles rotated to values different from that of MolMol's default ϕ and ψ angle values. The residues comprising 5RES were chosen randomly, with the exception of proline, which was specifically placed in the center of the peptide for its properties discussed previously (section 2.1). Next the two input files required by SCOPE (DIANA and psf file) were created. The DIANA input file was constructed in MolMol and a structure file was created using the CHARMM program. These files were then input into SCOPE. The resulting .pdb file generated by SCOPE was then compared to the original 5RES .pdb created in MolMol by calculating backbone root mean square deviation (RMSD) and also by comparing ϕ and ψ dihedral angles between the two.

Next, SCOPE's ability to construct energetically favorable structures was tested using 12 different proteins (1A1Z, 1DP3, 1TGR, 2J5Y, 1A1W, 3LAY, 1G10, 1J4V, 2EZM, 2EZN, 2MOB, 2PTV) from the Protein Data Bank (PDB) [17]. These particular proteins were selected so that it would be able to test a variety of secondary structures (i.e., alpha-helical, beta-strand, and alpha-beta mix). Both a DIANA file (using MolMol) and a structure file (using CHARMM) were created in order to generate a SCOPE .pdb file for each of the 12 PDB proteins. The resulting SCOPE generated .pdb file was then compared to its original .pdb file for each protein by calculating the backbone RMSD between the two. 1000 similar structures for each protein were created by perturbing or randomly altering the ϕ and ψ angles of the DIANA file; the resulting perturbed structures were all within 6Å of the SCOPE generated protein. Next, the Van der Waals potential was calculated for the SCOPE generated protein as well as the 1000 perturbed structures for each protein (therefore, 1001 structures for each protein) using both SCOPE and CHARMM. Similarly, the electrostatic potential was calculated in both SCOPE and CHARMM.

3 Results

3.1 Structure Generation

In order to test SCOPE's ability to generate structures comparable to other programs, 5RES peptide generated in MolMol was compared to the 5RES peptide generated in SCOPE. The resulting backbone RMSD between the 2

structures is 0.019Å. Comparison of ϕ and ψ angles between the two structures, as well as the peptide sequence, are listed in Table 1. The difference in the angles shown is due to numerical precision error between MolMol and Scope. MolMol will read in the coordinates of the atoms but when displayed within MolMol many of the coordinates have slight differences in the hundredths and thousandths place. Some examples of these numerical precision errors are listed table 1.

Table 1: Peptide of 5 residues(5RES) created MolMol to test the ϕ and ψ angles assigned in MolMol to the ϕ and ψ angles created with SCOPE.

Residue	Original ϕ	Original ψ	SCOPE ϕ	SCOPE ψ
TYR	180	150	180	149.956
GLN	90	60	90.041	59.985
PRO		30		29.926
LYS	-30	0	-29.927	-0.001
ALA	-60	180	-59.954	180

3.2 Protein Model Generation & Non-bonded Energy Evaluations

The previously mentioned 12 proteins were used to test SCOPE's accuracy in representation of protein structures and calculation of potential energies. For each protein, comparisons between the protein obtained from the PDB (original) and the same protein generated by SCOPE (SCOPE) were made by calculating the backbone RMSD between the two. The resulting structural similarity results are shown in Table 2.

Table 2: The different calculations of backbone RMSD between 12 test proteins obtained from the PDB and the same proteins recreated using SCOPE.

Protein	Secondary Structure	Size (Amino Acids)	BB RMSD to DIANA file
1A1Z	α	83	0.633
1DP3	α	55	0.226
1TGR	α	52	0.468
2J5Y	α	61	0.318
1A1W	α	83	0.654
3LAY	α	79	0.631
1G10	α/β	102	0.548
1J4V	β	101	0.441
2EZM	β	101	0.667
2EZN	β	101	0.676
2MOB	α/β	94	0.536
2PTV	β	96	0.354

Each SCOPE generated protein was then perturbed into 1000 structures. The phi and psi angles were rotated by +/- 2 degrees to create 1000 different structures within 6 angstroms. The Van der Waals and electrostatic potential energies were calculated separately in both SCOPE and CHARMM for each of the 1000 derivative structures. Figure

3 displays the correlation for the Van der Waals potential calculated by CHARMM and SCOPE for protein 3LAY and figure 4 reveals the correlation between the electrostatic potential calculated by CHARMM and SCOPE for the same protein (3LAY). Figures 5 – 8 contain the correlation between the Van der Waals potential and the electrostatic potential calculated by CHARMM and SCOPE for proteins 1G10 and 1J4V.

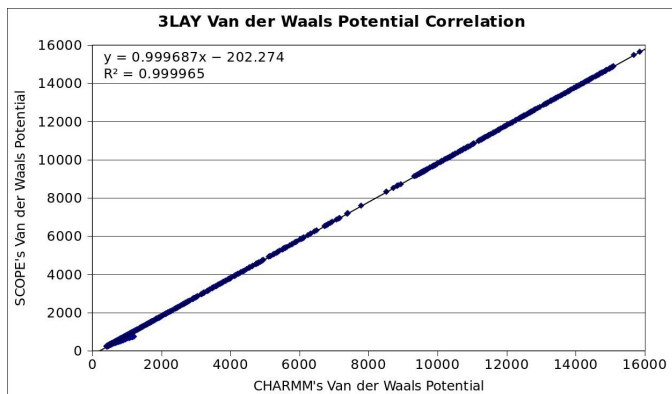


Figure 3: The Van der Waals Correlation between the CHARMM program and SCOPE for protein 3LAY.

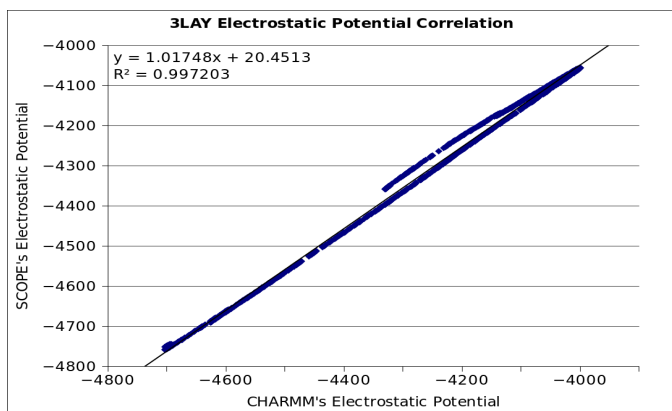


Figure 4: The electrostatic Potential between the CHARMM program and SCOPE for protein 3LAY.

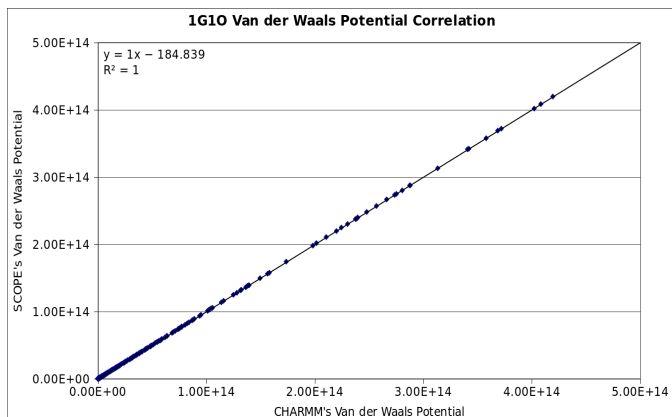


Figure 5: The Van der Waals Potential between the CHARMM program and SCOPE for protein 1G10.

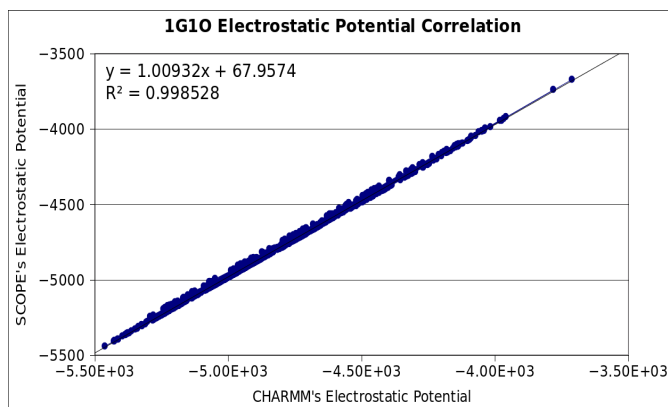


Figure 6: The electrostatic Potential between the CHARMM program and SCOPE for protein 1G10.

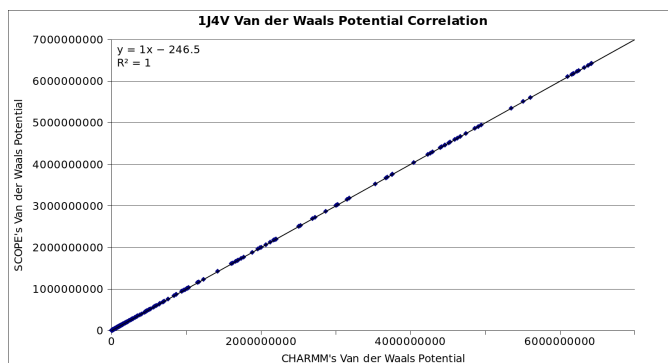


Figure 7: The Van der Waals potential correlation between the CHARMM program and SCOPE for protein 1J4V

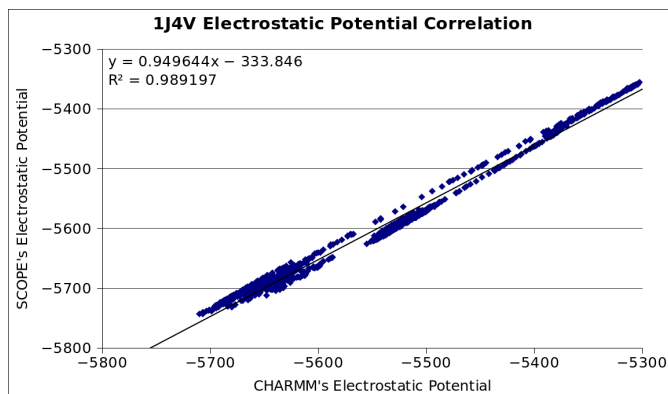


Figure 8: The Electrostatic potential correlation between the CHARMM program and SCOPE for protein 1J4V

4 Discussion

SCOPE's ability to generate structures comparable to those constructed in MolMol is demonstrated using the constructed 5RES peptide and 12 proteins representing different structural categories and sizes. In all of these exercises, the constructed structures by SCOPE are nearly identical to their original counterparts generated by MolMol. The subtle differences that are observed are due to more precise representation of structures by SCOPE. Inherently,

PDB file format imposes a limited numerical precision in representing the atomic coordinates in the Cartesian space. The backbone RMSD between the 5RES peptide generated in MolMol and the 5RES peptide generated in SCOPE is very low (0.019Å).

It is important to note that due to peculiarities of MolMol, the ϕ angles of prolines are not computed and therefore not reported in the DIANA format. Manual editing of the DIANA file is to capture the ϕ angle of prolines. In some instances other violations of standard peptide geometry causes a significant distortion of structures. For example, our preliminary calculations of backbone RMSD between original proteins and that same protein generated with SCOPE (data not shown) revealed problematic values (in excess of 15Å), which is explained using the 3LAY protein as an example. 3LAY contains two prolines at residues 20 and 43 of the protein. After carefully examining these specific residues some interesting observations were made as to why there appeared to be such huge diversions in backbone RMSD. In the case of residue 20, the original structure has an ω angle of -165 degrees, yet SCOPE was not able to rotate the ω angle accordingly. With regard to residue 43, the original structure contains a ϕ angle that is rotated to -53 degrees, yet the corresponding ϕ angle generated in SCOPE defaults to -72.3 degrees. Because MolMol does not calculate the ϕ angles of proline, the Diana file created from MolMol does not write out a ϕ angle for the proline and the angle is not rotated properly. To circumvent the issue, our solution was to add the ϕ angle to the proline in the DIANA file and rotate the ω angles in the original structure to be 180 degrees. These changes allowed for a reduction in the backbone RMSD from 0.855Å to 0.631Å (Fig. 9).

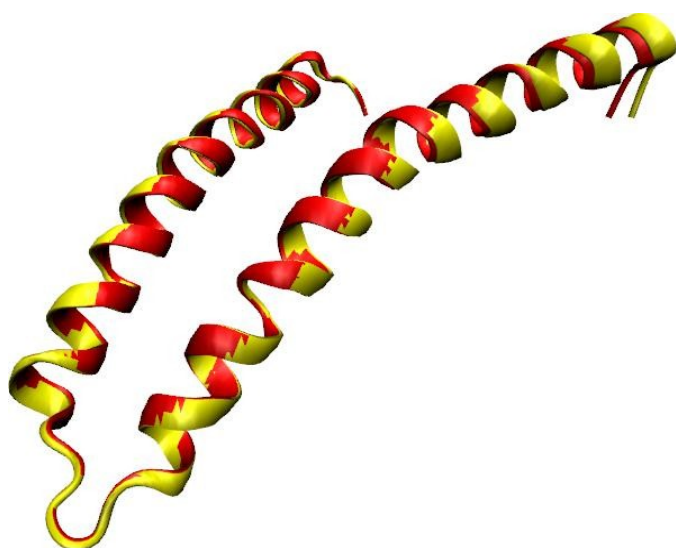


Figure 9: Comparisons between the 3LAY protein from MolMol created by the DIANA file (seen here in red) and the same protein generated by SCOPE using the DIANA file (seen here in yellow) after modifications produced a reduced backbone RMSD (0.631Å).

Protein representation in rotamer space has some distinct advantages. One such advantage is related to the reduced set of information that is needed to reconstruct a protein structure. The backbone only will have dihedral angles ϕ , ψ , and ω . If an all atom version is used then the x, y, z coordinates of 7 atoms need to be known for a total of 35 different parameters. So in the backbone alone, the rotamer representation reduces the number of parameters from 35 to 3.

The use of rotamer space to construct a protein also has some disadvantages, which primarily relate to the loss of information. Bond angles created under the rotamer geometry in both MolMol and SCOPE may differ from the bond angles that are present in the original file obtained from the Protein Data Bank. This was observed when bond angles were calculated between all bond angles for 3LAY, demonstrating bond angles that differed by as much as 65 degrees. Another problem may arise with bond lengths. When bond lengths were compared between all bond lengths for 3LAY, although small, the maximum difference was 0.07Å. The major differences in bond angles and possibly bond lengths led to major differences in atom coordinates, further contributing to high values in backbone RMSD (data not shown).

To alleviate this issue, structures made manually in MolMol (not input as a .pdb file) use the perfect geometry assumption. This allows for structures to be created in MolMol and then compared to structures generated by SCOPE. Using the amino acids of residues 18 – 66 of the 3LAY structure, since a proline is located centrally to the structure, a protein in MolMol was manually created and then the structure was re-created in SCOPE. Comparisons between the two structures revealed a backbone RMSD of 0.290 Å (Fig. 10). Therefore by assuming perfect geometry for bond lengths, bond angles, and the ω angles SCOPE can accurately depict the structure.

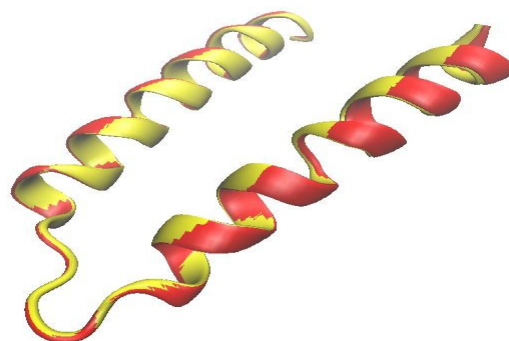


Figure 10: Residues 18 – 66 created by MolMol and SCOPE with an RMSD of 0.290.

Our initial non-bonded energy calculation comparisons between the CHARMM and SCOPE programs revealed large differences between the two. After further investigation it was realized that the coordinates of the atoms in CHARMM and the coordinates of the atoms in SCOPE contained different levels of accuracy. Though the difference in accuracies was only in a few decimal places, the energy spike was magnified since atoms at close distances cause the Van der Waals term to become basically exponential(1). Once the accuracy was fixed to the same number of decimal places, however, the non-bonded energies became extremely correlated.

In fact for all proteins, both non-bonded energies were very strongly correlated between CHARMM and SCOPE with R^2 values ranging from 0.99 to 1.0 and 0.96 to 0.99 for Van der Waals and electrostatic energies respectively (data not shown). Because correlations in non-bonded energies between the CHARMM and SCOPE programs were highly similar, these findings were demonstrated using only proteins 3LAY, 1G10, and 1J4V as examples (Figs. 3 - 8).

Not all of the protein's non-bonded energy terms have a perfect linear correlation. The reason for the discrepancy is that SCOPE does not use an N-terminus residue or a C-terminus residue of the protein while CHARMM, creates the protein with both terminal residues. As a result, the energies from the HT2, HT3, OT1, and OT2 atoms are ignored but the HT1 atom is calculated in CHARMM leaving only the H atom on the first residue to be calculated in SCOPE. Some proteins have the same coordinates for the H atom and the HT1 atom resulting in a higher correlation while the proteins that differed in the coordinates resulted in the lower correlations.

Future work on the SCOPE program will start with adding on to the forcefield. The next term to be added will be a hydrogen-bond term that can be used to help with refinement of protein structures. Also, the addition of a Levenberg-Marquardt minimization algorithm will facilitate refinement of protein structures.

SCOPE is a simple open source program that uses only structure and angle files to reconstruct proteins and output an energy analysis of the newly created structure. Because the program is written in C++, users are given the flexibility to make modifications, such as adding extra energy terms, that are relevant to the task at hand. SCOPE's utility can also be expanded by using it in combination with other protein folding programs, such as REDCRAFT, in order to determine energetically favorable structures.

5 References

- [1] Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J.. *Molecular Cell Biology*, 4th edition. . W.H. Freeman, 2000.
- [2] Dobson, C. M.. Protein misfolding, evolution and disease. *Trends in Biochemical Sciences* (1999) **24**: 329-332.
- [3] Kim J & Holtzman DM. Prion-Like behavior of amyloid-b. *Science* (2010) **330**: pp. 918-919.
- [4] Wu B, Chien EYT, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC, Hamel DJ, Kuhn P, Handel TM, Cherezov V & Stevens RC. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science* (2010) **330**: pp. 1066-1071.
- [5] Powers ET & Balch WE. Protection from the outside. *Nature* (2011) **471**: pp. 42-43.
- [6] Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T.R.. The protein folding problem. *Annu Rev Biophys* (2008) **37**: 289-316.
- [7] Radford, S. E.. Protein folding: progress made and promises ahead.. *Trends in Biochemical Sciences* (2000) **25**: 611-618.
- [8] Andreeva A & Murzin AG. Structural classification of proteins and structural genomics: new insights into protein folding and evolution. *Acta Crystallogr Sect F Struct Biol Cryst Commun* (2010) **66**: pp. 1190-1197.
- [9] Cellmer T, Buscaglia M, Henry ER, Hofrichter J & Eaton WA. Making connections between ultrafast protein folding kinetics and molecular dynamics simulations. *PNAS* (2011) **108**: pp. 6103-6108.
- [10] Dobson, C. M.. Protein folding and misfolding. *Nature* (2003) **426**: 884-890.
- [11] Anfinsen CB. Principles that govern the folding of protein chains. *Science* (1973) **181**: pp. 223-230.
- [12] Bryson M, Tian F, Prestegard JH & Valafar H. REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. *J Magn Reson* (2008) **191**: pp. 322-334.
- [13] Guntert P, Mumenthaler C & Wuthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA.. *J Mol Biol* (1997) **273**: pp. 283-298.
- [14] Brooks, B. R., Brooks III, C. L., Mackerell Jr, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archonits, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., PU, J. Z., Schaefer, M.,

Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. & Karplus, M.. CHARMM: The Biomolecular Simulation Program. *Journal Computational Chemistry* (2009) **30**: 1545-1614.

[15] Schwieters CD, Kuszewski JJ, Tjandra N & Clore G. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* (2003) **160**: pp. 65-73.

[16] Koradi R, Billeter M & Wuthrich K. MOLMOL: A program for display and analysis of macromolecular structures. *J Mol Graphics* (1996) **14**: pp. 51-55.

[17] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE. The Protein Data Bank. *Nucleic Acids Res* (2000) **28**: pp. 235-242.

Comparative Study of Alternative Energy Functions for the HP Model of Protein Structure Prediction

Mario Garza-Fabre, Gregorio Toscano-Pulido and Eduardo Rodriguez-Tello

Information Technology Laboratory, CINVESTAV-Tamaulipas. Parque Científico y Tecnológico TECNOTAM
Km. 5.5 carretera Cd. Victoria-Soto La Marina. Cd. Victoria, Tamaulipas 87130, MÉXICO

Abstract—*Protein structure prediction is the problem of finding the functional conformation of proteins given only their amino acid sequence. The HP model is an abstract formulation of this problem, which captures the fact that hydrophobicity is the major driving force in the protein folding process. It represents a hard combinatorial optimization problem, widely addressed with metaheuristics. The conventional energy function of the HP model does not provide an effective discrimination among candidate solutions. Therefore, alternative energy formulations have been proposed. We inquire into the effectiveness of several of such alternative approaches. The discrimination potential of each of the studied functions is analyzed as well as their impact on the behavior of a basic local search algorithm.*

1. Introduction

Proteins are at the heart of cellular function, carrying most of the key processes associated with life. The functional properties of a protein are dictated by its three-dimensional conformation. To fully understand the biological roles of a protein it is imperative, therefore, to determine its structure.

The *Protein Structure Prediction* (PSP) problem aims to determine the native conformation of proteins given only their linear chain of amino acids. Such a structure is assumed to be the one minimizing the overall free energy [1]. Solving PSP at atomic resolution requires a prohibitive computational effort even for relatively small proteins. Thus, simplified protein models have emerged as valuable tools for studying the most general principles of the folding process.

One of such simplified formulations of PSP is the HP model [2, 3]. However, even a so abstract model represents a hard combinatorial optimization problem which has been proved to be \mathcal{NP} -complete [4, 5]. This has widely motivated the use of metaheuristics to address this problem [6].

Metaheuristics rely on an effective evaluation scheme to guide the search process. However, the conventional energy function of the HP model enables a very poor discrimination. Thus, no preferences can be set among potential solutions, leading the search to be oriented almost at random. This problem is expected to have a major impact on the performance of local search algorithms. The low discrimination

of the conventional function produces large plateaus in the energy landscape, on which local search strategies could fail to detect a promising search direction [7].

Alternative HP energy functions have been proposed to improve the performance of search algorithms [7]–[12]. Nevertheless, there are no reported results on the advantages of using most of such approaches. In this paper, a comparative study is presented where seven different formulations of the HP energy function are considered. The discrimination potential of these approaches is first analyzed. Then, the effectiveness of each of the studied functions to guide the search process is evaluated. A basic local search algorithm was adopted for this sake.

This paper is organized as follows. The HP model is defined in Section 2. In Section 3, the studied approaches are described. Our experimental results are discussed in Section 4. Finally, Section 5 concludes.

2. The HP model

Amino acids, the building blocks of proteins, can be classified on the basis of their affinity for water. *Hydrophilic* or *polar* amino acids (P) are usually found at the outer surface of proteins. By interacting with the aqueous environment, these residues contribute to the solubility of the molecule. In contrast, *hydrophobic* or *nonpolar* residues (H) tend to pack on the inside of proteins, where they interact with one another to form a water-insoluble core. These properties of the amino acids represent, therefore, one of the major driving forces responsible for the folded state of proteins.

In the Hydrophobic-Polar (HP) model, proposed by Dill in 1985 [2, 3], proteins are represented as sequences of the form $S \in \{H, P\}^L$, where L denotes the number of amino acids. The subsets of H and P residues in S are here referred to as S_H and S_P , respectively. Valid conformations are modeled as *Self-Avoiding Walks* of the HP sequence S on a lattice. That is, 1) lattice nodes are labeled by the amino acids, 2) a lattice node can be assigned to at most one residue and 3) adjacent residues in S are also adjacent in the lattice. This study focuses on the two-dimensional square lattice.

By emulating hydrophobic interactions, the HP model aims to find a valid conformation where the number of H - H *topological contacts* (HHtc) is maximized. Two residues

$s_i, s_j \in S$ are said to form a topological contact, denoted by $tc(s_i, s_j)$, if they are nonconsecutive in S (i.e., $|i - j| \geq 2$) but adjacent in the lattice. The free-energy function in the HP model is defined as the negative of HHtc; maximizing HHtc is equivalent to minimize such an energy function.

Formally, PSP in the HP model is defined as the problem of finding the conformation $c^* \in C(S)$ such that $E_{D85}(c^*) = \min\{E_{D85}(c) \mid c \in C(S)\}$, where $C(S)$ is the set of all valid conformations of S . $E_{D85}(c)$ denotes the free energy of conformation c , which is given by:¹

$$E_{D85}(c) = \sum_{s_i, s_j \in S_H} e(s_i, s_j) \quad (1)$$

where

$$e(s_i, s_j) = \begin{cases} -1 & \text{if } tc(s_i, s_j) \\ 0 & \text{otherwise} \end{cases}$$

An example of the optimal conformation for an HP protein of length $L = 20$ on the square lattice is shown in Figure 1.

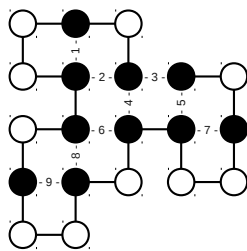


Fig. 1: Optimal conformation for sequence HPHPPHHPH-PPHPHHPH of length $L = 20$. Black and white balls denote H and P residues, respectively. H - H topological contacts (HHtc) have been numbered. The free energy of this conformation is $E_{D85}(c) = -9$, since HHtc = 9.

Despite its apparent simplicity, finding the optimal conformation for a protein in the HP model is a hard combinatorial optimization problem, proved to be \mathcal{NP} -complete [4, 5].

3. Alternative HP energy functions

This section describes the alternative HP energy functions considered for this study. A three-letter acronym has been assigned to each of the studied approaches. The acronyms are based on first author's initial and publication year.

3.1 Krasnogor et al., 1999 (K99)

In the conventional HP energy function, only H - H topological contacts (HHtc) contribute to the quality assessment of conformations. Given two conformations with the same HHtc value, it is possible, however, that one of them has better characteristics (more compact) than the other. Krasnogor

¹The acronym D85 is used to distinguish this conventional function from the other approaches considered in this study.

et al. [7] proposed the following distance-dependent energy function:

$$E_{K99}(c) = \sum_{s_i, s_j \in S_H} e(s_i, s_j) \quad (2)$$

where

$$e(s_i, s_j) = \begin{cases} -1 & \text{if } tc(s_i, s_j) \\ -1/(d(s_i, s_j)^k |S_H|) & \text{otherwise} \end{cases}$$

where $d(s_i, s_j)$ denotes the distance between residues s_i and s_j . In [7], the value of $k = 4$ was used for the square lattice.

In [7], no significant improvements were achieved when using the modified energy function. As the authors pointed out, the superiority of the approach is expected to become more evident for larger instances and, particularly, when local search strategies are implemented. The relevance of using this proposal needs to be further investigated.

3.2 Custódio et al., 2004 (C04)

The conventional HP energy function maximizes only H - H interactions, thus the positioning of P residues is not directly optimized. This may result in unnatural structures for sequences with long P segments and, especially, when P segments are located at the ends of the chain.

Custódio et al. [8] proposed a modified energy function based on the assumption that it may be preferable for an H residue to have a P neighbor than to be in contact with the aqueous solvent. In the proposed function, the energy of a conformation is computed as the weighted sum of the number of hydrophobic-hydrophobic (HHc), hydrophobic-polar (HPc) and hydrophobic-solvent contacts (HSc).² Formally:

$$E_{C04}(c) = \omega_1 HHc + \omega_2 HPc + \omega_3 HSc \quad (3)$$

where ω_1, ω_2 and ω_3 denote the relative importance of HHc , HPc and HSc .

In [8], the proposed function allowed to improve the performance of a genetic algorithm for some of the adopted test cases.

3.3 Lopes and Scapin, 2006 (L06)

Lopes and Scapin [9] proposed an energy function which is based on the concept of *radius of gyration*. The radius of gyration is a measure of the compactness of conformations; the more compact the conformation, the smaller the value for this measure. The proposed function is given by:

$$E_{L06}(c) = HnLB \cdot RadiusH \cdot RadiusP \quad (4)$$

The $HnLB$ term comprises the number of H - H topological contacts (HHtc) and a penalty factor which accounts for the violation of the self-avoiding constraint. Formally:

$$HnLB = HHtc - (NC \cdot PW) \quad (5)$$

²A free lattice location is said to be occupied by the solvent.

where NC is the number of collisions and the penalty weight PW can be computed as $PW = (0.033 \cdot L) + 1.33$ [13].

Before defining the $RadiusH$ and $RadiusP$ terms, let us first define RgH as the radius of gyration for H residues:

$$RgH = \sqrt{\frac{\sum_{s \in S_H} [(x_s - \bar{X})^2 + (y_s - \bar{Y})^2]}{|S_H|}} \quad (6)$$

where x_s and y_s are the coordinates of residue s while \bar{X} and \bar{Y} denote the mean coordinates for H residues. Analogously, we can compute RgP , the radius of gyration for P residues, by considering only P rather than H residues in (6).

The $RadiusH$ term measures how compact the hydrophobic core of the conformation is. This term is given by:

$$RadiusH = MaxRgH - RgH \quad (7)$$

where $MaxRgH$ is the radius of gyration of a totally unfolded conformation; *i.e.*, the maximum possible RgH value.

Finally, the $RadiusP$ term aims to push P residues away from the hydrophobic core. Given the previously defined RgH and RgP measures, the $RadiusP$ term is computed as:

$$RadiusP = \begin{cases} 1 & \text{if } (RgP - RgH) \geq 0 \\ \frac{1}{1 - (RgP - RgH)} & \text{otherwise} \end{cases} \quad (8)$$

$RadiusP$ lies in the range $[0, 1]$. A value of $(RgP - RgH) > 0$ means that P residues are more exposed than H residues. This is a convenient scenario, so the $RadiusP$ term has no contribution to the final energy value ($RadiusP = 1$). Otherwise, $(RgP - RgH) < 0$ suggests H residues to be more spread than the P ones, so the energy value of the conformation is decreased. Note that (4) is to be maximized.

In [9, 13], no results are provided on the impact of using this function rather than the conventional approach.

3.4 Berenboym and Avigal, 2008 (B08)

Berenboym and Avigal [10] proposed an alternative energy function, called by them the *global energy*. In this function, each pair of nonconsecutive H residues contributes to the energy value, even if they are not topological neighbors:

$$E_{B08}(c) = \sum_{s_i, s_j \in S_H} e(s_i, s_j) \quad (9)$$

where

$$e(s_i, s_j) = \begin{cases} \frac{-1}{(x_{s_i} - x_{s_j})^2 + (y_{s_i} - y_{s_j})^2} & \text{if } |i - j| \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

In [10], the effects of using a local search operator within a genetic algorithm were investigated for both, the conventional and the proposed energy functions. However, an explicit comparison to demonstrate the advantages of using a particular energy function was not reported.

3.5 Cébrían et al., 2008 (C08)

Cébrían *et al.* [11] proposed an alternative formulation of the HP energy function which measures the deviation from the unit distance (*i.e.*, topological contact distance) for each pair of H residues. Let $d(s_i, s_j)^2 = (x_{s_i} - x_{s_j})^2 + (y_{s_i} - y_{s_j})^2$ be the distance between residues s_i and s_j , and let $dv(s_i, s_j) = d(s_i, s_j)^2 - 1$ denote its deviation from the unit distance. The energy value of a conformation c is given by:

$$E_{C08}(c) = \sum_{s_i, s_j \in S_H} dv(s_i, s_j)^k \quad (10)$$

where $k \geq 1$ is a parameter of the function, whose larger values give more weight to unit distances. We used $k = 2$, since this value seems to provide the best behavior based on the results reported in [11]. $E_{C08}(c^*) = 0$ would refer to the ideal (potentially unrealistic) scenario where all pairs of H residues are at a unit distance in conformation c^* . In [11], no experimental results were reported about the benefits of using the proposed energy function instead of the conventional one.

3.6 Islam and Chetty, 2009 (I09)

Islam and Chetty [12] proposed a modified HP function based on two measures: *H-compliance* and *P-compliance*.

H-compliance measures the proximity of H residues to the center of a hypothetical rectangle enclosing all H residues, denoted by the reference point (x_r, y_r) . Formally:

$$H\text{-comp}(c) = \frac{\sum_{s \in S_H} (x_r - x_s)^2 + (y_r - y_s)^2}{|S_H|} \quad (11)$$

where x_s and y_s denote the lattice coordinates of the s residue.

P-compliance is a measure of how close P residues are to the boundaries of a hypothetical rectangle enclosing all P residues, defined by x_{min} , x_{max} , y_{min} and y_{max} . Formally:

$$P\text{-comp}(c) = \frac{\sum_{s \in S_P} \min \left\{ \begin{array}{l} |x_{min} - x_s|, |x_{max} - x_s|, \\ |y_{min} - y_s|, |y_{max} - y_s| \end{array} \right\}}{|S_P|} \quad (12)$$

Finally, the energy of a given conformation c is defined as:

$$E_{I09}(c) = \alpha E_{D85}(c) + H\text{-comp}(c) + P\text{-comp}(c) \quad (13)$$

where $E_{D85}(c)$ is the conventional HP energy function (see Section 2) and α is a high value integer constant to ensure this will be the dominant term in (13). We used $\alpha = 10,000$.

In [12], the advantages of using the proposed energy function were demonstrated for a 85-length HP benchmark sequence. However, the impact of using this function should be carefully investigated for a larger set of test cases.

4. Experimental Results

In this section, we investigate the effectiveness of the studied approaches. Note, however, that even when an alternative evaluation function is used, the goal of the optimization process remains to maximize HHtc, which is the singular objective in the HP model. In this study, the exclusive purpose for using alternative energy functions is to guide the search process in a more effective manner. Table 1 presents the 9 HP benchmark sequences adopted for this study.

Table 1: Benchmarks, length (L) and optimal value (HHtc*).

	Sequence	L	HHtc*
S1	HPHP ₂ H ₂ PH ₂ PH ₂ P ₂ HPH	20	9
S2	P ₂ HP ₂ H ₂ P ₄ H ₂ P ₄ H ₂ P ₄ H ₂	25	8
S3	P ₃ H ₂ P ₂ H ₂ P ₅ H ₇ P ₂ H ₂ P ₄ H ₂ P ₂ HP ₂	36	14
S4	P ₂ HP ₂ H ₂ P ₂ H ₂ P ₅ H ₁₀ P ₆ H ₂ P ₂ H ₂ HP ₂ H ₅	48	23
S5	H ₂ (PH) ₄ H ₃ P(HP) ₃ (P ₃ H) ₃ PH ₄ (PH) ₄ H	50	21
S6	P ₂ H ₃ PH ₈ P ₃ H ₁₀ PH ₃ H ₁₂ P ₄ H ₆ PH ₂ PH ₂	60	36
S7	H ₁₂ PHPH(P ₂ H ₂ P ₂ H ₂ P ₂ H) ₃ PHPH ₁₂	64	42
S8	H ₄ P ₄ H ₁₂ P ₆ (H ₁₂ P ₃) ₃ HP ₂ H ₂ P ₂ H ₂ HPH	85	53
S9	P ₆ HPH ₂ P ₅ H ₃ PH ₅ PH ₂ P ₄ H ₂ P ₂ H ₂ PH ₅ P	100	48
	H ₁₀ PH ₂ PH ₇ P ₁₁ H ₇ P ₂ HPH ₃ P ₆ HPH ₂		

4.1 Degree of discrimination

The discrimination strategy directly impacts the performance of search algorithms. That is, if it is not possible to set preferences among solutions the search process will be guided practically at random.

The degree of discrimination that each of the studied functions provides is investigated. We analyzed the distribution of ranks that these approaches induce on a set of candidate solutions. A ranking expresses the relationship among a set of items according to a given property. In the context of this study, potential conformations are ranked according to their quality. The first rank is assigned to the best solution, the next rank to the second best solution, and so on. Solutions with the same quality will share the same rank.

We adopted the *relative entropy* (RE) measure proposed by Corne and Knowles [14]. Given a set of n ranked solutions (there are at most n ranks, and at least 1), the relative entropy of the distribution of ranks D is defined as:

$$RE(D) = \frac{\sum_r \frac{D(r)}{n} \log\left(\frac{D(r)}{n}\right)}{\log(1/n)} \quad (14)$$

where $D(r)$ denotes the number of solutions with rank r . $RE(D)$ tends to 1 as approaching to the ideal situation where each solution has a different rank (*i.e.*, the maximum possible discrimination). On the other hand, when all the solutions share the same ranking position (*i.e.*, the poorest discrimination), $RE(D)$ takes a value of zero.

In this experiment, 1,000 different valid structures were generated at random. For each of the studied energy functions, these solutions were evaluated and ranked to finally compute the RE measure. We performed 100 repetitions of this experiment for all the benchmarks. The box plots in Figure 2 present the overall statistics of this experiment.

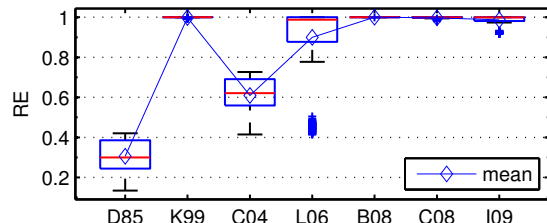


Fig. 2: Relative entropy (RE). Overall statistics.

From Figure 2, it is possible to note that the conventional HP function, D85, achieved the lowest RE values. This confirms the poor discrimination capabilities of this function, which has been the main factor motivating the exploration of alternative approaches. C04 showed the worst performance among the alternative functions. Function L06 achieved high RE values most of the time, but the outliers indicate a low performance of this function for some of the benchmarks. Finally, it is important to remark the high discrimination provided by functions B08, K99, C08 and I09.

The above results can be better understood by analyzing Figure 3. This figure presents the histograms with the distribution of ranks achieved by each function for the first repetition of this experiment regarding sequence S1. From this figure, it is possible to note how poor the distribution of ranks achieved by function D85 is. Only five different ranking positions were enough to classify the 1,000 generated solutions. It can be seen a peak where there are almost 400 solutions sharing the same rank. In fact, no matter the amount of generated solutions, the maximum number of ranks which can be assigned through function D85 is 9, since $HHtc^* = 9$ for this benchmark sequence (S1). The second worst scenario is presented by function C04, where less than 40 different ranking positions were required, out of which two were each assigned to at around 100 conformations.

Functions L06 and I09 showed an increased discrimination, since about 720 and 650 ranking positions were occupied to classify the totality of solutions, respectively. In the case of function I09, a maximum of eight solutions were assigned to the same rank. On the other hand, the histogram for L06 presents a high peak indicating that there are about 250 equally ranked conformations. Function L06 is defined as the product of three terms, out of which one corresponds to HHtc (see Section 3.3). All solutions for which $HHtc = 0$ will have the same energy value, 0. To some extent, this can be seen as a drawback. Function L06 will not be able to discriminate among these solutions even if some of them have better chances than others to further improve.

Finally, the histograms for B08, K99 and C08 confirm the high degree of discrimination these approaches provide. We can see that function C08 allowed roughly 950 different ranks to be assigned. B08 showed the strongest discrimination among all the studied functions, followed by K99. The corresponding histograms for these functions reveal that almost all solutions were mapped to a different rank. Only a few ranks were assigned to at most two solutions.

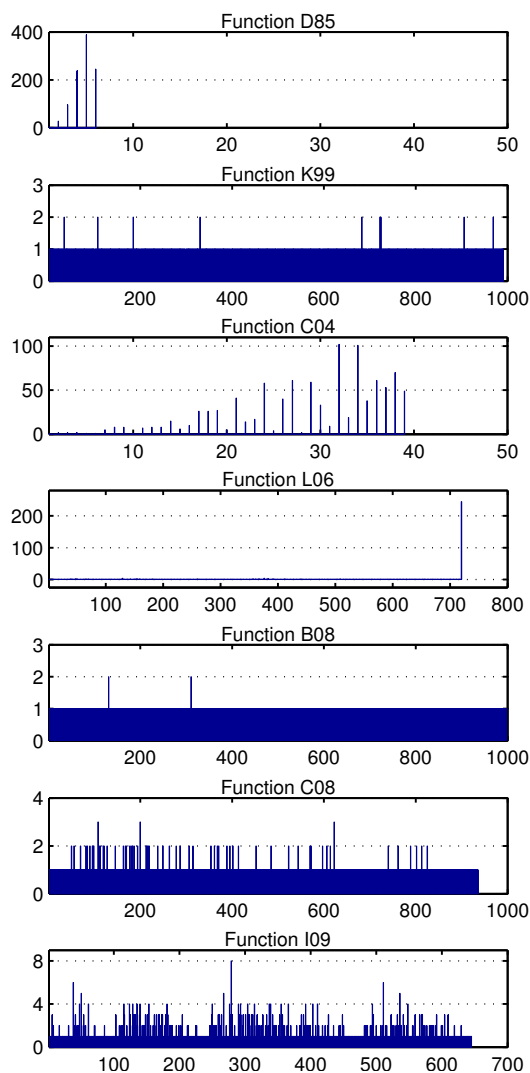


Fig. 3: Density of the distribution of ranks achieved by the studied evaluation functions. Sequence S1, run 1.

4.2 Search performance

We implemented a *Steepest Descent Hill Climbing* algorithm (SDHC) to evaluate the effectiveness of the studied energy functions at guiding the search process. SDHC is a parameter-free algorithm, whose motivation in this study is to avoid affecting (neither negatively nor positively) the performance of the approaches through parameter settings. Given that SDHC is a local search technique, functions providing a finer discrimination are expected to perform better. As pointed out by Krasnogor *et al.* [7], a poor discrimination will produce large plateaus in the energy landscape, on which local search strategies could fail to identify a descent direction. Algorithm 1 describes the implemented SDHC.

The algorithm starts with a valid conformation generated at random, denoted by c . Once c is generated, we identify c' , the best conformation among all defined in the neighborhood of c , $N(c)$. Then, solutions c and c' are compared with respect to their energy values. At this point is where the

Algorithm 1 Steepest Descent Hill Climbing (SDHC).

```

BEGIN SDHC()
1:  $c \leftarrow \text{getRandomValidSolution}()$ 
2: loop
3:    $c' \leftarrow \text{getBest}(N(c))$ 
4:   if  $E(c') < E(c)$  then
5:      $c \leftarrow c'$ 
6:   else
7:     Stop()

```

END

different energy functions come to play a decisive role in the behavior of the algorithm. If c' has a better energy value than c ($E(c') < E(c)$), then a replacement occurs and the process repeats. Otherwise, the process ends, since given the current solution and the adopted neighborhood it is not possible to achieve an improvement (c is locally optimal).

An internal coordinates representation with absolute moves was adopted [15]. Candidate conformations are encoded as sequences in $\{U, D, L, R\}^{L-1}$, denoting the up, down, left and right possible locations for a residue with regard to the preceding one (solutions are decoded to Cartesian coordinates for evaluation). The implemented neighborhood structure $N(c)$ is defined by all solutions that can be reached through 1-variable perturbations of c . Given a sequence of length L , the size of such a neighborhood is $|N(c)| = 3(L-1)$. However, only valid conformations are considered.

It is important to remark that the aim of using the SDHC algorithm is not to improve the state-of-the-art results for this problem. In this study, SDHC serves only as a tool to measure the impact of using each of the energy functions.

The behavior of the SDHC algorithm was evaluated when using each of the studied functions. A total of 100 independent executions were performed for all the adopted benchmarks. The results of this experiment are presented in Figure 4. Each plot in this figure shows the average number of H - H topological contacts (HHtc) achieved by the algorithm as the search progressed (iteration by iteration).

From Figure 4, it is possible to derive some general conclusions. The poorest performance for this experiment was presented by function C08, whose results were even worse than those of function D85 in most of the considered test cases. This behavior can be explained by the fact that function C08 is not consistent with the conventional objective of the HP model. As stated at the beginning of Section 4, even when alternative functions are used to guide the search process, the goal remains to maximize HHtc; or, which is equivalent, to minimize function D85. The alternative function should not contradict D85 when discriminating among potential conformations, otherwise we will probably be pursuing a different optimum. Nevertheless, given two conformations c_1 and c_2 , it is possible the case where $E_{D85}(c_1) < E_{D85}(c_2)$ but $E_{C08}(c_1) > E_{C08}(c_2)$,

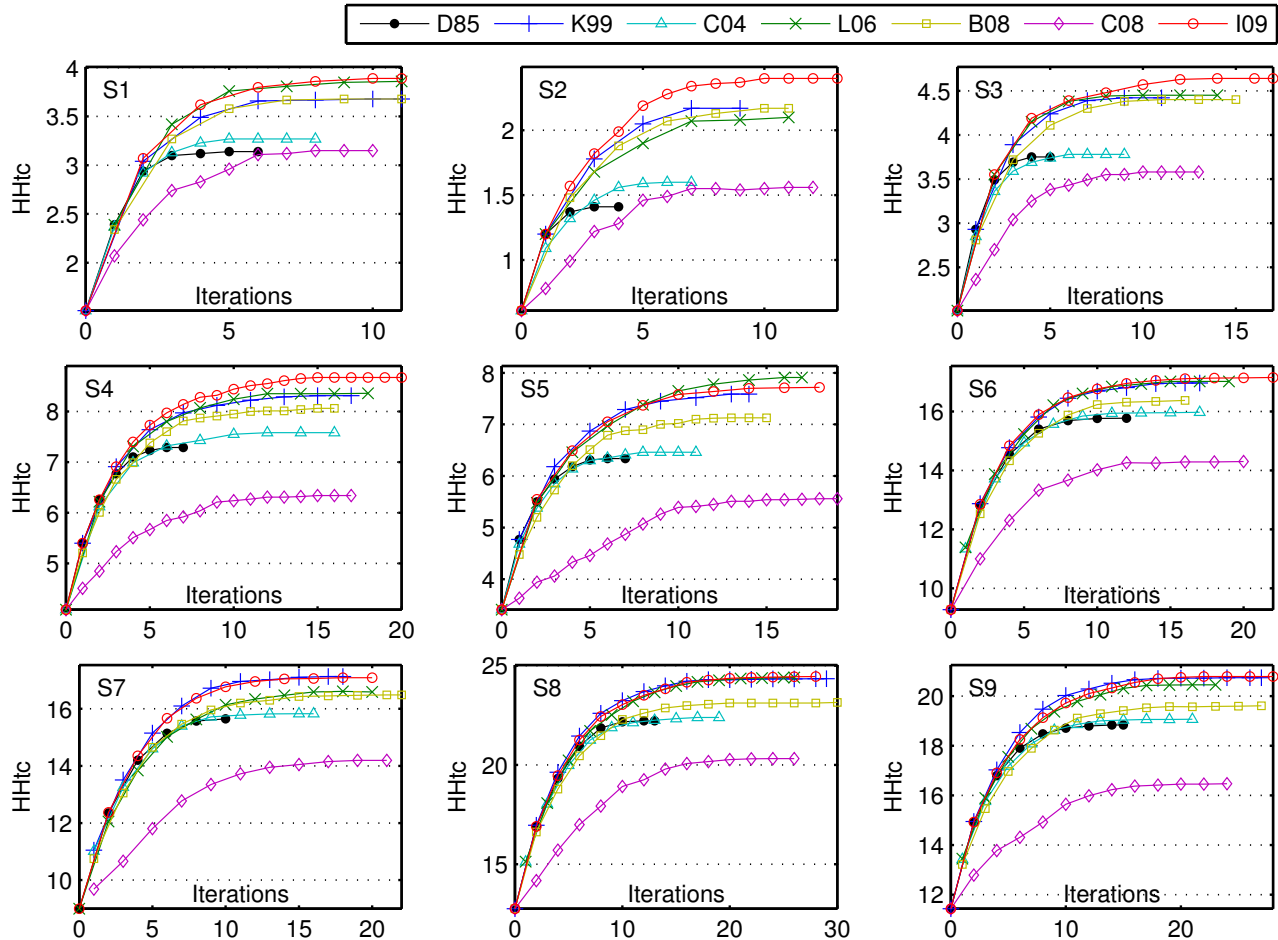


Fig. 4: Results of the SDHC algorithm. Achieved number of H - H topological contacts (HHtc) at each iteration. Average of 100 independent executions.

which is a contradiction.³ An example of this scenario is presented in Figure 5. This can be seen as a drawback, so function C08 is not expected to steer the search in an effective manner. Such an important issue needs to be further explored for all the studied approaches.

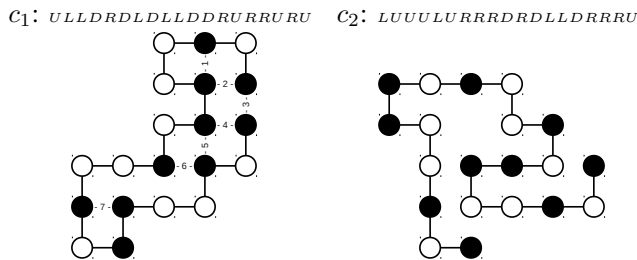


Fig. 5: C08 contradicts D85, since $E_{D85}(c_1) = -7 < E_{D85}(c_2) = 0$ but $E_{C08}(c_1) = 5548 > E_{C08}(c_2) = 5308$.

As expected, function D85 showed a low performance for

³Note that the case where $E_{D85}(c_1) = E_{D85}(c_2)$ but $E(c_1) \neq E(c_2)$ is not a contradiction. This is a convenient scenario, since the aim of using the alternative function E is to enable a more fine-grained discrimination.

this experiment. For all instances, the algorithm achieved the lowest number of iterations due to the poor discrimination this function provides. Function D85 exposed the second worst overall behavior. C04 reached slight improvements, but its limited performance was comparable with that of function D85 in some cases. Note that functions D85 and C04 were previously identified in Section 4.1 because of their low discrimination capabilities. To some extent, this explains the poor performance presented by these approaches.

Functions K99 and B08 behaved similarly for the smallest benchmarks, but their performance curves diverged as the size of the problem was increased. The results of B08 deteriorated for the largest test cases, while the increasing performance of K99 allowed this function to compete at the top of the ranking. L06 obtained very competitive results most of the time. Finally, we can highlight the outstanding behavior that function I09 consistently showed for all the considered test cases. Our results indicate that the best performers were I09, L06 and K99, in this order.

Functions I09, K99, B08 and C08 were all identified in Section 4.1 to provide a strong discrimination. However, only K99 and I09 are among the best performers of

this experiment. That some equally discriminative functions performed better than others suggests that more important than the strength is the effectiveness of the discrimination (intensity does not imply effectiveness).

5. Conclusions and Future Work

The conventional energy function of the HP model enables a very poor discrimination among potential conformations. Nevertheless, an effective evaluation scheme is an essential requirement for metaheuristics in order to guide the search process towards promising regions of the solutions space. Alternative HP energy functions have been proposed to enhance the performance of search algorithms. However, for most of these approaches there are not reported experimental results where the benefits of their usage are demonstrated.

This paper presented the results of a comparative study where seven different formulations of the HP energy function were considered. Our first experiment was concerned with the analysis of the degree of discrimination that each of these functions provides. The obtained results confirmed the poor discrimination capabilities of the conventional function, which has been the main motivation for exploring alternative approaches. All the alternative functions demonstrated to provide a more fine-grained discrimination. The most discriminative function according to our results is B08, followed by the K99, C08 and I09 approaches, in this order.

In our second experiment, we evaluated the impact of using the studied functions on the performance of a parameter-free local optimizer. The aim of using a parameter-free algorithm was to avoid influencing the behavior of the approaches through parameter settings. In general, most of the alternative functions allowed to increase the performance of the implemented algorithm. As expected, the conventional D85 function exhibited a low performance for this experiment. However, the C08 approach behaved even worse for most of the adopted test cases. On the other hand, functions I09, L06 and K99 consistently achieved very competitive results, being the best performers in this test.

From this study, it is possible to derive some general conclusions. First, intensity of discrimination does not necessarily imply effectiveness at guiding the search process. Even when functions I09, K99, B08 and C08 were all identified to provide a strong discrimination, only I09 and K99 behaved favorably. In contrast, B08 and particularly C08 presented a limited search performance. That the less discriminative approaches (D85 and C04) showed a low overall performance confirmed, however, that a tighter evaluation scheme is important to improve the behavior of search algorithms.

The fact that D85 consistently exposed a poor performance supports the relevance of exploring the use of alternative approaches. To the best of our knowledge, this research is producing the first results that have been reported in this direction. Nevertheless, this research is in progress. The preliminary results presented in this paper suggest that

functions I09, L06 and K99 are very promising approaches for studies on the HP model. However, the impact of using these approaches needs to be further investigated for more sophisticated search algorithms. Also, it is important to extend this study to the three-dimensional cubic lattice, or to other lattice configurations (for example, the face-centered cubic lattice), in order to generalize our conclusions.

References

- [1] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.
- [2] K. A. Dill, "Theory for the Folding and Stability of Globular Proteins," *Biochemistry*, vol. 24, no. 6, pp. 1501–9, 1985.
- [3] K. F. Lau and K. A. Dill, "A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins," *Macromolecules*, vol. 22, no. 10, pp. 3986–3997, 1989.
- [4] B. Berger and T. Leighton, "Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-complete," in *RECOMB '98: Proceedings of the second annual international conference on Computational molecular biology*. New York, NY, USA: ACM, 1998, pp. 30–39.
- [5] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, "On the Complexity of Protein Folding," in *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, 1998, pp. 597–603.
- [6] X. Zhao, "Advances on Protein Folding Simulations Based on the Lattice HP models with Natural Computing," *Appl. Soft Comput.*, vol. 8, no. 2, pp. 1029–1040, 2008.
- [7] N. Krasnogor, W. E. Hart, J. Smith, and D. A. Pelta, "Protein Structure Prediction With Evolutionary Algorithms," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 1999)*, W. Banzhaf, J. M. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. J. Jakiela, and R. E. Smith, Eds. Orlando, Florida, USA: Morgan Kaufman, July 1999.
- [8] F. L. Custódio, H. J. C. Barbosa, and L. E. Dardenne, "Investigation of the Three-dimensional Lattice HP protein Folding Model Using a Genetic Algorithm," *Genetics and Molecular Biology*, vol. 27, pp. 611–615, 2004.
- [9] H. Lopes and M. Scapin, "An Enhanced Genetic Algorithm for Protein Structure Prediction Using the 2D Hydrophobic-Polar Model," in *Artificial Evolution*, ser. Lecture Notes in Computer Science, E.-G. Talbi, P. Liardet, P. Collet, E. Lutton, and M. Schoenauer, Eds. Springer Berlin / Heidelberg, 2006, vol. 3871, pp. 238–246.
- [10] I. Berenboym and M. Avigal, "Genetic Algorithms with Local Search Optimization for Protein Structure Prediction Problem," in *GECCO '08: Proceedings of the 10th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2008, pp. 1097–1098.
- [11] M. Cebrián, I. Dotú, P. Van Hentenryck, and P. Clote, "Protein Structure Prediction on the Face Centered Cubic Lattice by Local Search," in *Proceedings of the 23rd national conference on Artificial intelligence - Volume 1*. AAAI Press, 2008, pp. 241–246.
- [12] K. Islam and M. Chetty, "Novel Memetic Algorithm for Protein Structure Prediction," in *AI 2009: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, A. Nicholson and X. Li, Eds. Springer Berlin / Heidelberg, 2009, vol. 5866, pp. 412–421.
- [13] H. Lopes and M. Scapin, "A Hybrid Genetic Algorithm for the Protein Folding Problem Using the 2D-HP Lattice Model," in *Success in Evolutionary Computation*, ser. Studies in Computational Intelligence, A. Yang, Y. Shan, and L. Bui, Eds. Springer Berlin / Heidelberg, 2008, vol. 92, pp. 121–140.
- [14] D. Corne and J. Knowles, "Techniques for Highly Multiobjective Optimisation: Some Nondominated Points are Better than Others," in *2007 Genetic and Evolutionary Computation Conference (GECCO'2007)*, D. Thierens, Ed., vol. 1. London, UK: ACM Press, July 2007, pp. 773–780.
- [15] R. Unger and J. Moult, "Genetic Algorithms for Protein Folding Simulations," *Journal of Molecular Biology*, vol. 231, no. 1, pp. 75–81, May 1993.

EST-PAC^{HPC} – a web portal for high-throughput EST annotation and protein sequence prediction

Adam K.L. Wong¹, Andrzej M. Goscinski¹, Christophe Lefèvre^{2,3}

¹*School of Information Technology, Deakin University, Geelong, Australia*

aklwong_angf@deakin.edu.au

²*Institute for Technology Research and Innovation (ITRI), BioDeakin, Deakin University*

³*Victorian Bioinformatics Consortium, Monash University*

clefevre@deakin.edu.au

Abstract – Expressed Sequence Tags (ESTs) are short DNA sequences generated by sequencing the transcribed cDNAs coming from a gene expression. They can provide significant functional, structural and evolutionary information and thus are a primary resource for gene discovery. EST annotation basically refers to the analysis of unknown ESTs that can be performed by database similarity search for possible identities and database search for functional prediction of translation products. Such kind of annotation typically consists of a series of repetitive tasks which should be automated, and be customizable and amenable to using distributed computing resources. Furthermore, processing of EST data should be done efficiently using a high performance computing platform. In this paper, we describe an EST annotator, EST-PAC^{HPC}, which has been developed for harnessing HPC resources potentially from Grid and Cloud systems for high throughput EST annotations. The performance analysis of EST-PAC^{HPC} has shown that it provides substantial performance gain in EST annotation.

Keywords: Expressed Sequence Tag, High Throughput EST Annotations, EST Data Mining, Grid and Cloud Computing, Performance Evaluation.

BIOCAMP 2011

I. INTRODUCTION

High-end computing facilities such as grids and clouds [8] are the key to enabling bioinformatics projects in the next generation sequencing era. Different research groups both nationally and internationally could benefit by sharing research data, computing platforms and experiment results cost-effectively. Many research laboratories will have many terabytes if not petabytes of data to transfer, store and analyse. Handling and analysing such huge amount of genomic data require fast and reliable computer networks as well as a huge amount of computation power and storage. Although high-end supercomputers are now easily available to a broad scientific community, users without in depth I.T. knowledge are often forced to cope with many low-level details when using those machines for

scientific investigations. Cloud technologies in particular promise to provide seamless access to high performance computer clusters through the abstractions of services and brokers that hide the details of the underlying software and hardware infrastructure. In this paper, we describe an expressed sequence tag (EST) annotator, EST-PAC^{HPC}, which has been developed for EST annotation on a high performance computing platform.

BLAST [13] is probably the most worldwide used bioinformatics tool for sequence alignment and BLAST searching of ESTs is a key component task of EST annotation. A typical EST annotation procedure often needs to perform BLAST searching for a large volume of ESTs repeatedly on different genomic databases. Thus, such procedure should be executed on a HPC platform to leveraging the power of parallel processing. There are a number of programs and hardware solutions for efficient high-throughput BLAST searching in Grids [4] and Clouds [5]. However, there is a lack of generic software solutions for personalized management, presentation and mining of the search results. For this reason, downstream analysis remains a task to be solved in ad hoc ways by different users. On the other hand, other EST annotators [17] have concentrated on providing an intergraded annotation and data mining environment but have failed to handle the high throughput computational requirement of EST annotation. EST-PAC^{HPC} is a fully functional EST annotator, which performs using HPC resources potentially from various grid and cloud systems.

The rest of this paper is organized as follows. Section 2 provides the background knowledge of EST annotation. It also describes the EST-PAC and EST-PAC^{HPC} software packages. Section 3 explains the approach taken by EST-PAC^{HPC} for high throughput BLAST searching. Section 4 covers the performance evaluation of EST-PAC^{HPC} for EST annotation using BLAST searching on a HPC platform. The experimental

test-bed, workload construction and results of the performance evaluation are discussed. Finally, Section 5 presents the conclusions and our future work.

II. BACKGROUND

An expressed sequence tag (EST) is a short DNA sequence, usually 200 to 500 nucleotides long, that is generated by sequencing the transcribed cDNA sequence of an expressed gene. ESTs were used for the first time as a primary resource for human gene discovery [1]. Since then, there has been an exponential growth in the generation and accumulation of EST data, with approximately 69 million ESTs now available in public databases (GenBank 01 March 2011, all species). Since ESTs can provide significant functional, structural and evolutionary information, there are a lot of worldwide biological projects and laboratories that continually produce ESTs for different researching tasks. Many EST sequencing projects are underway for numerous organisms and extensive computational strategies have been developed to organize and analyse both small- and large-scale EST data for gene discovery, transcript and single nucleotide polymorphism analysis as well as functional annotation of putative gene products [17].

With the decreasing cost of DNA sequencing technology and the vast diversity of biological resources, researchers increasingly face the basic challenge of annotating a huge amount of EST data from a variety of species. EST annotation basically refers to the analysis of unknown ESTs that can be performed by database similarity search for possible identities and database search for functional prediction of translation products. Such kind of annotation typically consists of a series of repetitive tasks, which should be automated, and all these operations should be self-installing, platform independent, easy to customize and amenable to using distributed computing resources. Furthermore, processing of EST data should be done efficiently on high performance computing platforms.

A. EST-PAC

EST-PAC was developed as a web oriented multi-platform software package for EST annotation which can run on a single compute-server [19]. It has integrated the BLASTALL suite [13], EST-Scan2 [10] and HMMER [7] in a relational database system accessible through a web portal. The system allows users to customize annotation strategies and provides an open-source data-management environment for research and education in bioinformatics.

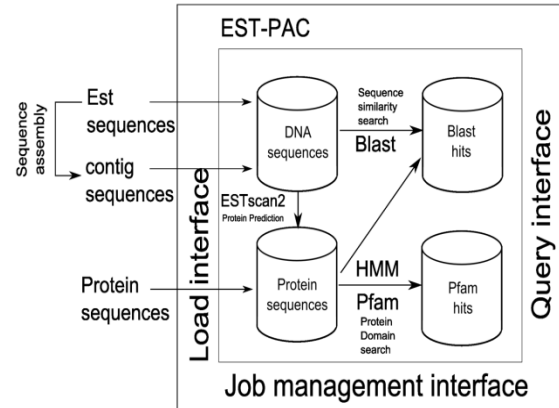


Fig. 1 Workflow and interfaces available in EST-PAC

The core of EST-PAC consists of an open source relational database management system that uses Structured Query Language, MySQL 5 [11], and a number of PHP 5 [15] programs, which allow the storage and management of ESTs using a web interface. The workflow of ESTs annotation is shown in Fig.1. User login is available for visualization and query, with additional privileges to run annotation tools. Sequences in FASTA format [9] are loaded into the database through a web interface and annotation tasks can be requested. A set of continuously running programs checks the database and extracts sequences to be processed using the BLASTALL suite, ESTScan2 or, HMMER.

The coding content of the EST can be evaluated with the Hidden Markov Model approach of ESTScan2 and the predicted translation products can then be compared against protein sequence databases. A report can be obtained from a web query page. As all results are stored in a relational database, users are able to query on every value returned by the annotation process. An interface is also available to assist the construction and storage of database queries. In addition to the public databases which can be downloaded and installed locally or accessed through web based blast services such as NCBI [12], users have the possibility to create their own databases from EST-PAC in order to make more precise and relevant comparisons.

B. EST-PAC^{HPC}

We have extended EST-PAC into EST-PAC^{HPC} which can utilize HPC resources such as computer clusters in Grids; and with a potential of using clouds resources for bioinformatics computing. The web portal approach of EST-PAC^{HPC} has enabled biologists who are not IT specialists to benefit directly from the use of high-performance computing technology. EST-PAC^{HPC} supports both high throughput and high performance computation of the selected bioinformatics applications.

To achieve high throughput computation, bioinformatics jobs from many users can run on different processors of a cluster concurrently. This solution has shortened the service waiting time. To achieve high performance computation, many of the submitted bioinformatics jobs can run on multiple processors of a cluster as parallel applications.

As shown in Fig. 2, an Apache [3] web-server with MySQL database system and PHP language script interpreter form the backbone of the EST-PAC^{HPC} Bio-Server, which is currently providing computation services to the Bioinformatics Research Group [6] at Deakin University. The heart of EST-PAC^{HPC} lays on its novel job-scheduling mechanism that integrates transparently with most of the existing cluster and grid resource managers such as PBS [16] and Sun Grid Engine [18]. Currently, the openMPI [14] parallel programming environment is adopted for parallel computation.

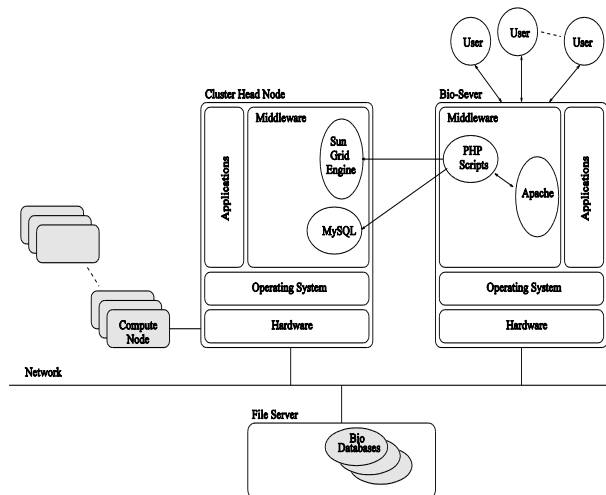


Fig. 2 Architecture of EST-PAC^{HPC} running as a bioinformatics computation server at Deakin University

A web-portal interface is provided in EST-PAC^{HPC} to release biologists from performing tedious I.T. tasks such as hardware setup, software installation and configuration as well as data management. Most importantly, it hides completely the details of bioinformatics application deployment in the underlying HPC platform. Fig. 3, Fig. 4 and Fig. 5 are the snapshots extracted from EST-PAC^{HPC} web portal; each of them shows the main functions, EST sequence handling and EST annotation correspondingly.

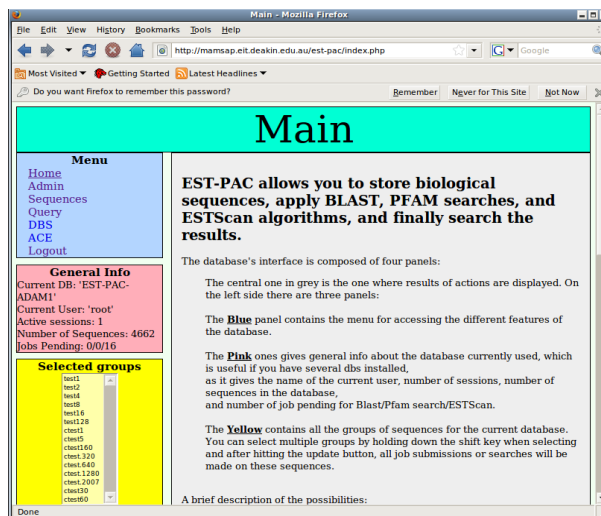


Fig. 3 Main page of the EST-PACHPC web portal

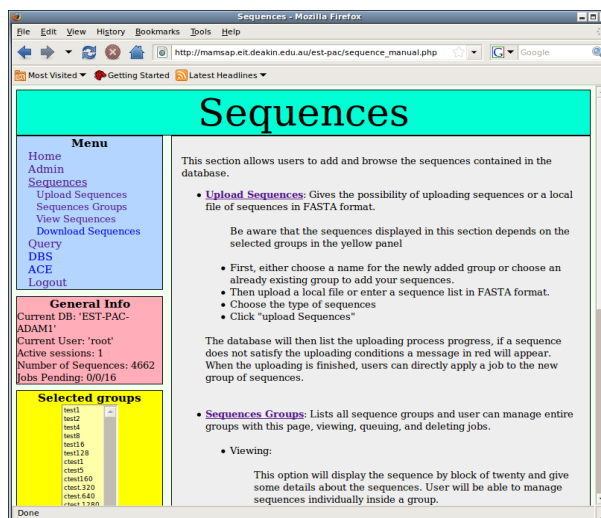


Fig. 4 Web page showing major functions for EST sequence handling

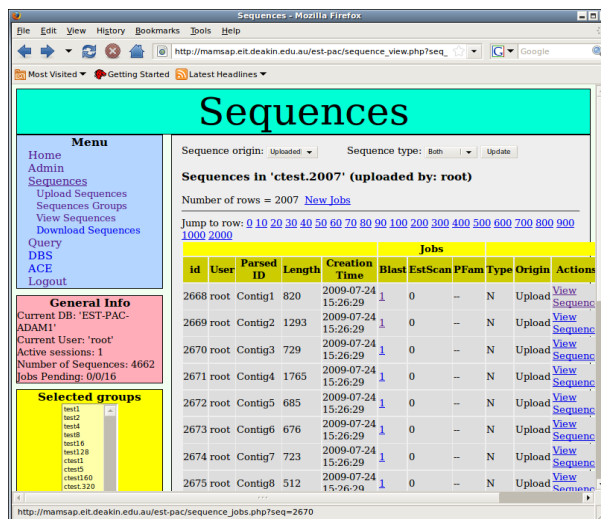


Fig. 5 Web page showing EST annotation job deployment

III. HIGH THROUGHPUT EST ANNOTATION IN EST-PAC^{HPC}

As mentioned in Section II, the core operation in EST analysis is a database similarity search, which assigns possible identities to the unknown ESTs. This can be done by the BLAST program. There are many publicly available resources for users to carry out BLAST operations that can be found such as a free resource from the National Centre for Biotechnology Information, NCBI-Blast [13] on the one side, to a pay-per-use resource from Windows Azure Blast [5] on the other.

The NCBI-Blast server, which is backed by a high-end supercomputer, provides a real time and high performance BLAST service to users. However, this free service has restricted users from carrying out high throughput BLAST searches. A submitted BLAST job of more than 50 sequences will be penalized in term of its dispatch time as the server is shared world wide. Besides, users' search results will not be kept in the NCBI databases indefinitely. The pay BLAST service from Windows Azure seems to be a flexible and cost-effective solution for carrying out high-throughput BLAST despite that it is still a trial service from Azure. Nevertheless, those service providers do not mean to provide EST annotation service to users. The downstream result analysis remains a task to be solved in ad hoc ways by users.

A. EST annotation: An ad hoc approach

Assuming BLAST is used to carry out the sequence similarity search, the basic steps of performing a high-throughput EST annotation are as follows:

1. Obtain and prepare copies of known genomic databases.
2. Obtain and prepare ESTs (short sequences).
3. Carry out BLAST search of ESTs on the known genomic databases.
4. Post-process BLAST search results: this refers to i) store results into DBMS system for further data mining process; and ii) visualize BLAST search results for ESTs analysis.
5. If necessary, repeat step 3 and 4 by replacing BLAST with different search tools such as EST-Scan2 and HMMER for coding region detection of DNA and protein sequence alignment.

Fig. 6 shows a workflow of high-throughput EST annotation via an ad hoc approach. As can be seen, users (mostly biologists) have to cope with many low-level details of BLAST parallelization as well as handling of annotation post-processing, which is tedious and time consuming.

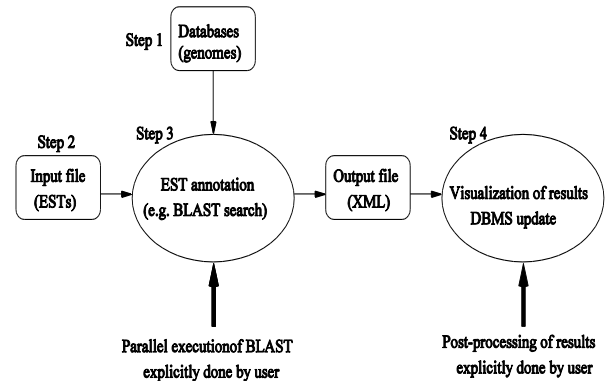


Fig. 6 Workflow of high-throughput EST annotation

B. EST annotation: The HPC-PAC^{HPC} approach

The web-portal interface provided in EST-PAC^{HPC} has simplified the tasks of Step 1 and Step 2 as described in the previous subsection (See Fig. 7). Once EST data are uploaded to the system, users can easily deploy an EST annotation, as corresponding to Step 3, via the web-portal (See Fig. 8). The running of BLAST searches on a computer system, e.g. HPC clusters, is completely transparent to users. The current implementation of EST-PAC^{HPC} has provided a job-scheduler, which can be integrated to most of the existing cluster and grid resource managers such as PBS and Sun Grid Engine, thus harnessing HPC resources potentially from various grids and clouds. Results of the BLAST searches are permanently stored in the MySQL DBMS and can be visualized in real time via the web-portal (See Fig. 9).

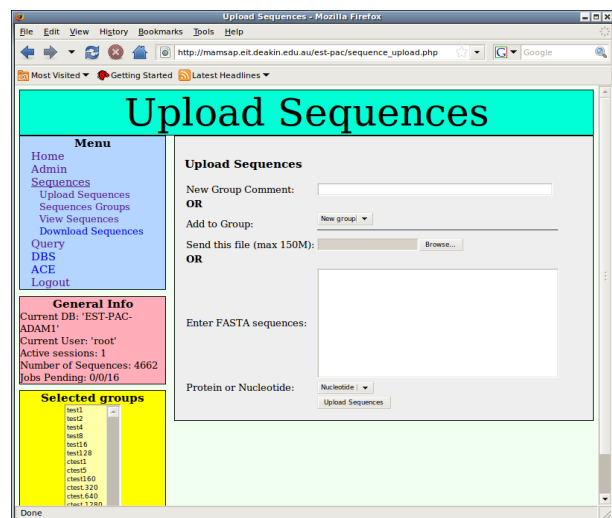


Fig. 7 Web page showing uploading of EST data

This calibration has also shown that reasonable speedup of EST annotation is achievable even for data sets of small number of EST sequences. However, we believe that there is still room for improvement in the speedup, especially in handling the concurrence of DBMS update operations.

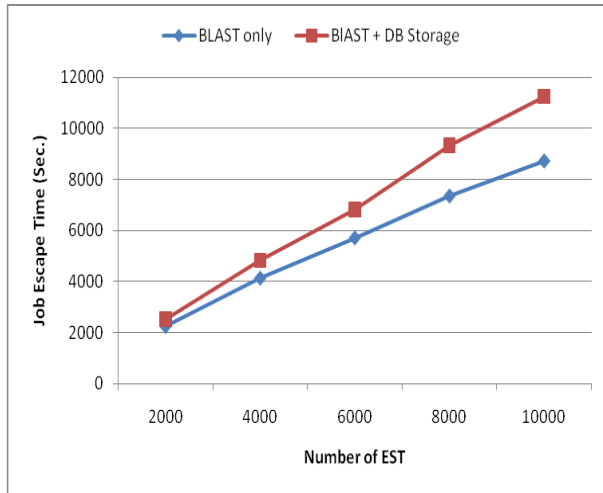


Fig. 10 Job Escape Time of EST annotation against EST size (Single computer server)

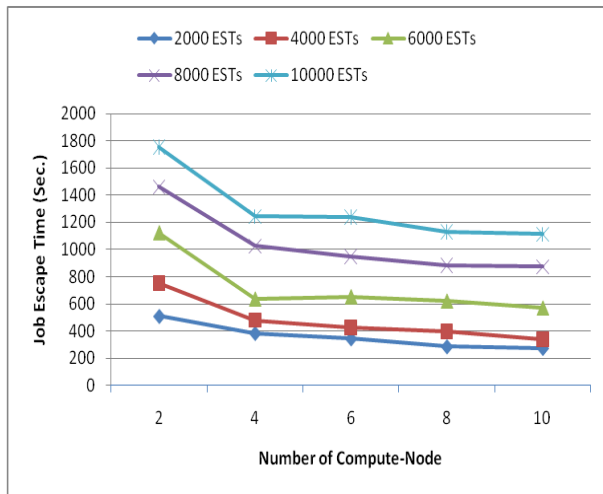


Fig. 11 Job Escape Time of EST annotation against No. of Compute-Node with different EST sizes (Computer server with HPC cluster)

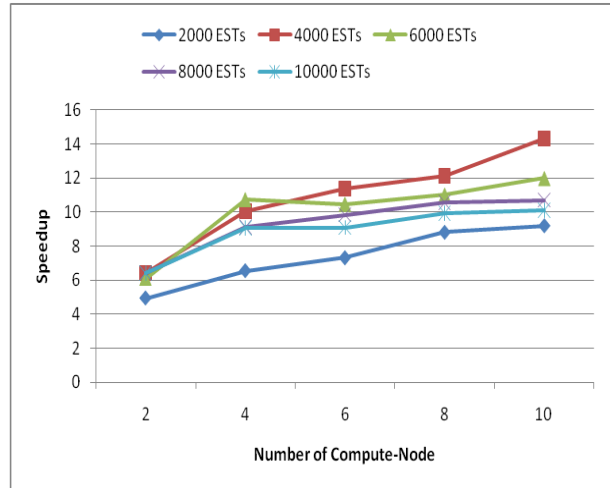


Fig. 12 Speedup of EST annotation against No. of Compute-Node with different EST sizes (Computer server with HPC cluster)

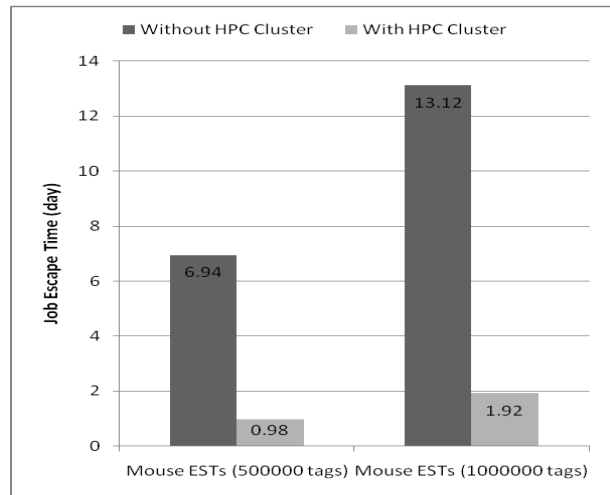


Fig. 13 Job Escape Time of Mouse EST annotations against Mouse Genomic Database

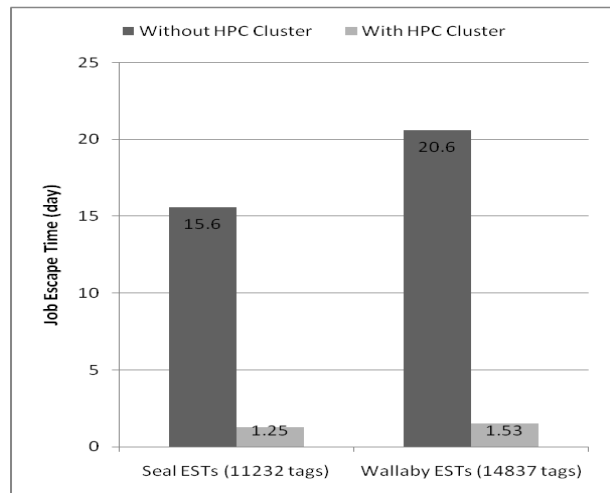


Fig. 14 Job Escape Time of Seal and Wallaby EST annotations against All-organism Genomic Database

Fig. 13 and Fig. 14 present the results obtained from the second experiment set. A promising improvement has been achieved in each of the EST annotations. Job escape time for the annotation of the Wallaby ESTs can be reduced from 20 days to less than 2 days.

V. CONCLUSIONS AND FUTURE WORK

We have extended the EST annotation software package EST-PAC to EST-PAC^{HPC} which can harness HPC resources potentially from various grid and cloud systems for high throughput EST annotations. The performance gain is substantial. The web-portal based approach of EST-PAC^{HPC} can remove the burden of biologists from performing tedious I.T. tasks such as hardware setup, software installation and configuration as well as data management. Even more, it also hides all the details of high performance computing from the users. In conclusion, EST-PAC^{HPC} provides an open framework for rapid prototyping of data mining and on-line visualization of sequence data, presenting an expandable data-management environment for research and education in bioinformatics.

Currently, we are extending the job-scheduling mechanism and the HPC job scheduler of EST-PAC^{HPC} to make it become cloud-enabled. Preliminary work has begun to study Amazons Elastic Compute Cloud (EC2) for HPC [2].

REFERENCE

- [1] Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* Vol. 252, Issue 5013, pp.1651–1656. Jun 1991.
- [2] Amazon. Amazon ec2 high performance computing. <http://aws.amazon.com/ec2/hpc-applications/>. Last access: April 2011.
- [3] Apache HTTP Server Project. <http://httpd.apache.org/>. Last access: April 2011.
- [4] Arun Krishnan. GridBLAST: a Globus-based high-throughput implementation of BLAST in a Grid computing framework: Research Articles. *Concurrency and Computation: Practice & Experience* archive, Vol. 17, Issue 13, pp. 1607-1623. John Wiley and Sons Ltd. UK. November 2005.
- [5] Azure NCBI Blast. <http://research.microsoft.com/en-us/projects/azure/azureblast.aspx>. Last access: April 2011.
- [6] Bioinformatics Research Group, Deakin University. http://mamsap.eit.deakin.edu.au/wiki/index.php/Main_Page. Last access: April 2011.
- [7] Eddy SR: Profile hidden Markov models. *Bioinformatics*. Vol. 14, pp. 755-763. 1998.
- [8] Gartner. Gartner highlights five attributes of cloud computing: <http://www.gartner.com/it/page.jsp?id=1035013>, last access: April 2011.
- [9] HUP0-PSI Standard FASTA Format. http://en.wikipedia.org/wiki/FASTA_format. Last access: April 2011.
- [10] Iseli C, Jongeneel CV, Bucher P: ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*. pp. 138-148. 1999.
- [11] MySQL. <http://dev.mysql.com>. Last access: April 2011.
- [12] National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov/>. Last access: April 2011.
- [13] NCBI Blast. Basic Local Alignment Tool from NCBI: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYP E=BlastHome, last access: April 2011.
- [14] OpenMPI. <http://www.open-mpi.org/>. Last access: April 2011.
- [15] PHP. <http://www.php.net>. Last access: April 2011.
- [16] Portable Batch System. <http://www.openpbs.org/>. Last access: April 2011.
- [17] Shivashankar H. Nagaraj, Robin B.Gasser and Shoba Ranganathan. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*. Vol. 8, No. 1, pp. 6-21. Advance Access Publication. May 23, 2006.
- [18] Sun Microsystems Inc. Sun Grid Engine. URL: <http://gridengine.sunsource.net/>. Last access: April 2011.
- [19] Yvan Strahm, David Powell and Christophe Lefèvre. EST-PAC a web package for EST annotation and protein sequence prediction. *Source Code for Biology and Medicine* 2006.

Collagen Type XI $\alpha 1$ Chain Amino Propeptide Structural Model and Glycosaminoglycan Interactions *in Silico*

Chris Mallory^{1,3}, Owen McDougal^{1,3}, Julia Thom Oxford^{2,3}

¹Department of Chemistry and Biochemistry, ²Department of Biological Sciences, ³Biomolecular Research Center, Boise State University, Boise, ID 83725, USA

Abstract-*Modeling of the collagen $\alpha 1(XI)$ amino propeptide (NPP) domain was performed to better understand how dimerization and glycosaminoglycan binding are coordinated. The program MODELLER was used to generate a homology model of collagen $\alpha 1(XI)$ NPP domain based on the crystal structure of the closely related NC4 domain of collagen $\alpha 1(IX)$ (PDB:2UUR) to a root mean square deviation (rmsd) of 0.785 Å resolution. A model of collagen $\alpha 1(XI)$ NPP domain dimer was constructed in two alternative templates; 1) the thrombospondin dimer template (PDB:1Z78), and 2) by submission of two monomer subunits based on PDB:2UUR to ClusPro. Calculation of relative binding energy for the interaction between each collagen $\alpha 1(XI)$ NPP model and glycosaminoglycans as ligands was performed using AutoDock4. Results support a higher affinity between heparan sulfate and the dimer compared to the monomer. Sequential point mutation studies in the putative binding site (147-KKKITK-152) indicated the importance of each basic lysine residue in the binding of heparan sulfate. Two orders of magnitude change in binding affinity was predicted when comparing wild type to the mutation K152A.*

Keywords: collagen, heparin, molecular interaction, glycosaminoglycan, protein

1 Introduction

Collagen is a triple helical protein comprising approximately 25% of the protein contained in the human body. The triple helix of collagen is unique due to its formation from three left-handed helical strands to compose the right-handed triple helix. The strands that make up this triple helix have the sequence Gly-Xxx-Yyy; where approximately 30%

of the Xxx and Yyy are proline and hydroxyproline, respectively [1-3]. To date more than 27 different collagens have been reported in the literature, of which 16 have non-collagenous domains attached to the extended collagen triple helix. Collagen type XI is a minor fibrillar collagen involved in regulating the diameter of collagen fibrils [1,4]. Composed of three different left-handed helical strands $\alpha 1$, $\alpha 2$, and $\alpha 3$, each of the alpha chains contains non-collagenous domains. The $\alpha 1$ amino terminal non-collagenous domain (NPP) is proteolytically cleaved at a much slower rate than $\alpha 2$ or $\alpha 3$ and is therefore resident on the surface of collagen fibrils for an extended period of time in tissues [1]. Currently, a protein data base structure file is not publically available for the Npp $\alpha 1$, but a recently published structure for the NC4 domain of collagen IX is available that demonstrates remarkable structural similarity to collagen XI [5]. The NPP domain is included in a family of laminin, neurexin, sex hormone binding globulin (LNS) domains, in which a crystal structure of thrombospondin has also recently appeared in the literature in a monomer and dimer form [6].

There is experimental evidence to support the hypothesis that the Npp domain of collagen $\alpha 1(XI)$ interacts with glycosaminoglycans such as heparan sulfate. This interaction is proposed to be significant in determining the thickness of the fibril as it forms [4]. These interactions occur at a glycosaminoglycan binding within collagen $\alpha 1(XI)$ consisting of 147-KKKITK-152. Independently, experimental evidence indicates the formation of an NPP $\alpha 1(XI)$ dimer. Using a combination of homology modeling and protein-protein docking, a computational model for the $\alpha 1(XI)$ NPP was created. Models were used in the docking program AutoDock4 to calculate the energy of interaction between collagen $\alpha 1(XI)$ NPP (monomer and dimer) with the heparan sulfate ligand [7]. The results of the docking study provide a

theoretical inhibition constant (K_i), molecular binding energies, and predicted atomic interactions such as hydrophobic, electrostatic, and hydrogen bonding.

The conservation and importance of each basic lysine residue in the putative binding site (147-KKKITK-152) were further evaluated by point mutations. Each positively charged residue was suspected to be critical for glycosaminoglycan binding and was analyzed by computational docking analysis of the point mutants in AutoDock4.

2 Methods

2.1 Collagen $\alpha 1(XI)$ NPP homology model monomer

The sequence for Collagen $\alpha 1(XI)$ NPP was submitted to BLAST against available PDB structures for template identification [8]. Blast returned two possible templates; thrombospondin (PDB:1Z78) and the NC4 domain of collagen $\alpha 1(IX)$ (PDB:2UUR) [5,6,8]. The sequences for each template were aligned independently to the Collagen $\alpha 1(XI)$ NPP domain, and then collaboratively using the '*salign*' command in MODELLER [9]. Homology models for Collagen $\alpha 1(XI)$ NPP domain using both 2UUR and 1Z78 as templates were then created using the MODELLER method with disulfide bonding specified between Cys 25–Cys 207 and Cys 146–Cys 200 [1,9]. The returned models were evaluated using PROCHECK for areas of high energy, restricted points on the Ramachandran plot, and residue clashes [10]. Loop rebuilding and energy minimization were performed in MODELLER, while corrections to amino acids contained in restricted areas of the Ramachandran plot were amended in Chimera [11]. An rmsd was calculated for the homology model generated from each template (2UUR & 1Z78), using Chimera [11]. The homology model created from 2UUR (HM1) was selected as the best model based on sequence alignment, query match, E value, and an rmsd of 0.785Å from template structure. Swiss Deep View was used to make the following single-point mutations into the Collagen $\alpha 1(XI)$ NPP domain putative binding site: K147A, K148A, K149A, and K152A [12].

2.2 Collagen $\alpha 1(XI)$ homology model dimer

A dimer model of the Collagen $\alpha 1(XI)$ NPP was created with MODELLER and ClusPro using the thrombospondin dimer (PDB:2ES3) as a template.

The homology model was created by the same process as the monomer model with the exception that a repeated sequence of Collagen $\alpha 1(XI)$ NPP was used. Model evaluation was performed using PROCHECK, Verify3d, and Ramachandran plot analysis. A second dimer model was created by submitting two identical homology model monomers to the ClusPro server [13]. The overall lowest energy model was selected from the balanced interaction cluster.

2.3 Docking Studies

Docking studies between the Collagen $\alpha 1(XI)$ NPP monomer homology model, single-point putative binding site mutants, MODELLER dimer, and the ClusPro dimer with the glycosaminoglycan heparan sulfate (disaccharide, decasaccharide, and Arixtra, a low molecular weight heparin) as the ligand were performed using AutoDock4 [7]. The AutoDock standard conditions for a large run were used with a grid spacing of 0.379 Å. Binding energies were evaluated using WordPad, with residue interactions visualized using Chimera [11].

3 Results

3.1 Homology model

BLAST search results indicated that the NC4 domain of collagen IX (PDB:2UUR) was the best template with an E value of 5.0×10^{-5} . Alignment and model creation was performed using MODELLER with PDB:2UUR as a template. The computational model of the Collagen $\alpha 1(XI)$ NPP monomer from 2UUR was found to have an rmsd of 0.785 Å (Fig. 1a) [5,11]. Point mutations were introduced into the putative binding site sequentially using Swiss Deep View. Docking interactions using a heparan sulfate disaccharide and Collagen $\alpha 1(XI)$ NPP monomer models indicated that all lysine residues were necessary for best binding and lowest estimated inhibition constant (Table I). Submission of the wild type structural monomer to ClusPro, resulted in clusters returned based on hydrophobic, electrostatic, van der Waals + electrostatic, and balanced interactions. Without previous dimerization knowledge the lowest energy balanced model with a weighted score of -918.2 was selected, as recommended by ClusPro (Fig. 1b) as the optimal dimer model [13-16].

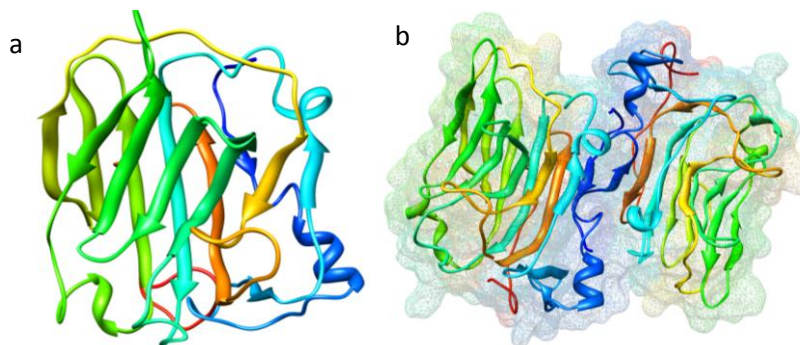


Fig. 1: Structural depictions of: a) Homology model monomer built from the template 2UUR; b) Lowest energy balanced interaction dimer from ClusPro [5,13].

Table I

Monomer homology model and mutant interactions between heparan sulfate disaccharide.

Model	Estimated Free Energy (kcal/mol)	Estimated K_i	Hydrogen bond residue interactions	Polar residue interactions
Wild Type	-8.08	1.20 μ M	ARG119, LYS149	ARG119, LYS149
K147A	-8.37	727.94 nM	PHE118, LYS152	PHE118, LYS152
K148A	-8.41	687.57 nM	LYS152	LYS152
K149A	-7.42	3.64 μ M	LYS148	LYS148
K152A	-5.34	121.85 μ M	-	LYS149

*The estimated inhibition constant (K_i) describes the binding affinity for the ligand to the receptor. This is not to be confused with the dissociation constant (K_d) that provides insight into how easily the receptor-ligand complex separates into its individual components, the receptor and ligand.

3.2 Binding Studies

Docking results for heparan sulfate and each dimer model confirmed the hypothesis that dimerization

increases affinity for glycosaminoglycans relative to the monomer, as shown in Table II.

Table II

Interactions between ClusPro dimer and a heparan sulfate disaccharide, deca-saccharide, and Arixtra[7,13-16].

Ligand	Est. Free Energy of Binding (kcal/mol)	Est. K_i	Electrostatic Energy (kcal/mol)	Total Intermolecular Energy (kcal/mol)	Hydrogen bond residue interactions	Polar residue interactions
Heparan sulfate disaccharide	-17.53	142.31 fM	-10.57	-12.65	-	LYS73, ASN122, LYS149
Heparan sulfate deca-saccharide	-21.13	327.17 aM	-22.22	-23.18	LYS148	LYS73, ARG97, ARG119, LYS148, TYR197
Arixtra	-7.93	1.53 μ M	-5.07	-7.92	LYS73	LYS72, ARG97, ASP125

4 Discussion

Point mutation docking studies revealed that K148A had the lowest estimated binding energy, providing a

preliminary hypothesis that K148 is the least critical residue involved in binding of glycosaminoglycans. Furthermore, K152 appears to be the most important residue in the putative binding site involved in glycosaminoglycan

interaction as evidenced by a predicted two orders of magnitude increase in the K_d . Initial experimental results suggest that a greater affinity for glycosaminoglycans occurs upon dimerization or oligomerization of Collagen $\alpha 1(XI)$ NPP. Binding results for heparan sulfate disaccharide, decasaccharide, and Arixtra support the experimental results of increasing affinity for glycosaminoglycans upon dimerization of the Collagen $\alpha 1(XI)$ NPP. While it is clear that they are coordinated, it is still unclear whether dimerization is facilitated upon interactions with glycosaminoglycans or alternatively, if dimerization occurs prior to binding, and subsequently facilitates interaction with glycosaminoglycans.

5 Conclusion

To date there is no crystal structure available for the Collagen $\alpha 1(XI)$ NPP domain. Using computational modeling, we have created a homology model based on the NC4 domain of collagen IX (PDB:2UUR) as a template. The resulting model was used to investigate glycosaminoglycan binding and provided insight into protein:glycosaminoglycans interactions. *In silico* prediction provided preliminary insight into the importance of K152 in the binding interactions with glycosaminoglycans, from a 100 fold decrease in binding affinity compared to that of the wild type.

Furthermore, dimerization of the Collagen $\alpha 1(XI)$ NPP domain has been observed experimentally to increase the binding affinity to glycosaminoglycans. This experimental result was successfully replicated through the modeling of a dimer and docking interactions of heparan sulfate and its derivatives. Additional studies are being conducted to determine if glycosaminoglycan binding induces dimerization, or alternatively, if increased affinity for glycosaminoglycans is a result of Collagen $\alpha 1(XI)$ NPP dimerization.

Acknowledgments

Authors wish to acknowledge technical and editorial support from Luke Woodbury. This work was supported in part by grants from the National Institutes of Health/National Center for Research Resources Grant P20RR16454, NIH/NICHD R15HD059949, NASA (NNX10AN29A), Arthritis Foundation, Research Corporation Cottrell College Scholars, and Mountain States Tumor Medical Research Institute, National Institutes of Health/NIAMS RO1AR47985 and KO2AR48672, the M. J. Murdock Charitable Trust, and Lori and Duane Stueckle Professorship.

References

- [1] A. Fallahi, B. Kroll, L.R. Warner, R.J. Oxford, K.M. Irwin, L.M. Mercer, S.E. Shadle, J.T. Oxford, "Structural model of the amino propeptide of collagen XI $\alpha 1$ chain with similarity to the LNS domain," *Prot. Sci.*, 14, 1526-1537, 2005.
- [2] N.P. Morris, H.P. Bachinger, "Type XI Collagen Is a Heterotrimer with the Composition (1α , 2α , 3α) Retaining Non-triple-helical Domains," *J. Biol. Chem.*, 262, 11345-11350, 1987.
- [3] M.K. Gordon, R.A. Hahn, "Collagens," *Cell Tissue Res.*, 339, 247-257, 2010.
- [4] D.R. Keene, J.T. Oxford, N.P. Morris, "Ultrastructural Localization of Collagen Types II, IX, and XI in the Growth Plate of Human Rib and Fetal Bovine Epiphyseal Cartilage: Type XI Collagen is Restricted to Thin Fibrils," *J. Histochem. Cytochem.*, 43, 967-979, 1995.
- [5] K. Tan, M. Duquette, J.H. Liu, K. Shanmugasundaram, A. Joachimiak, J.T. Gallagher, A.C. Rigby, J.H. Wang, J. Lawler, "Heparin-induced cis- and trans-dimerization modes of the thrombospondin-1 N-terminal domain," *J Biol Chem.*, 283:3932-41, 2008.
- [6] V.-M. Leppanen, H. Tossavainen, P. Permi, L. Lehtio, G. Ronnholm, A. Goldman, I. Kilpelainen, T. Pihlajamaa, "Crystal structure of the N-terminal NC4 domain of collagen IX, a zinc binding member of the laminin-neurexin-sex hormone binding globulin (LNS) domain family," *J. Biol. Chem.*, 282, 23219-23230, 2007.
- [7] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, "Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function," *J. Comput. Chem.*, 19, 1639-1662, 1998.
- [8] Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezuk, Y.; McGinnis, S.; Madden, T. L. "NCBI BLAST: A Better Web Interface," *Nucleic Acids Res.*, 36, W5-W9, 2008.
- [9] M.A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali, "Comparative Protein Structure Modeling of Genes and Genomes," *Annu. Rev. Biophys. Biomol. Struct.*, 29, 291-325, 2000.
- [10] R.A. Laskowski, "PROCHECK: a program to check the stereochemical quality of protein structures," *J. Appl. Cryst.*, 26, 283-291, 1993.
- [11] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, "UCSF Chimera-A Visualization System for Exploratory Research and Analysis," *J. Comput. Chem.*, 25, 1605-1612, 2004.
- [12] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling," *Bioinformatics*, 22, 195-201, 2006.
- [13] D. Kozakov, D.R. Hall, D. Beglov, R. Brenke, S.R. Comeau, Y. Shen, K. Li, J. Zheng, P. Vakili, I.C. Paschilidis, S. Vajda, "Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19," *Proteins: Struct.,Funct., Bioinf.*, 78, 3124-3130, 2010.
- [14] D. Kozak, R. Brenke, S.R. Comeau, S. Vajda, "PIPER: An FFT-Based Protein Docking Program with Pairwise Potentials," *Proteins: Struct.,Funct., Bioinf.*, 65, 392-406, 2006.
- [15] S.R. Comeau, D.W. Gatchell, S. Vajda, "Camacho, C. J. ClusPro: An Automated Docking and Discrimination Method for the Prediction of Protein Complexes," *Bioinformatics*, 20, 45-50, 2004.
- [16] S.R. Comeau, D.W. Gatchell, S. Vajda, C.J. Camacho, "ClusPro: A Fully Automated Algorithm for Protein-Protein Docking," *Nucleic Acids Res.*, 32, W96-W99, 2004.

Probability Density Profile Analysis: A Method for Identifying Novel Protein Structures

Arjang Fahim¹, Stephanie Irausquin¹, Matthew Fawcett¹, Mikhail Simin¹, and Homayoun Valafar¹

¹Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA

Abstract - Although a number of scientific advances have been made in the area of structural biology, a few obstacles continue to impede our ability to quickly and efficiently characterize protein structure-function relationships. Probability Density Profile Analysis (PDPA) is a method which rapidly quantifies the structural novelty of a protein, based on the statistical analyses of a minimal amount of empirical data. Here we present findings related to the sensitivity and range of applicability of PDPA. Our results support the conclusion that two dimensional PDPA (2D-PDPA) can reliably be utilized for identification of a protein structure to within 3Å of the known structure, using a library of existing structures. Furthermore, the sensitivity of 2D-PDPA has been tested using proteins containing different secondary structural characteristics (α , β , and α/β) and our preliminary investigations support the conclusion that 2D-PDPA is equally applicable to all general classes of proteins.

Keywords: Residual Dipolar Couplings, Parzen Density Estimation, Probability Density Profile Analysis, Structural Homology Detections

1 Introduction

Proteins are often referred to as the working molecules of a cell, performing many important structural, functional and regulatory processes [1]. Yet, revealing the function of proteins is a particularly challenging problem. Sequence-based approaches are an option, but identifying functionally characterized homologs is only feasible for less than half of the proteins predicted from genome sequencing projects [2] and is often compounded by the fact that proteins tend to be multi-functional [3]. Since a protein's structure often dictates its function, an alternative approach is to determine the structure of the protein of interest in order to identify functionally important sites [3]. This is believed to provide a solution for many of the remaining proteins, since structure is more evolutionarily conserved than sequence [2, 3].

Although the characterization of any protein adds to repositories of structural data, most structural biologists would concur that novel structures are particularly important for a number of reasons: they generate models of similar proteins for comparison; identify evolutionary relationships; further contribute to our understanding of protein function and mechanism; and allow for the fold of other family members to be inferred [4-6]. Considering the evolutionary mechanisms responsible for the generation of new structures in proteins, it has been speculated that there may be a limited

number of unique protein folds - as few as ten thousand families [7-9]. Currently the Protein Data Bank (PDB; [10]) consists of nearly 68,000 protein structures, but less than 1,400 families are represented and approximately no new fold families have been reported since 2008 [11, 12]. Ideally, solved protein structures for new protein families [6] would be used as templates for *in silico* structure prediction methods [4, 13] and the results of both solved and predicted structures would in turn be used to infer function [2, 14, 15]. However, such an approach requires new, efficient and cost-effective computational methods for target selection and structure determination.

Traditional methods of structure determination, such as X-ray crystallography and NMR spectroscopy, are expensive and time-consuming techniques. Previously we presented a method, referred to as Probability Density Profile Analysis (PDPA), which rapidly quantifies the structural novelty of a protein using only a minimal amount of empirical data. PDPA is a potentially important tool that provides investigators with fast, cost-effective, easy to interpret results while also further contributing to our understanding of structure-function relationships in proteins. The interpretation of PDPA scores, as well as the effective applicable range of PDPA, had not been known previously. In this report, we provide the means to interpret PDPA results and establish both the sensitivity and applicability of this method [19, 23].

2 Methods

2.1 Residual Dipolar Coupling (RDC)

Residual Dipolar Couplings are the result of dipolar interactions in a partially ordered system [16] and are defined in Equation (1):

$$D_{ij} = D_{max} \cdot \left\langle \frac{3 \cos^2(\theta) - 1}{2} \right\rangle \quad (1)$$

In this equation, D_{ij} is the magnitude of calculated RDC in hertz that is between two $\frac{1}{2}$ spin nuclei in the presence of a magnetic field; θ signifies the angle between the magnetic field vector and the inter-spin vector (nuclei i and j); brackets represent the time average for a specific coupling; and D_{max} denotes the maximum magnitude of a coupling that is further defined in Equation (2).

$$D_{max} = -\left(\frac{\mu_0}{4\pi}\right) \frac{\gamma_i \gamma_j h}{2\pi^2 r_{ij}^3} \quad (2)$$

In Equation (2), μ_0 signifies the magnetic permeability; γ_i and γ_j are the gyromagnetic ratios of two nuclei (i and j); r is the intranuclear distance between two nuclei; and h is Planck's constant.

The RDC equation can be manipulated into a matrix form (Equation (4)) as shown in Equation (3):

$$D_{ij} = v_{ij} \cdot S \cdot v_{ij}^T \quad (3)$$

$$S = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{bmatrix} \quad (4)$$

A unit vector that joins two corresponding nuclei is represented by v_{ij} and S is the traceless and symmetric Saupe order tensor matrix (OTM) [17]. S can be further decomposed into $S = RS'R^T$ such that R is a Euler rotation matrix, whose columns are the eigenvectors of S ; and S' (Equation 5) is a traceless diagonal matrix of the eigenvalues of S , whose diagonal elements $S'_{xx}, S'_{yy}, S'_{zz}$ are the principle order parameters (POP).

$$S' = \begin{bmatrix} S'_{xx} & 0 & 0 \\ 0 & S'_{yy} & 0 \\ 0 & 0 & S'_{zz} \end{bmatrix} \quad (5)$$

The rotation matrix R can be decomposed into three different rotations related to x , y and z as shown in Equation (6):

$$R(\alpha, \beta, \gamma) = R_z(\alpha) R_y(\beta) R_x(\gamma) \quad (6)$$

Using the previous equations, the order tensor can be rewritten in five parameters: $S'_{xx}, S'_{yy}, \alpha, \beta, \gamma$. This particular parameterization is used in our experiment to generate RDC data sets.

2.2 1D-PDP Analysis

Our initial work with PDPA was conducted using One Dimensional Probability Density Profile Analysis (1D-PDPA) and was based on unassigned RDC data from one alignment medium [18]. This proof of concept established the feasibility of identifying homologous structures from unassigned RDC data, however it lacked the potential for large scale applications. In summary, 1D-PDPA established structural similarity on the basis of comparing the distribution of experimental and computed RDC data. 1D-

PDPA requires a collection of experimental unassigned RDCs as well as a library of potential structures.

2.3 2D-PDPA

2D-PDPA extends the analysis of 1D-PDPA by utilizing RDC data from two alignment media. The additional set of experimental RDC data has obvious advantages over 1D-PDPA. 2D-PDPA limits the search space to seven parameters [19] and is capable of generating a more accurate and unique PDP for a given structure. A 2D-PDPA analysis session requires a collection of RDC data from two alignment media along with a library of homologous structures. A two dimensional Parzen density estimation (or kernel density estimation) is used to generate a two dimensional PDP (2D-PDP) by considering both alignment media [19]. Figure 1 illustrates a sample 2D-PDP for the protein Pf2048, a structure which has not yet been characterized. The two dimensional distribution of RDCs that is generated from the experimental data is denoted as the query PDP (qPDP) and is used, in addition to the estimated order tensors, as input to the 2D-PDPA. Incorporation of RDC data from the second alignment medium requires an extension of the search space by three more variables representing possible orientations of the second alignment medium with respect to the first one. Traditional inclusion of these three additional variables would have increased computation time by a factor of $2.5657e+09$. This intractable increase in computation time has been eliminated based on new technology that has been recently introduced [19, 20].

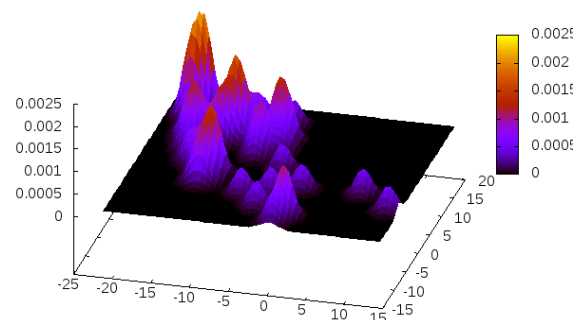


Figure 1. An example 2D-PDP signature for a protein (Pf2048) of unknown structure.

2D-PDPA calculates PDP for every rotation and a scoring method is used to find the best structure in terms of the similarity to the qPDP. To calculate fitness scores we consider three metric systems: Manhattan Block, Chi-Square, and Modified Chi-Square. The Manhattan Block method is defined in Equation (7):

$$S(qPDP, cPDP) = \sum_{i \in M} |(q_i - c_i)| \quad (7)$$

In Equation (7), q_i represents the i^{th} value of $qPDP$ and c_i represents the i^{th} value of computed PDP ($cPDP$). M denotes the number of sampled points in both query and calculated

PDP sets. The Chi-Square method is defined in Equations (8) and (9):

$$\delta_i^2(q_i, c_i) = \frac{(q_i - c_i)^2}{q_i} \quad (8)$$

$$\chi^2(qPDP, cPDP) = \sum_{i \in M} \delta_i^2(q_i, c_i) \quad (9)$$

In Equation (8), q_i represents the i^{th} value of $qPDP$ and c_i represents the i^{th} value of $cPDP$. Due to the asymmetric nature of the χ^2 metric ($\chi^2(A, B) \neq \chi^2(B, A)$), a modified Chi-Square has been introduced and shown in Equation (10):

$${}_m\chi^2(qPDP, cPDP) = \frac{[\chi^2(qPDP, cPDP) + \chi^2(cPDP, qPDP)]}{2} \quad (10)$$

In equation (10), ${}_m\chi^2$ denotes the modified χ^2 metric and $qPDP$ and $cPDP$ represent the experimental and computed PDPs, respectively. The modified χ^2 metric is a symmetric measure of distance and it therefore constitutes a formal metric space. During our early investigations, no preference was given to any one of the scoring metrics described above. However, based on the investigation that is presented here, the Manhattan Block metric was able to demonstrate slightly better results (shown in Figure 3a-c) in terms of the distribution of scores over bb-rmsd and as well as greater R^2 values.

2.4 Data Preparation

In this experiment, three reference proteins of different sizes and structural types (Table 1) were utilized in order to assess the sensitivity and selectivity of 2D-PDPA. This step is necessary due to the influence of secondary structures on orientation of the backbone N-H vectors. Traditionally, RDC data from helical regions have been reported to carry less information relative to other secondary structures. The proteins listed in (Table 1), were obtained from PDB [10]; Figure 2 provides a cartoon representation of each of the structures listed in Table 1.

Table 1. Protein structures obtained from the Protein Data Bank.

Protein	Secondary Structure	Number of Residues	CATH Classification
1A1Z	α	91	1.10.533
1OUR	β	114	2.60.120.400
1GB1	α/β	56	3.10.20.10

For each protein structure listed in Table 1, a set of one thousand structural variations were created by randomly altering the backbone ϕ and ψ torsion angles. Each dataset represented structural variations in the range of 0-8 Å with respect to the corresponding reference structure and were generated in the PDB file format. To obtain the RDC data for the three reference proteins, we utilized REDCAT [21]. The assignment information was discarded prior to providing

these data to 2D-PDPA. The PDB files were exported in REDCAT [21] to retrieve the RDC data in the 2D-PDPA program. Two sets of ^{15}N - ^1H backbone RDC data, representing two typical alignment media, were calculated for each reference protein by using REDCAT and the initial order parameters shown in Table 2. The RDC sets were calculated separately under three conditions: with one set containing no error, the second set corrupted through the addition of uniform noise in the range of $\pm 1\text{Hz}$, and the third set consisted of randomly eliminating 15% of RDC data that is normally expected during pragmatic conditions. The first set serves to simulate the ideal conditions (no error) versus the real conditions ($\pm 1\text{Hz}$ and 15% of RDC gap for the second and third sets).

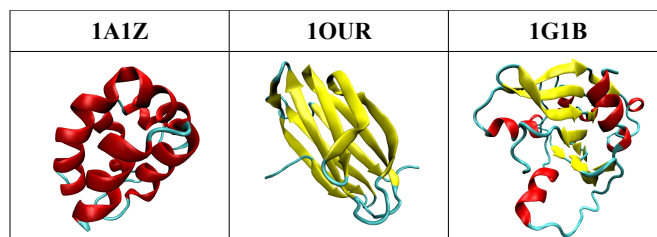


Figure 2. Illustrates the structures listed in Table 1.

Table 2. List of initial order parameters used to calculate two RDC sets.

	Sxx	Syy	Szz	Alpha	Beta	Gamma
Set1	3.00e-4	5.00e-4	-8.00e-4	0°	0°	0°
Set2	-4.00e-4	-6.00e-4	1.00e-3	40°	50°	-60°

The 2D Parzen Density Estimation [18] program was used to analyze RDC data and to create the 2D Probability Density Profile (2D-PDPA) [16] finger prints of each protein. Order tensors were calculated in two ways: First, the optimal 2D order tensors are obtained from REDCAT using structure and calculated RDC data; Second, order tensors are estimated using RDC data from two alignment media [22]. These two approaches represent a transition from ideal to more pragmatic conditions.

3 Results and Discussion

3.1 Experiment 1

The main objective of this experiment was to identify differences between the various metrics in order to establish the most appropriate metric for use. Experiment 1 used protein 1GB1 and its corresponding calculated RDC data, using no error or noise to demonstrate the ideal conditions. The experiment was repeated 3 times with different metrics each time. Order tensor matrices were obtained from REDCAT [21] for each RDC set (Table 3).

The relationship between 2D-PDPA structure scores and bb-rmsd for the one thousand variable structures generated is shown for each metric in Figure 3; the corresponding least squares regression logarithmic line and R^2 values are also shown for each metric. For bb-rmsd values up to 2.5Å, a linear correlation between PDPA scores and bb-rmsd exists

(Figure 3). For structures with bb-rmsd greater than 2.5\AA , PDPA scores remain in the same range: [0.8-1] for Manhattan Block, [2-3] for Modified Chi-Square, and [10-15] for Chi-Square (Figure 3). For all metrics tested, the Manhattan Block obtained the highest R^2 value (0.65, Figure 3). Therefore, the Manhattan-Block metric was selected and utilized exclusively for all remaining experiments.

Table 3. List of order parameters for each RDC set (alignment medium) obtained from REDCAT.

Order Tensor	No Error (1G1B)	$\pm 1\text{Hz}$ Error (1G1B)	15 RDC Gap (1G1B)	$\pm 1\text{Hz}$ Error (1A1Z)	$\pm 1\text{Hz}$ Error (1OUR)
Sxx1	$3e-4$	$2.966e-4$	$2.967e-4$	$3.091e-4$	$3.022e-4$
Syy1	$4e-4$	$5.08e-4$	$5.061e-4$	$4.985e-4$	$5.053e-4$
Sxx2	$7.99e-5$	$9.624e-5$	$9.726e-5$	$8.665e-5$	$-3.235e-5$
Sxy2	$3.89e-4$	$3.863e-4$	$3.795e-4$	$3.936e-4$	$4.05e-4$
Sxz2	$5.42e-4$	$5.412e-4$	$5.428e-4$	$5.44e-4$	$6.325e-4$
Syy2	$-1.70e-4$	$-1.784e-4$	$-1.82e-4$	$-1.856e-4$	$-5.998e-5$
Syz2	$5.414e-4$	$5.445e-4$	$5.396e-4$	$5.369e-4$	$4.348e-4$

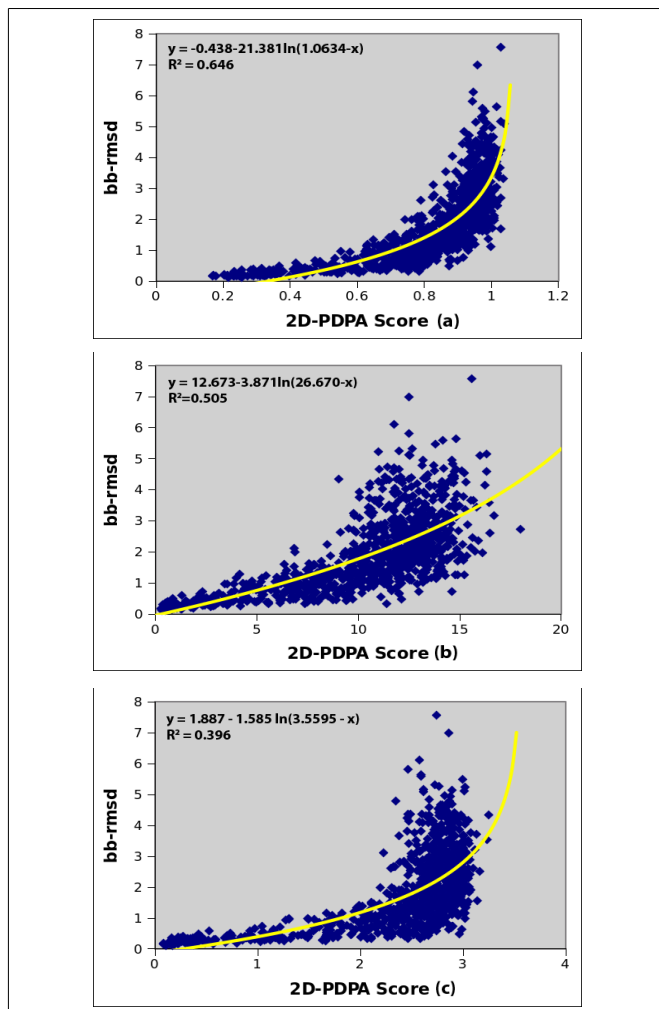


Figure 3. Calculated 2D-PDPA scores vs bb-rmsd using different scoring methods for 1GB1 protein: (a) Block scoring method, (b)

Chi-square scoring method, and (c) Modified chi-square scoring method.

3.2 Experiment 2

The objective of this experiment was to study the behavior of 2D-PDPA as a function of experimental noise. Experiment 2 used 1GB1 and calculated the RDC data using $\pm 1\text{Hz}$ error to demonstrate noisy conditions. Order tensor matrices were obtained from REDCAT [21] for each RDC set (Table 3).

Figure 4 shows the relationship between the 2D-PDPA's score (Manhattan-Block distance) for one thousand structures and their corresponding bb-rmsd with respect to the original structure; the least squares regression line and R^2 for the data are also shown in Figure 4.

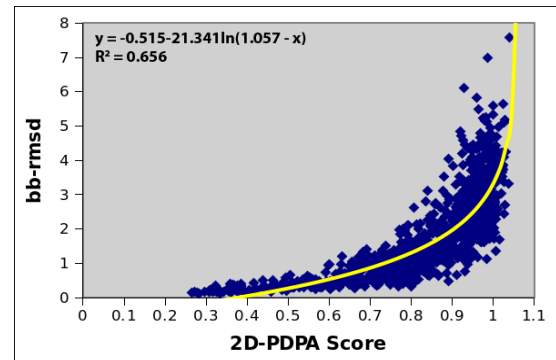


Figure 4. Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1GB1 with $\pm 1\text{Hz}$ error added.

This experiment was repeated by randomly removing 15 (28%) RDC values from both synthetic RDC data sets. Order tensor matrices were obtained from REDCAT for each RDC set (Table 3). The plot of the 2D-PDPA scores using block metric against the bb-rmsd is seen in Figure 5 along with the least squares regression line and R^2 value. The PDPA scores increase as a result of the random removal of RDC data, however a correlation still exists between PDPA score and bb-rmsd ($R^2=0.573$, Figure 5).

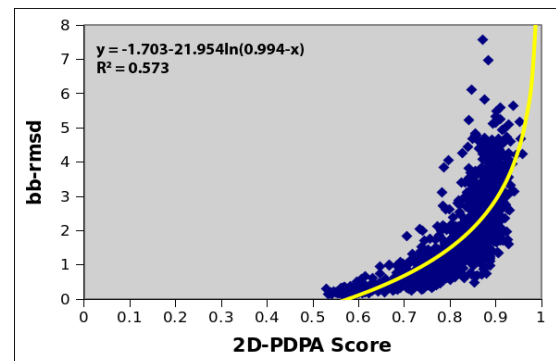


Figure 5. Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1GB1 with 15 (28%) of RDC data removed from RDC sets.

3.3 Experiment 3

Experiment 3 used protein 1A1Z, which is an α -helical structure, and calculated the RDC data with $\pm 1\text{Hz}$ of uniformly added error. Order tensors were obtained from REDCAT for each RDC set (Table 3). Figure 6 shows the correlation between the bb-rmsd of the structures and the 2D-PDPA scores; least squares linear regression line and R^2 values are included.

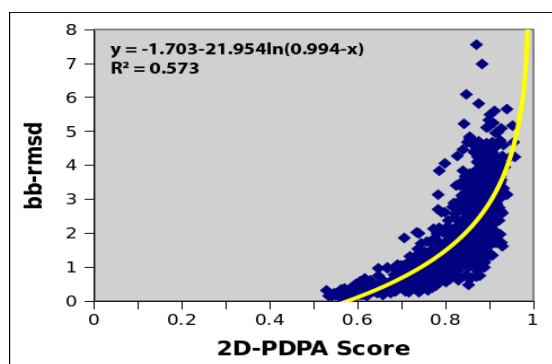


Figure 6. Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1A1Z with $\pm 1\text{Hz}$ error added.

3.4 Experiment 4

Experiment 4 used protein 1OUR and calculated the RDC data using $\pm 1\text{Hz}$ error to demonstrate noisy conditions. Order tensor matrices were obtained from REDCAT [21] for each RDC set (Table 3). Figure 7 shows the relationship between one thousand 2D-PDP structure scores and bb-rmsd with the Manhattan Block metric. The least squares regression line and the R^2 value are also shown in Figure 7.

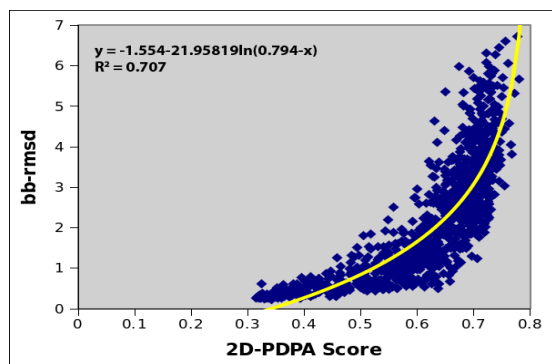


Figure 7. Calculated 2D-PDPA scores vs bb-rmsd using Manhattan Block metric for 1OUR with $\pm 1\text{Hz}$ error added.

4 Conclusion

2D-PDPA is a powerful method which can be utilized to identify homologous structures using only a minimal set of experimental data prior to a full structure determination protocol. Therefore, 2D-PDPA is a viable method for ascertaining a protein's structural novelty to within 3\AA , relative to the existing library of structures. The main contribution of our method demonstrates the correlation

between scored PDP and bb-rmsd of the corresponding structure. This also confirms the reliability of the 2D-PDPA identification and scoring, up to a threshold of 3\AA . To conduct our experiments we chose 3 structures representing three distinct CATH families. The experiment repeated for RDCs with no error and RDCs with error and missing data has confirmed 2D-PDPA's capability for pragmatic conditions. In all cases, the correlation between bb-rmsd and calculated PDP scores are clear. In the case of noisy RDCs data, our experiments show a slight shift of 2D-PDPA's score, yet a correlation is maintained. A-priori determination of score thresholds allows for interpretation and reliability of the 2D-PDPA's performance. The observed threshold of 3\AA also extends the use of the presented method to confirmation of computationally modeled structures. A hybrid approach of 2D-PDPA based selection of best computed structures can be envisioned, which allows for combined strengths of computational and experimental methods of structure determination while maintaining low cost.

5 Acknowledgements

This work was supported by NSF-Career grant MCB-0644195 from NSF to Dr. Homayoun Valafar. The high-performance computational environment used for this work was funded by NSF grant CNS-0708391.

6 References

- [1] Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D & Darnell J. Molecular Cell Biology, 4th edition. W.H. Freeman, 2000.
- [2] Hvidsten TR, Laegreid A, Kryshchuk A, Andersson G, Fidelis K & Komorowski J. A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS One* (2009) 4: p. p. e6266.
- [3] Skolnick J & Fetrow JS. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in Biotechnology* (2000) 18: pp. 34-39.
- [4] Baker D & Sali A. Protein structure prediction and structural genomics. *Science* (2001) 294: p. pp. 93-96.
- [5] Brenner SE, Chothia C, Hubbard TJ & Murzin AG. Understanding protein structure: using scop for fold interpretation. *Methods Enzymol* (1996) 266: p. pp. 635-643.
- [6] Chandonia J & Brenner SE. The impact of structural genomics: expectations and outcomes. *Science* (2006) 311: p. pp. 347-351.
- [7] Orengo CA, Todd AE & Thornton JM. From protein structure to function. *Current Opinion in Structural Biology* (1999) 9: pp. 374-382.
- [8] Sali A & Kuriyan J. Challenges at the frontiers of structural biology. *Trends in Biochemical Sciences* (1999) 24: p. M20-M24.
- [9] Service RF. A dearth of new folds. (2005) 307: pp. 1555-1555.

- [10] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE. The Protein Data Bank. *Nucleic Acids Res* (2000) **28**: pp. 235-242.
- [11] Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM & Orengo CA. Assigning genomic sequences to CATH. *Nucleic Acids Research* (2000) **28**: pp. 277-282.
- [12] Murzin AG, Brenner SE, Hubbard T & Chothia C. SCOP - A Structural Classification Of Proteins Database For The Investigation Of Sequences And Structures. *J Mol Biol* (1995) **247**: pp. 536-540.
- [13] Kopp J, Bordoli L, Battey JND, Kiefer F & Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* (2007) **69 Suppl 8**: p. pp. 38-56.
- [14] Murzin AG & Patthy L. Sequences and topology: From sequence to structure to function. *Curr Opin Struct Biol* (1999) **9**: p. p. 359-362.
JH & Valafar H. Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA). *J Magn Reson* (2008) **192**: pp. 60-68.
- [17] Saupe, A & Englert, G. Phys. Rev. Lett; High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules. *Phys. Rev. Lett* (1963) **11**: pp. 462-464.
- [18] Valafar H & Prestegard JH. Rapid classification of a protein fold family using a statistical analysis of dipolar couplings. *Bioinformatics* (2003) **19**: pp. 1549-1555.
- [19] Yandle R MR&VH. Using Residual Dipolar Coupling from two Alignment Media to Detect Structural Homology. *BIOCAMP* (2009) : p. pp. 90-95.
- [20] Mukhopadhyay R, Miao X, Shealy P & Valafar H. Efficient and accurate estimation of relative order tensors from lambda-maps. *J Magn Reson* (2009) **198**: pp. 236-247.
- [21] Valafar H & Prestegard J. REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* (2004) **167**: pp. 228-241.
- [22] Miao X, Mukhopadhyay R & Valafar H. Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application. *J Magn Reson* (2008) **194**: pp. 202-211.
- [23] Bansal S, Miao X, Adam M, Prestegard J & Valafar H. Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA). *J Magan Reson* (2008) **192**: pp. 60-68.

Mimicking Transcription Process to Recognise Promoters in E.coli

T.Sobha Rani

Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, India

Abstract—Promoter prediction is a computationally interesting and complex problem. Various groups have tried promoter prediction with different sequential and structural features of promoters. The structural aspects of DNA in promoter recognition are gaining popularity of late. First step in transcription process is the binding of RNA polymerase with the promoter. Here in this work, a preliminary study of interactions between RNA polymerase and specifically the binding sites within the promoter is carried out. Interaction values between RNA polymerase and DNA are used to identify the -35 and -10 binding sites in the promoter. A set of windows around these regions are extracted. Bi-gram features of these windows are used to test the validity of using such interactions in promoter recognition. Two types of encoding, Electron-ion interaction potential (EIIP) and amino acid-base pair interaction values are used to quantify the interaction between RNA polymerase and the promoter. Current results are comparable to earlier results obtained with n-grams. The experiments seem to point to a signal global in nature is much more efficient than local signal in promoter recognition. The results also confirm that the basic interactions between RNA polymerase and DNA (promoter) have the capability to identify the promoters in a whole genome.

Keywords: Classification ; EIIP encoding ; amino acid-base pair integration; machine learning

1. Introduction

Promoter prediction is complex and several groups of researchers have attempted to solve this problem by extracting different features which can be used to characterize the promoters. Some of the features that have been used for this task are position weight matrices [1], [2], [3], n-mers [4], [5], [6] which are statistical in nature. There are methods that have used DNA structural features such as enthalpy [7], thermal stability [8], stress induced duplex destabilization [9], roll-angle [7], base stacking energy [10] etc. Ponomarenko et al. have listed a wide variety of structural properties [11]. A wide range of classifiers such as neural networks [13], [1], SVM [12], hidden Markov model [14] and graph based induction [15] are also used.

Even though there is a huge amount of work done, the promoter prediction problem is far from being solved. The accuracy of predictions is not very high. In case of eukaryotes a group of promoters called GC rich promoters

are easier to predict than other promoters which are not GC rich. We want to investigate this problem from the point of view of the basic chemical interactions that arise between the RNA polymerase and the promoter irrespective of the nature of the promoters present in the genome. As a consequence, DNA-RNA polymerase interactions and bi-grams are used in the promoter identification in this work.

1.1 DNA-RNA Polymerase Interaction

In prokaryotes, the first step in transcription is the binding of RNA polymerase with the promoter. RNA polymerase is a large molecule consisting of five subunits α_1 , α_2 , β , β' and ω . In order to bind promoter-specific regions, the core enzyme requires another subunit, sigma (σ). The sigma factor greatly reduces the affinity of RNAP for nonspecific DNA while increasing specificity for certain promoter regions, depending on the sigma factor. This way, transcription is initiated at the right region. The complete holoenzyme therefore has 6 subunits: $\alpha_1\alpha_2\beta\beta'\omega\sigma$ (480 kDa). The structure of RNAP exhibits a groove with a length of 55 (5.5 nm) and a diameter of 25 (2.5 nm). This groove fits well the 20 (2 nm) double strand of DNA.

Promoter specific transcription on RNA polymerase is conferred by σ subunit. Based on sequence analysis these σ factors are divided into two broad classes σ -70 factors and σ -54 factors. Four highly conserved regions are identified by aligning σ 70 family of proteins [16], [17], [18]. Of these regions 2 and 4 are highly conserved and basic in nature and regions 1 and 3 exhibit low conservation and are acidic in nature. The secondary structures of regions 1 and 2 are predicted to be β -sheets with helices and regions 3 and 4 are predicted to be helical [19].

A series of studies revealed that sub-region 2.4 (located at the C-terminal end of region 2) interacts directly with promoter -10 hexamer elements, whilst sub-region 4.2 (located at the C-terminal end of region 4) interacts directly with promoter -35 hexamer elements. A number of studies using a variety of primary and alternative σ factors from E.coli and B.subtilis have identified residues of region 2.4 (a sub region of region 2) interacting with -10 hexamer and these interactions are depicted in figure 1 [20], [21], [22], [23]. Genetical analysis studies explain the interactions between the residues of RNA polymerase and nucleotides of -35 region in DNA [24], [25]. Figure 2 illustrates these interactions between the residues of σ 4.2 region and the -35 region of the promoter. Eventhough a lot of other interactions

are involved, only the interactions between RNA polymerase and the binding sites is considered here as a starting point.

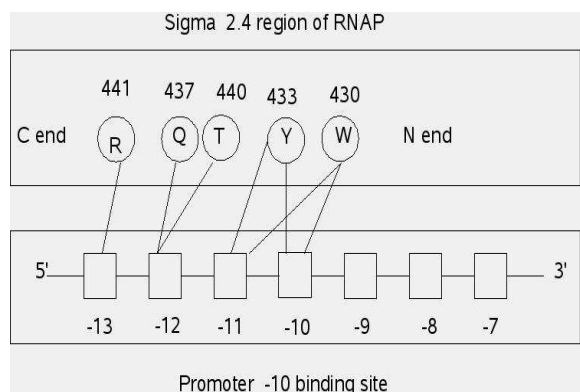


Figure 1: Pictorial depiction of the interactions between -10 binding site and amino acids of σ subunit.

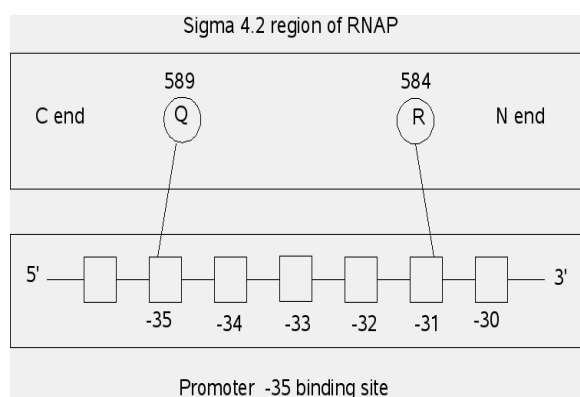


Figure 2: Pictorial depiction of the interactions between -35 binding site and amino acids of σ subunit.

A systematic study of n-grams in promoter prediction for $n = 2, 3, 4, 5$ [6] was carried out by us. We have obtained 68% promoter prediction accuracy for E.coli with $n = 3$. We got a very good prediction of promoters on forward-strand of E.coli taken from NCBI data base [6].

The main difference between the work that is being proposed in this paper and the work reported by Sobha et al. [6] is that in this paper emphasis is on identifying the binding sites through interaction between the DNA and RNA polymerase whereas in the later work it is just the occurrence of n-grams in the whole promoter without distinguishing the binding sites and non-binding sites.

2. Approach

A preliminary study of DNA-RNA polymerase interaction information in promoter recognition is performed by us [26]. We have attempted to compute the interaction through cross-correlation between promoter and RNA polymerase sigma subunit. We have not considered the three-dimensional

aspect of RNA polymerase then. Hence, the results of classification were not good for promoters. Here, in this paper we have tried to identify a subset of amino acids in RNA polymerase sigma subunit that takes part in the interaction between promoter and RNA polymerase. These appear mostly as part of α helix in $\sigma 2$ and $\sigma 4$ regions of σ subunit.

Interaction between RNA polymerase and promoter is quantified in two ways. One is by computing the cross-correlation between the DNA and RNA polymerase signals converted into numerical sequences. Second one is by considering the values obtained considering the interaction between amino acids and nucleotides. Cross-correlation between the residues of sigma subunits of RNA polymerase which interact with -10 and -35 hexamer regions are converted into numerical sequence using the EIIP values for the amino acid [27]. Similarly the nucleotides which take part in this interaction are also converted into numerical sequences using EIIP encoding [28]. Since we have no knowledge about the nucleotides which interact with the amino acids in a sigma subunit, we have followed the spacer scheme of Ma et al. [13]. They have considered a varying space of 15-21 bp (7 bp) between -35 and -10 regions and 3-11 bp (9 bp) between -10 region and TSS. In the same way, we have constructed our sigma subunit segment of length 80, consisting of zeroes except at the positions -35, -31 and -13, -12, -11, -10 positions with different spacings between them. This would result in a set of 63 combinations. Maximum correlation coefficient of the 63 combinations is chosen to fix the spacers between -35 and -10 regions and also between -10 region and TSS. Once the spacers are fixed, we can identify the binding regions in a promoter. Windows of certain length are extracted around these binding sites. Bi-gram features of these windows are extracted as features for a multi-layer feed forward neural network to train and identify the promoters in a genome.

3. Methodology

E.coli promoter data set is used for experimentation. We consider sequences of length 80 bp with 60 base pairs upstream of the Transcription Start Site (TSS) and the rest downstream [12]. Positive data set consists of 669 promoter sequences of length 80 bp [12]. Negative data sets of Gordon et al. who have chosen these in a biologically meaningful way by taking sequence fragments outside the promoter region. They also have built negative data sets with 709 sequence fragments from coding region and 709 sequence segments from intergenic portions.

3.1 Feature Extraction

Features are extracted in two stages. In the first stage, DNA-RNA polymerase interaction is used. In the second stage, windows around binding sites are identified and bi-gram features of these are extracted. These features are used

as input for a multi-layer feed forward network to learn about promoter. Here, the promoter recognition is posed as a binary classification problem.

3.1.1 Step 1: Identification of binding sites using DNA-RNA polymerase interaction

From literature the amino acids that interact with -10 and -35 binding regions are identified. The residues that interact with -35 binding site are taken from the work of Campbell et al. [25]. Similarly the residues that participate in interaction with -10 binding site are taken from the work of Malhotra et al. [23]. These are depicted in Figure 1.

In order to compute the interaction between DNA and RNA polymerase, we have chosen the cross-correlation as the means. Cross-correlation between the two can be computed by converting both DNA and RNA polymerase sequences into numerical sequences. In this method the amino acid residues and nucleotides are encoded into numerical format using EIIP values [27], [28]. EIIP encoding is chosen since it can be used to encode both amino acids and nucleotides. Table 1 lists the EIIP values of the relevant amino acids and nucleotides.

Table 1: EIIP values for amino acids [27] and nucleotides [28].

Amino acid	EIIP	Nucleotide	EIIP
Tyrosine(Y)	0.0516	A	0.1260
Tryptophan(W)	0.0548	T	0.1335
Glutamine(Q)	0.0761	G	0.0806
Threonine(T)	0.0941	C	0.1340
Arginine (R)	0.0959		

Another way of encoding using values provided by Mandel et al. [29]. Mandel et al. [29] have analyzed protein-dna complexes to extract all non-homologous pairs of amino acid-base pairs that are in close contact. A quantitative measure of the likelihood of the interaction between each pair of amino acid and base is computed. A score can be computed by summing up the individual measures of amino acid-base pairs assuming additivity in their contributions to binding. This score can be used a measure of the compatibility between the protein and its dna target. Table 2 lists these amino acid-base pair interaction values.

Table 2: Amino acid-base pair interaction values [29].

	G	A	T	C
Trp	-1.96	-3.93	-1.96	-3.93
Tyr	-2.87	-2.87	0.54	0.13
Gln	-0.09	1.16	0.31	-3.09
Thr	-3.46	-0.06	-0.06	-1.16
Arg	2.74	0.34	1.25	-3.93

In order to obtain the interaction between promoter and residues in the sigma subunit, $similar(j)$ defined in 1 is computed.

$$similar(j) = \min(\sum abs(s_1 - s_2)); j = 1, 2, 3, \dots, 63 \quad (1)$$

Here $j=1$ denotes the spacing between -35 and -10 regions as 15, and spacing between -10 and TSS as 3 bp. Similarly, $j=2$ denotes spacing between -35 and -10 regions as 16, and spacing between -10 and TSS as 3 bp and so on. Final $j=63$ denotes spacing between -35 and -10 regions as 21, and spacing between -10 and TSS as 11 bp. These are listed in Table 3. $similar(j)$ will be close to zero if s_1 and s_2 are close to each other. That is, if we suppose s_1 as promoter and s_2 as the set of residues that interact with the promoter, when they are compatible with each other, then $similar(j)$ values will also be zero.

Table 3: Spacing between -35 and -10 binding sites (SP35) and -10 and TSS (SP10) for different j values.

j	SP35 (bp)	SP10 (bp)
1	15	3
2	16	3
..
7	21	3
8	15	4
..
63	21	11

The values of the 63 combinations for various spacings between -35 and -10 regions and -10 and TSS can be treated as the compatibility between the sigma subunit and promoter binding regions. Of the 63 combinations obtained from the above calculations, the highest score is considered to arrive at the spacers between binding sites. Fixing up the spacers, binding sites can be identified.

3.1.2 Step2: Bi-gram feature extraction

Regions with high information content are selected, specifically 17 positions around the -35 binding site and 11 positions around the -10 binding site and 7 positions around the transcription start site are extracted. A bi-gram is a combination of contiguous two letters. DNA consist of four bases and therefore 16 bigrams (AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC) are formed [6]. For each window 16 bigrams are computed. In total 48 bi-gram features are obtained for all the three windows. Two types of experiments are performed. One in which the original 48 bigram features are given as input features to the multi-layer feed-forward (MLFF) perceptron. The output of the neural network is ≥ 0.5 if the given sequence is predicted as a promoter. Second one in which the bi-gram feature values for each of the windows are combined together into 16 bigram features. Simulations are done using SNNS package [30].

3.2 Training and Testing

Bi-gram features extracted from the windows around the binding sites and TSS are used as input features for the MLFF neural network. We have carried out 5-fold cross-validation procedure in which the total data set is divided into 5 parts. In each fold, 1 part will form the test set while the remaining four will be used for training. Precision (Pr), Sensitivity (Sn) and Specificity (Sp) are used as measures of classification performance. Specificity is the proportion of the negative test sequences that are correctly classified and sensitivity is the proportion of the positive test sequences that are correctly classified. Precision is the proportion of the correctly classified sequences of the entire test data set.

3.3 Extension to Whole Genome Promoter Prediction

The real test for any promoter recognition is its ability to identify promoters in a whole genome. Towards this end, we have used *section1* and *section3* of E.coli. Total genome of E.coli is divided into 400 sections. Out of these sections two sections *section1* and *section3* are chosen to extend the promoter recognition algorithm. These are chosen for the purpose of comparison with the results obtained using n-gram features [6]. A sliding window of 80 bp is used to segment these sections into segments of size 80 bp. We consider a sliding window of length 80 extracting segments from the start of the DNA sequence considered, that is, 1–80, 2–81, 3–82 and so on. These are represented as the bi-gram feature vectors which are used by the neural network classifier. Each of the segments gets classified as promoter (P) or non-promoter (NP). If a segment $m - (m + 79)$ is classified as a promoter, then the nucleotide m is annotated as P and if it is classified as non-promoter then m is annotated as NP . This process of annotation is continued for the entire sequence to get a sequence of P 's and NP 's. We propose that if a contiguous segment of length more than a certain threshold has all P 's then we annotate that region as promoter region otherwise as non-promoter region. For the verification purpose we have considered the *section1* and *section3* of E.coli [31]. It also denotes the set of promoters present in these segments.

4. Discussion

Table 4 shows the average of 5-fold cross validation results for both 48 bigram features as well as 16 bigram features extracted using $similar(j)$ (refer to equation 1). Sensitivity and specificity using 48 bigram-features are close to what was obtained using bi-grams for the entire promoter [6] than the ones obtained with 16-bigram features. But *section1* and *section3* results using 48 features and 16 features extracted from the windows using $similar(j)$ values present a different scenario. False-positives, that is non-promoters identified as promoters are much less with 16

features compared to 48 features. This fact is evident from figures 3 and 4. In these figures X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window. Only output greater than 0.5 is considered as a promoter. It is not only an output greater than 0.5 that is essential we also need to have a stretch of continuous ones over a threshold value of 20-25 is required to annotate the stretch of base pairs as a promoter. In this context, we could identify a clear stretch of positives with 16 features compared to 48 features. These results are comparable to what was obtained using n-grams [6]. We have also carried out one more experiment wherein the nucleotides in the windows are straightaway used as input features to the neural network after converting them into numerical values using EIIP codes. This experiment results are not as good as the results obtained with bi-grams.

Table 4: Classification results using $similar(j)$ features. SetA: Bi-grams from each window used separately and SetB: Bi-grams from each window combined together.

Features	Number	Pr	Sp	Sn
SetA	48	79.15	86.92	62.76
SetB	16	76.47	86.53	55.14

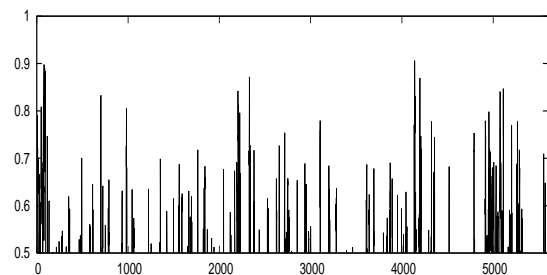


Figure 3: *section1* of E.coli tested with 48 bi-gram features. X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window.

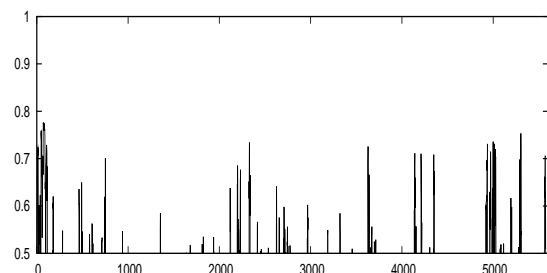


Figure 4: *section1* of E.coli tested with 16 bi-gram features. X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window.

As in the case of n-grams extracted from the whole pro-

motor, we have obtained satisfactory results with the features extracted from the interaction between promoter and certain residues of sigma subunit. Through this interaction, we have extracted the binding sites and the windows around the binding sites and TSS. Whole genome promoter prediction results using 16 bi-gram features in fact assures that the binding sites that are extracted are of relevance since we obtain similar results as in the case of bi-grams extracted from the whole promoter. Results obtained with 16 features compared to 48 features indicates that a global signal is much more powerful than a local signal.

Annotation of same *section1* and *section3* of E.coli obtained with features extracted from interactions derived by Mandel et al. is done and the results of the annotation for *section1* is shown in Figure 5. Similar results are predicted by these features also. But, the stretch of promoters is about 15-25 only. None of these results are predicting as well as 3-grams [6].

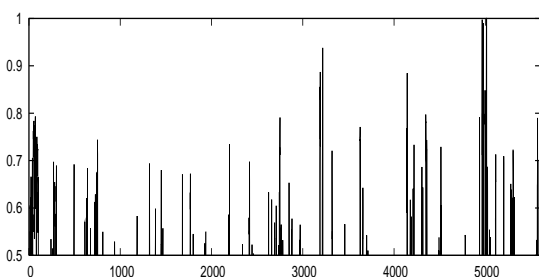


Figure 5: *section1* of E.coli tested with 16 bi-gram features extracted from interactions proposed by Mandel et al. X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window.

Moreover, frequency analysis of the binding sites extracted using EIIP and Mandel values, indicates a marked bias towards certain bases in positions -35 and -31 and also -13, -12, -11 and -10. Table 5 and Table 6 represent the frequency of occurrence each base pair at each position in -35 binding site. The consensus at the -35 binding site of each position using EIIP gives a closer similarity to the general consensus TTGACA observed in literature. Since EIIP values for T and C are close, that could explain some distribution between T and C at -35 position and -31. In case of interactions obtained through Mandel's values, since, Glutamine favours A in comparison to the others, we could observe, a bias towards A. So is Arginine at position -31 which favours G.

The annotation results of these sections of E.coli compare with that of results obtained using 3-grams in the earlier work [6]. The distinct aspect of this work is the identification of the binding sites through the interactions between RNA polymerase and the binding sites of the promoter. If the binding sites were not identified correctly, the resulting bi-grams around the windows would not lead to the correct

Table 5: Frequency of occurrence of bases in -35 binding site for promoters using Mandel et al. interaction values

	-35	-34	-33	-32	-31	-30
A	0.503	0.282	0.298	0.230	0.018	0.242
T	0.381	0.285	0.367	0.430	0.228	0.317
G	0.113	0.175	0.089	0.094	0.753	0.145
C	0.0014	0.257	0.243	0.245	0.000	0.296

Table 6: Frequency of occurrence of bases in -35 binding site for promoters using EIIP encoding for interaction

	-35	-34	-33	-32	-31	-30
A	0.051	0.291	0.260	0.341	0.052	0.294
T	0.412	0.375	0.288	0.323	0.374	0.269
G	0.000	0.224	0.309	0.224	0.000	0.291
C	0.537	0.109	0.142	0.112	0.574	0.145

identification of the promoters. This was verified through experiments where the binding sites were incorrectly identified and the resulting classification accuracy of promoters was down to 45%. Only in the case of correct identification, we get to identify binding sites correctly hence can identify promoters much better.

5. Conclusions

Promoter recognition is attempted using the interactions between RNA polymerase and promoter. Experiments with similarity and cross correlation between RNA polymerase and promoter are tried. Experiments used to obtain similarity and cross-correlation using EIIP values show that a global signal (Figure 4) is rather more effective than a local signal (Figure 3). Eventhough the test data results indicate a higher sensitivity value, generalization capability of the 16 features is better than 48 features. The results also point to the fact that similarity measure between the signal is more efficient in promoter recognition. Interactions derived using amino acid-base pairs are not as powerful as the signal derived using EIIP values. The analysis of frequency distribution of bases in the binding sites shows that EIIP values have a distribution closer to the predicted consensus sequence compared to amino acid-base pair interactions. Additivity of interactions is assumed in these cases. Whether there is a stable conformation possible, with a lower interaction value is to be investigated further. And also addition of more interactions to the set will increase the accuracy much further. A committee machine using these different features can be designed to annotate a segment as a promoter or a non-promoter based on voting. In addition the same sections are used for annotation with GLIMMER and Genemark packages which annotate the coding regions in the given DNA segment. Most of our promoters identified are occurring upstream of these coding regions giving credence to our annotation scheme.

Same arguments can be extended for eukaryotes in which

lot of transcription binding factors (TBP) bind to a promoter before a RNA polymerase is summoned. In this case, the interaction between TBPs and promoter can be modeled through the protein-promoter binding interactions and can be used to identify the promoters.

The main aim of the work is to prove the efficacy of the interaction between the RNA polymerase and the DNA in identifying the promoters in a whole genome. Even though the n-gram features are being used, it is very important to correctly identify the binding site regions through the interaction between DNA and RNA polymerase to get good accuracies. Hence, the main assumption that the interaction between DNA and RNA polymerase is proven to be very useful in promoter identification.

Acknowledgements

I would like to acknowledge the help of my student Mr. Naresh in the survey of literature and the financial assistance provided by my university, University of Hyderabad.

References

- [1] V. Bajic, A. Chong, S. Seah, V. Brusic, "An Intelligent System for Vertebrate Promoter Recognition," *IEEE Intelligent Systems*, pp. 64-70, 2002.
- [2] Q. Li, H. Lina, "The recognition and prediction of $\sigma 70$ promoters in Escherichia coli K-12," *Journal of Theoretical Biology*, vol. 242, pp. 135-141, 2006.
- [3] Y. Huang, C. Wang, "Integration of knowledge discovery and artificial intelligence approaches for promoter recognition in DNA sequences," *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)*, 2005.
- [4] F. Leu, N. Lo, L. Yang, "Predicting Vertebrate Promoters with Homogeneous Cluster Computing," *Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, 2005, p. 143.
- [5] H. Ji, D. Xinbin, Z. Xuechun, "A systematic computational approach for transcription factor target gene prediction," *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology CIBCB '06*, 2006, p1.
- [6] T. Sobha Rani, S. Bapi Raju, "Analysis of n-gram based promoter recognition methods and application to whole genome promoter prediction," *In Silico Biology*, vol. 9, pp. s1-s16, 2009.
- [7] I. Deyneko, E. Alexander, B. Helmut, G. Kauer, "Signal-theoretical DNA similarity measure revealing unexpected similarities of E. coli promoters," *In Silico Biology*, vol. 5, online 2005.
- [8] K. Aditi, B. Manju, "A novel method for prokaryotic promoter prediction based on DNA stability," *BMC Bioinformatics*, vol. 6, doi: 10.1186/1471-2105-6-1, 2005.
- [9] H. Wang, C. Benham, "Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress," *BMC Bioinformatics*, vol. 7, doi: 10.1186/1471-2105-7-248, 2006.
- [10] T. Abeel, P.R. Yvan Saeys, Y.V. de Peer, "ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles," *Bioinformatics*, vol. 24, pp. i24-i31, 2008.
- [11] J. Ponomarenko, M. Ponomarenko, A. Frolov, D. Vorobyev, G. Overton, N. Kolchanov, "Conformational and physicochemical DNA features specific for transcription factor binding sites," *Bioinformatics*, vol. 15, pp. 654-668, 1999.
- [12] L. Gordon, A.Y. Chervonenkis, A.J. Gammerman, I.A. Shahmurradov, V. Solovyev, "Sequence alignment kernel for recognition of promoter regions," *Bioinformatics*, vol. 19, pp. 1964-1971, 2003.
- [13] Q. Ma, J.T.L. Wang, D. Shasha, C.H. Wu, "DNA sequence classification via an expectation maximization algorithm and neural networks: a case study," *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, Special Issue on Knowledge Management*, vol. 31, pp. 468-475, 2001.
- [14] A. Pedersen, P. Baldi, Y. Chauvin, S. Brunak, "The biology of eukaryotic promoter prediction - a review," *Computers & Chemistry*, vol. 23, pp. 191-207, 1999.
- [15] T. Matsuda, H. Motoda, T. Washio, "Graph based induction and its applications," *Advanced Engineering Informatics*, vol. 16, pp. 135-143, 2002.
- [16] M. Gribskov, R.R. Burgess, "Sigma factors from E. coli, B. subtilis, phase SPO1, and phage T4 are homologous proteins," *Nucleic Acids Res.*, vol. 14, pp. 6745-6763, 1986.
- [17] J.D. Helmann, M.J. Chamberlin, "Structure and function of bacterial sigma factors," *Ann. Rev. Biochem.*, vol. 57, pp. 839-872, 1988.
- [18] J.A. Jaehing, "Sigma factor relatives in eukaryotes," *Science*, vol. 253, pp. 859, 1991.
- [19] M.J. Zvelebil, G.J. Barton, W.R. Taylor, M.J.E. Sternberg, "Prediction of protein secondary structure and active sites using the alignment of homologous sequences," *J.Mol. Biol.*, vol. 195, pp. 957-961, 1987.
- [20] D.A. Siegel, J.C. Hu, W.A. Walter, C.A. Gross, "Altered promoter recognition by mutant forms of the sigma 70 subunit of Escherichia coli RNA polymerase," *J. Mol. Biol.*, vol. 206, pp. 591-603, 1989.
- [21] T.J. Kenney, K. York, P. Youngman, C.P.J. Moran, "Genetic evidence that RNA polymerase associated with A factor uses a sporulation-specific promoter in Bacillus subtilis," *Proc. Natl. Acad. Sci. USA*, vol. 86, pp. 9109-9113, 1989.
- [22] C. Waldburger, T. Gardella, R. Wong, M.M. Susskind, "Changes in conserved region 2 of Escherichia coli sigma 70 affecting promoter recognition," *J. Mol. Biol.*, vol. 215, pp. 267-276, 1990.
- [23] A. Malhotra, E. Severinova, S.A. Darst, "Crystal structure of a $\sigma 70$ subunit fragment from E. coli RNA polymerase," *Cell*, vol. 87, pp. 127-136, 1996.
- [24] T. Gardella, T. Moyle, M.M. Susskind, "A mutant Escherichia coli sigma 70 subunit of RNA polymerase with altered promoter specificity," *J. Mol. Biol.*, vol. 206, pp. 579-590, 1989.
- [25] E.A. Campbell, O. Muzzin, M. Chlenov, J.L. Sun, C.A. Olson, O. Weinman, M.L. Trester-Zedlitz, S.A. Darst, "Structure of the Bacterial RNA Polymerase Promoter Specificity σ Subunit," *Molecular Cell*, vol. 9, pp. 527-539, 2002.
- [26] T. Sobha Rani and S. Bapi Raju, "E.coli promoter recognition through wavelets," *Proceeding of BioComp'08, 2008 International conference on Bioinformatics and Computational Biology*, 2008, p. 256.
- [27] H.T. Chafia, F. Qian, I. Cosic, "Protein sequence comparison based on the wavelet transform approach," *Protein Engineering*, vol. 15, pp. 193-203, 2002.
- [28] I. Cosic, *Macromolecular Bioactivity: Is It Resonant Interaction Between Macromolecules? - Theory and Applications*, *IEEE Transactions on Biomedical Engineering*, vol. 41, pp. 1101-1114, 1994.
- [29] Y. Mandel-Gutfreund, H. Margalit, "Quantitative parameters for amino acid-DNA base interaction: implications for prediction of protein-DNA binding sites," *Nucleic Acids Research*, vol. 26, pp. 2306-2312, 1998.
- [30] Stuttgart Neural Network Simulator. Available: <http://www-ra.informatik.uni-tuebingen.de/SNNS/>
- [31] NCBI Viewer Available: <http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=1786181>

Bio-Medical Data Integration Based on MetaQuerier Architecture

¹Khondker Shajadul Hasan, ²Munirul Islam, ³M Samiullah Chowdhury, ³Eusuf Abdullah Mim, and ³Naieem Khan

¹School of Computer Science, University of Oklahoma, 110 W. Boyd St., Norman, OK 73019, USA
shajadul@ou.edu

²Department of Computer Science, Wayne State University, Detroit, MI 48202, USA
munirul@wayne.edu

³Department of EECS, North South University, Bashundhara, Dhaka, Bangladesh, {samiullah.chowdhury, eusuf9001, ntkhan}@yahoo.com

Abstract - *The emergence of a large number of bio medical datasets on the Internet has resulted in the need for flexible and efficient approaches to integrate information from multiple bio medical data sources and services. Thus data are scattered in different web sites and web databases. User struggling hard and for them it is extremely difficult for them to find accurate data from the web efficiently. In this paper, we tried to present our approach to establish an architecture which will automatically generate web data integration, optimize the composition, and execute the required output efficiently. While data integration techniques have been applied to the bio medical data domain, the focus has been on answering specific user queries. Thus we have found the indication towards large scale data integration. So the issue arises for which data integration architecture can be used. There are so many proposed large scale data integration architecture are available. Among all of them we designed our paper based on the MetaQuerier architecture. It's large scale integration over web databases. MetaQuerier architecture has five basic processes which will be clarified in this paper briefly. We used this architecture to implement our bio medical data integration and try to generate a well structured output. Here our first task is to explore the MetaQuerier architecture and secondly we will explore the design in terms of bio medical data.*

Keywords: *MetaQuerier architecture, Data crawling, Source clustering, Schema etc.*

I. INTRODUCTION

Biologists are now faced with the problem of integrating information from multiple heterogeneous public sources with their own experimental data contained in individual sources. The selection of the sources to be considered is thus critically important. There is a compelling demand for the integration and exploitation of heterogeneous biomedical information for improved clinical practice, medical research, and personalized healthcare across the EU. The ultimate goal of the project is to provide uninhibited access to universal biomedical knowledge repositories, large-scale information-based biomedical research and training.

Now-a-days new treatments come about as a result of other, earlier discoveries. They are often unconnected to each other, and in various field. Sometimes the research was done for non-medical purpose and only by accident contributes to the field of medicine. Like the discoveries of *penicillin*. But now all the treatment has to be done through research. In the terms of Bio-Medical, we are considering the data from medical diseases, different kind of elements of human being and analysis of various medicine, the experiments and result obtained from them, analyzing zinc, proteins, bacteria and many more for advance research. For example, changes in genomic DNA, presence of various protein modification, mRNA and protein levels etc. The possibilities from bio-medical data integration are enormous. For example, the central tumor suppressor protein p53 provides a potential target for new anti-cancer drugs. By integrating the datasets from different laboratory various result of protein p53 like its characteristic, behavior, effect, mutations, etc. in one single database.

Data for bio-medical researches integrated from the web. Data can be stored on the WEB in different form. The data can be non-structured or semi structured in the web. Like plain text files, HTML text files, native XML. Data might be found in online libraries, catalogues, etc. Databases in the research repositories are like genome databases, scientific databases, environmental databases, etc. There might be web services, semantic web, and knowledge base system. There are Ontologies, which are structurally and semantically research domain description with associated data. The following charts are important for our paper. Here we have shown some biomedical data sets. There are unlimited bio medical datasets. Here our main focus actually to introduce the fact that bio medical data are needed to be integrated and also it is possible to be integrated. So in order to clarify our paper we focus on specific data sets which are in the group of Protein.

Database systems except from the web must have to inter-operate, cooperate and coordinate with each other. These data have to shared, exchanged and ultimately integrated. In this regard our target to see the sights of the

Table 1: Contains an example of Bio-medical data and attributes. [2]

Concept	Source
Protein	HSPProtein(\$id, name, location, function, sequence, pubmedid)
	MMPProtein(\$id, name, location, function, sequence, pubmedid)
	MembraneProtein(\$id, name, taxonid, function, sequence, pubmedid)
	TransducerProtein(\$id, name, taxonid, location, sequence, pubmedid)
	DIPPProtein(\$id, name, function, location, taxonid)
	ProteinLocations(\$id, \$name, location)
Protein-Interactions	HSPProteinInteractions(\$fromid, toid, source, verified)
	MMPProteinInteractions(\$fromid, toid, source, verified)

MetaQuerier architecture. MetaQuerier architecture is one of the most recent data integration architecture. The following figure will give some basic idea about the data flow and data manipulation inside the MetaQuerier mechanism.

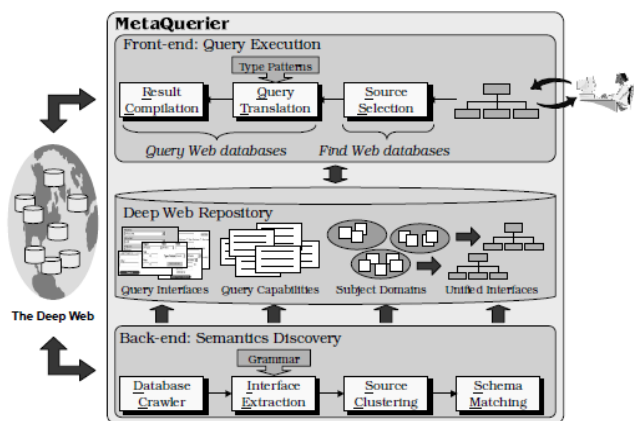


Figure 1: MetaQuerier: System Architecture. [1]

The above figure 1 is the complete scenario of the MetaQuerier architecture. Based on the MetaQuerier mechanism, our goal has two directions– First, to make the deep Web systematically accessible, it will help users find online databases useful for their queries. Second, to make the deep Web uniformly usable, it will help users query online databases [1]. In this paper we first describe the work flow of MetaQuerier engine and then explore the bio medical data inside MetaQuerier. At the end, Integration of related works will clarify the necessity of MetaQuerier in bio medical data integration. Actually the necessity of an efficient data integration engine arises due to the extremely huge volume of queryable databases. One side the data collection are dynamic another side they are non systemic.

To our knowledge, our goal of integration at a large scale has largely remained unexplored. The MetaQuerier engine actually integrates data from the web. One of the critical issues is that data are not predefined. Data are flourishing in every moment and datasets are getting larger. Since datasets are not predefined data discovery become dynamic. If one user search for different types of protein for example, for the next search he or she will not get the same datasets from the web resources. So this is a challenge while data searching. That’s why we need data crawler. Another major issue which steps in more complexity situation is that data are needed to be integrated on the fly. The engine will work at a time [1].

II. RELATED WORK

Our complete work has two basic direction and stands. One part deals with MetaQuerier engine and other part focus on bio medical data and how MetaQuerier engine will manipulate bio medical data. What do we understand about bio medical data? There exist a large number of bio medical datasets on the web in various formats. There is a need for flexible and efficient approaches to integrate information from these datasets. Unlike other domains, the bio medical domain has hold web standards, such as XML and web services. There exists a large number of bio medical data sources that are either accessible as web services or provide data using XML. For the bio medical data sources that provide their data as semi-structured web or text documents, we can use wrapper- based techniques to access the data.

For example, when a user queries the UniProt1 website for details of a protein, the user provides a uniprotid and gets back the information about the protein. The emergence of the large number of information providing services has highlighted the need for a framework to integrate information from the available data sources and services. In this paper, we describe our approach to automatically compose integration procedure to create new information-providing the MetaQuerier engine.

When the MetaQuerier receives a request based on bio medical data to create a new web service, it generates a parameterized integration that accepts the values of the input parameters such as protein name or its id and then retrieves and integrates information from relevant web pages, and returns the results to the user. The parameterized integration procedure is then hosted as a new web page what is known as data crawling. The discoveries of the pages according to the user requirement are dynamic and they are absolutely unsorted. This is the key challenge in composing web data for a new web based on the fly integration.

To further clarify these consider the example shown in Figure 2. We have access to three web services where each providing protein information for different organisms. We would like to create a new web service that accepts the name of an organism and the id of a protein and returns the protein information from the relevant web service. Given specific values of the input parameters, traditional data integration systems can decide which web service should be queried. However, without knowing the values of the parameters, the traditional integration systems would generate a procedure that requires querying all three web services for each request.

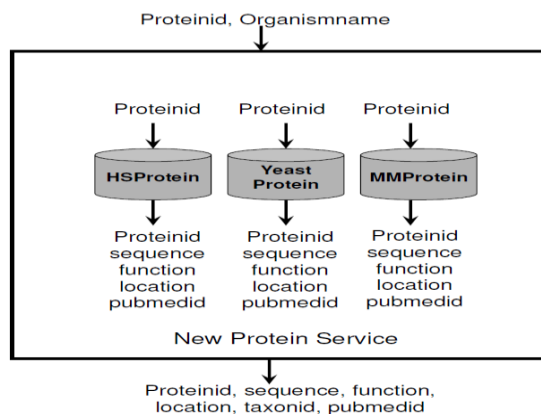


Figure 2: Protein Information. [2]

The key contribution of our approach is to extend the existing techniques to generate parameterized integration technique that can answer requests with different sets of values for the input parameters.

Now the key issue arises when it's needed to optimize the web data and in order to reduce the deep web data for optimizing the number of user request sent to existing data sources we need the help of MetaQuerier. Thus when data optimization with user satisfaction will be needed MetaQuerier will be in action. The existing optimization techniques means the MetaQuerier will utilize the searchable user query to filter out unnecessary source requests and/or reorder the joins to produce more efficient ordered web data. However, as we show with a detailed example later in the paper, the MetaQuerier techniques are enough when we apply them to the task of data integration.

Next section will describe each of the processes of the MetaQuerier techniques in more detail, show how they can be applied to the bio medical domain. We begin by describing a motivating example that we use throughout the paper to provide a detailed explanation of various concepts. Next, we discuss how existing data integration techniques can be extended to model web sources as data sources and reformulate web data creation requests into parameterized integrated data.

In MetaQuerier there are five basic processes. First is Database crawler, second Interface Extraction, third source clustering, fourth Schema Matching and fifth Result Compilation. [1]

There are three parts of the complete MetaQuerier. Front end, back end and deep web repository. But before understand the parts of MetaQuerier we need to understand its starting and action point. As it handles large volume of data, first, such integration is dynamic: Since sources are blooming and evolving on the Web, they cannot be statically configured for integration. Second, it is absolutely unsorted: Since queries are submitted by users for different needs, they will each interact with different sources. Thus, toward the large-scale integration, the MetaQuerier must achieve dual requirements—Dynamics discovery and on the fly integration. To our knowledge, MetaQuerier is the first one to present the overall system issues of building large scale integration. Next section we will elaborate about MetaQuerier architecture. [1]

III. SYSTEM DESCRIPTIONS

On the way towards bio-medical data integration, we have accounted the large scale of data and these data can be found on the web database. But data are not predefined. It means, we are doing a deep web searching as user's request but sources are not in a single domain, which we are calling "dynamic discoveries" and "on-the-fly" integration. Based on the processes used in MetaQuerier, we have structured the idea of bio-medical data integration. We will now describe the whole system of the application. [1]

The MetaQuerier was developed for large scale integration. In its way of integration, it basically search and collect the database on the web, extract the required data from the database and gather it into its own database and show users the output as requested. To understand the system easily, we have divided the whole process into five major parts. On sequence, Data Crawling, Interface extraction,

Source clustering, Schema Matching, Query Translation, Source Selection, and Result Compilation. Data Crawling, Interface extraction, Source Clustering, Schema Matching all these processes work at the back-end. Query Translation, Source selection and Result Compilation all these work on front-end. In this interim paper we are giving a short brief for ease of understand bio-medical data integration. [1]

A. Data Crawling

Collect data from enormous web environment is the main part of the challenge that we face while data integration. So actually we need data crawler. There is a difference between data crawler and web crawler. Existing and available search engines are efficient for necessary site searching. They search based on the root pages and also check user keywords as interface keyword [3]. Here if we go in that process we will be in the messy situation of managing terabyte of data. So in the MetaQuerier our task to find web pages that are exclusively important for us including the databases involved with these sites. Thus MetaQuerier design data crawling in two different segments to face the challenge of dynamic discovery. The first segment named site crawler and second segment is shallow crawler [1]. Together these two segments MetaQuerier named the data crawler as site based crawler. For site crawler the efficiency of query interfaces are very important.

Query interfaces are important because based on the interface keyword crawler will filter web sites. It will minimize unsorted and unnecessary data. Suppose the following interface can use to find more appropriate and mandatory data while search web sites. The more keyword will be used from the interface the more data filtering will be in action. Since we are not focusing on how efficient interface can be designed here, we just discuss the important of query interfaces and its necessity for data integration. In this paper we are actually trying to explore one of the important uses of data integration in the field of bio medical research. Site base crawler will go through the root pages and will identify IP addresses and shallow crawler will follow these IP addresses and will search web servers which will be found from site crawler [1].

Figure 3: Sample Search Interface. [9]

B. Interface Extraction

Interface extraction basically extracts the required data from the query interfaces. Query interface sometimes share similar or common query patterns but sometimes it shares different query patterns. In case of different query patterns the problem arises due to some hidden information or attributes. Hidden attributes are not visual on interface that's why its extraction normally out of interaction at the beginning. Thus for the hypothesized syntax, in metaQuerier the determined structure are rationalize by asserting the creation of query interfaces as guided by some hypothetical syntax [5]. Therefore handle this hypothetical syntax effectively creates new problem. So it's needed to be visualizing as a visual language whose composition conforms to a hidden non-prescribed, grammar. In this case MetaQuerier solve the problem in terms of parsing the visual language. Here the MetaQuerier approach is to introduce a parsing paradigm by assuming that there exists hidden syntax to describe the layout and semantic of query interfaces. Specifically, we develop the subsystem IE as a visual language parser, given a query interface in HTML format; IE tokenizes the page, parses the tokens, and then merges potentially multiple parse trees, to finally generate the query capability. [1,4]

Finally after parsing Interface Extraction basically extracts query capabilities from the query interfaces. The semantically related labels and elements of a search interface are viewed as logical attributes, though they are scattered in the html text or into the database without formal definitions. Therefore, attributes have to be identified by grouping associated labels and elements. Moreover, beyond the labels and elements, a significant amount of semantic/meta information for attributes exists on the query interfaces [5].

For example, in figure 4, "invention date" implies the Attribute is semantically a date data type, and its two elements are used to specify a range query condition with different roles in specifying the condition. Unlike the conventional database schemas, such semantic/meta information is "hidden" from computers and not formally defined on query interfaces. As such, the "hidden" information about each attribute needs to be revealed and defined to enrich the schema matching. [5]

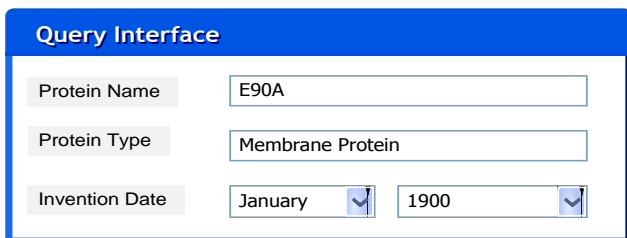


Figure 4: Sample query interface.

C. Source Clustering

Before move on to Schema Matching we need to understand Source Clustering. Source clustering collaborates with source selection which works in front end. These two processes help schema matching to get actual scenario. After determining query capabilities based on query interfaces source clustering sorted data as mediated process which provides data towards schema process. Here the second

challenge of MetaQuerier after the dynamic discovery, the on the fly integration comes in action. Source Cluster actually clusters sources according to subject domain. Going towards data integration, we need clustering sources by their query capabilities, specifically, given a set of query capabilities representing structured sources, our task is thus to construct a hierarchy of clusters, each representing an object domain of "structurally-consistent" sources. Thus we need to cluster the query interfaces into subject domains.

Domain elements and constraint elements have the following characteristics:

- Textboxes cannot be used for constraint elements.
- Radio buttons or checkboxes or selection lists may appear as constraint elements.
- An attribute consists of a single element cannot have constraint elements.
- An attribute consisting of only radio buttons or checkboxes does not have constraint elements.

Based on these characteristics, a simple two-step method has to be used to differentiate domain elements and constraint elements. First of all, we have to identify the attributes that contain only one element or whose elements are all radio buttons, or checkboxes or textboxes. Such attributes are considered to have only domain elements. Then an Element Classifier is needed to process other attributes that may contain both domain elements and constraint elements. [1,9]

D. Schema Matching

The schema of a database system is its structure described in a formal language supported by the database management system. In a relational database, the schema defines the tables, the fields in each table, and the relationships between fields and tables. Schemas are generally stored in a data dictionary. Although a schema is defined in text database language, the term is often used to refer to a graphical depiction of the database structure. Schema matching is the process of identifying that two objects are semantically related while mapping refers to the transformations between the objects.

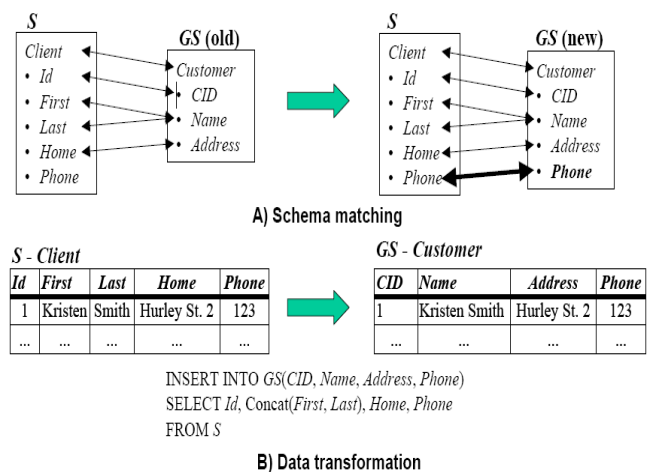


Figure 5: Schema matching for data integration.

In data integration process, schema matching find out the semantic domain values among the attributes, which we have found through query interfaces. In MetaQuerier, the

Bio-Medical Keywords: Exact Phrase

Production Date: All dates

Production Year: after: before:

Concentration (grams): between: and:

Platform: All Windows Macintosh Universal

Figure 8: Example of element relations.

When an attribute has multiple associated elements, we shall classify them into two types: *domain elements* and *constraint elements* because they usually play different roles in specifying a query. Domain elements are used to specify domain values for the attribute while constraint elements enforce some constraints to domain elements. For example, element “Exact phrase” is a constraint element while the textbox following Title keywords is a domain element.

Consider two interfaces. One interface contains an attribute protein Production date, and another interface contains an attribute protein production year, and they should be matched in terms of their semantics. But we cannot match them by only using names because they do not have exactly the same attribute name. However, if we can identify that the elements of both attributes are of *range type*, it would increase the confidence of matching them. When a user specifies a query on the global attribute Title of a MetaQuerier interface, during the query translation the query value should be mapped to the domain element of Bio-Medical keywords instead of the constraint element “Exact phrase”. [3]

B. Interface Extraction

Labels and elements are the basic components of a search interface, but it is insufficient to just extract individual labels and elements because many applications rely on the *logical attributes* formed by related labels and elements. In order to extract logical attributes, it is essential to determine the semantic associations of labels and elements. However, there are no explicit definitions of such associations in the HTML text of the search interface. We observe that labels and elements that represent the same attribute have a certain layout pattern and are usually close to each other and that in most cases they have some similar information in common. On the basis of this, we develop a three-step approach to tackle the problem of automatic interface extraction or in other words attribute extraction.

1) Extracting Individual Labels and Elements

This is the first step of our automatic attribute extraction method. Given a search interface, the extraction starts with its “<FORM>” tag. Each element itself contains its values (if available). Four types of input elements are considered: *textbox*, *selection list*, and *checkbox* and *radio button*. When a row delimiter like “
”, “<P>” or “</TR>” is encountered, a ‘|’ is appended to the Interface expression. This process continues until the “</FORM>” tag is encountered. In this process, some irrelevant texts may be included in the INTERFACE EXPRESSION even though some efforts are made to identify and discard them. [8]

2) Identifying the Names of Exclusive Attributes

Exclusive attributes are actually the ones whose names may appear as *values* in some elements, such as a group of *radio buttons* or a *selection list*. Correctly recognizing such attributes automatically is difficult because they do not appear on search interfaces as descriptive texts. [5]

Top screenshot: Bio-Medical Keyword Founder Producer Product Bar Code [Search]

Bottom screenshot: Bio-Medical Keyword [Search]
 Bio-Medical Keyword
 Founder
 Producer
 Product Bar Code

Figure 9: Examples of exclusive attributes. [9]

Exclusive attributes appear frequently on real Web search interfaces. A significant flaw of existing approaches for interface extraction is that they do not extract exclusive attributes.

The names of exclusive attributes are often the *most commonly used attribute names* of a domain. The basic idea is that we consider multiple interfaces in the same domain at the same time rather than separately. Then we use the extracted labels from all search interfaces of the same domain to construct a vocabulary for the domain. Finally we use the vocabulary to automatically identify and extract the names of exclusive attributes.

3) Grouping Labels and Elements

This step is to group the labels and elements that semantically correspond to the same attribute, and to find the appropriate attribute label/name for each group. For example, label “Bio-Medical Keywords”, the textbox, the three radio buttons and their values below the textbox all belong to the same attribute and this step aims to group them together and identify label “Bio-Medical Keywords” as the name of the attribute.

C. Source Clustering

Going towards MetaQuerier, we need clustering sources by their *query schemas*, i.e., attributes in their query interfaces.

Table 2: Translation Rules.

r ₁	[category; contain; \$s] → emit: [source; all; \$s]
r ₂	[name; contain; \$t] → emit: [name; contain; \$t]
r ₃	[concentration range; between; \$s, \$t] → \$p = ChooseClosestNum(\$s), emit: [concentration; less than; \$p]
r ₄	[onlooker’s age; between; \$s] → \$r = ChooseClosestRange(\$s), emit: [age; between; \$r]

For instance, for the advanced query interface of amazon.com, the query schema is specifically, given a set of query schemas representing structured sources, our task is thus to construct a *hierarchy* of clusters, each representing an object domain of “structurally-homogeneous” sources [1]. Apparently, we are focusing on Bio-Medical data. We explain a particular method of source clustering, which is quite efficient in terms of domain attributes.

1) Deriving information from Attributes

In our proposed interface schema model, we recommend four types of information for each attribute are defined: *domain type*, *value type*, *default value* and *unit*. These meta-data are only for *domain elements* of each attribute.

Domain type: Domain type indicates how many distinct values can be used for an attribute for queries. Four domain types are defined in our model: *range*, *infinite* and *Boolean*. [9]

Value type: Each attribute on a search interface has its own semantic value type even though all input values are treated as text values to be sent to Web databases through HTTP. [9]

Default value: Default values in many cases indicate some semantics of the attributes. A default value may occur in a selection list, a group of radio buttons and a group of checkboxes. It is always marked as “checked” or “selected” in the HTML text of search forms. Therefore, it is easy to identify default values. [9]

Unit: A unit defines the meaning of an attribute value (e.g., *kilogram* is a unit for *weight*). Different sites may use different units for values of the same attributes. For example, one search interface may use “Milligrams” as the unit of its Concentration attribute, while another may use “Liters” for its Concentration attribute. [9]

2) Translation Rules

Firstly, we have to consider another term named, query mediation. Query mediation works have been mainly focusing on mediating queries across multiple sources and thus abstract the problem as a paradigm of answering query using views. In particular, they assume each source has a wrapper, which encapsulates the tasks of extracting query capability, schema matching and constraint mapping for that source. The main focus of query mediation is thus on how to decompose a user query into sub-queries across multiple sources. In contrast, we have to focus on query translation between two sources other than mediating queries across multiple sources. In particular, we are dealing with the mapping of constraint heterogeneity. For our scenario of large scale integration, we have to on-the-fly translated queries and thus need the following mapping techniques.

Secondly, Schema mapping aims at translating a set of data values from one source to another one, according to given matching. Therefore, schema mapping only concerns about the equality relation between different schemas, based upon which data is converted. In particular, no constraint heterogeneity is considered in schema mapping. In contrast, constraint mapping focuses on translating specific queries other than the data values.

3) Discussion

We have proposed a generic type-based search-driven translation framework, which is well suited for the requirements of the on-the-fly constraint mapping among

large scale data sources and our concern here is mainly focused on Bio-Medical Data.

V. CONCLUDING DISCUSSION

This paper contains the core proposal we made through MetaQuerier architecture. Actually, the issue over here is that bio-medical data integration is an example of data integration. There are so many proposed data integration process. Our target is to deploy MetaQuerier as efficient data integration architecture and show one of its implementation. We proposed that we can use data integration for bio medical data or in the field of bio informatics. Here one part mainly gives idea about the MetaQuerier architecture and its sub processes and how each of the processes work. Although it's not focused how we can improve this MetaQuerier, we considered MetaQuerier is one most efficient data integration engine/design.

Inside the subsystem of metaquerier there are some conceptual changes of many common things to improve the efficiency of handling extremely huge and unsorted data. The three basic process of back-end were discussed elaborately. Other two processes are just briefly discussed. Our future work to make real time implementation based on some specific requirement and based on some ongoing bio medical research.

REFERENCES

- [1] Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. *Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web*.
- [2] Snehal Thakkar, Jos'e Luis Ambite, Craig A. Knoblock. *Composing, Optimizing, and Executing Plans for Bioinformatics Web Services*. In September 2, 2005.
- [3] Gautam Pant, Padmini Srinivasan, and Filippo Menczer. *Crawling the Web*.
- [4] Ping Wu, Ji-Rong Wen, Huan Liu, Wei-Ying Ma. *Query Selection Techniques for Efficient Crawling of Structured Web Sources*.
- [5] Hai He, Weiyi Meng, Clement Yu, Zonghuan Wu. *WISE-Integrator: A System for Extracting and Integrating Complex Web Search Interfaces of the Deep Web*.
- [6] Z. Zhang, B. He, and K. C.-C. Chang. On-the-fly constraint mapping across web query interfaces. In *Proceedings of the VLDB Workshop on Information Integration on the Web (VLDB-IIWeb'04)*, 2004.
- [7] Bin He and Kevin Chen-Chuan Chang. *Automatic Complex Schema Matching Across Web Query Interfaces: A Correlation Mining Approach*.
- [8] Chengyong Yang, Erliang Zeng, Tao Li, and Giri Narasimhan. *A Knowledge-Driven Method to Evaluate Multi-Source Clustering*.
- [9] Hai He, Weiyi Meng, Clement Yu, Zonghuan Wu. *Automatic Extraction of Web Search Interfaces for Interface Schema Integration*.

Expression Network Analysis of Abiotic Stress Responsive Myb in Rice

Shuchi Smita^{1,2}, Amit Katiyar^{1,2}, Dev Mani Pandey², Viswanathan Chinnusamy³, Kailash Chander Bansal^{1*1}

¹National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute Campus, New Delhi-110012, India

²Department of Biotechnology, B.I.T. Mesra, Ranchi, Jharkhand, India

³Department of Botany and Plant Sciences, 2150 Batchelor Hall, University of California, Riverside, USA

Abstract - In post-genomic era, bioinformatic tools allow us to explore and reconstruct the precise gene interaction network. To deduce the function of uncharacterized gene, genetic network by co-regulatory analysis from an expression data is a foremost approach. In this study, we report comprehensive identification of co-expressed MYB gene modules in rice. MYB transcription factor family is involved in phenylpropanoid and flavonoid biosynthesis and various other metabolic and developmental processes. By a reiterative database exploration, 249 potential OsMYB genes were retrieved. Computational analysis has shown the presence of several other functional domains including WD domain, G-beta repeat, response regulator receiver domain, BTB/POZ domain, SWIRM/Zinc finger domain and many more. Several studies have pointed out their involvement in a range of biological processes, revealing that a large number of MYB genes are transcriptionally regulated under conditions of biotic and/or abiotic stress. To investigate the existence of MYB co-regulatory network, a whole genome MYB expression study was carried out in rice. We identified the existence of co-expression clusters comprising phylogenetically related MYB genes, suggesting that specific sets of MYB genes might act in co-regulatory network. Thus, the co-expression networks identified in this study illustrate gene cooperation pathways that have not been identified by classical genetic.

Keywords: MYB gene, clusters analysis, *Oryza Sativa*

1 Introduction

These are Plant growth and development are regulated by the coordinated expression of thousand of genes. To infer the function of uncharacterized genes, coexpression analysis of gene-to-gene is a useful approach. Regulation of gene expression is highly complex process, influenced by genotype environment interactions. The huge biological information available publically forms a foundation for system biology study nowadays [2]. System networks are often analyzed using visualization and analysis of network to deduce gene function, pathway components and links between and genes [1]. Network can be analyzed by direct and module based methods as in graph [8]. On the basis of gene-to-gene

correlation coefficient derived from microarray hybridization data, cluster-based analysis give the idea of co-expressed gene

or connections between genes that respond simultaneously to various stimuli [7]. For network study, expression profiling data are seems as highly useful resource. Microarray gene expression data is analyzed by a variety of bioinformatics techniques. In addition to commonly studied gene-specific expression patterns, gene expression analysis can be used to elucidate module and system-level organization of the transcriptome. Gene clustering method for module detection based on similarity of expression levels in different set of condition (gene co-expression networks) were used in many studies [9]. Several clustering algorithms have been developed for this purpose. Here, we report comprehensive identification of coexpression gene modules of MYB genes in rice. Several studies have indicated that MYB significantly involved in stress induced responses in *Arabidopsis thaliana* and other plants also. Several studies have pointed out their involvement in a range of biological processes, revealing that a large number of MYB genes are transcriptionally regulated under biotic and/or abiotic stresses [6]. In computational biology, use of network has greatly changed the analytical ability of researcher. In the present studies an attempt has been made to study this relationship of MYB genes in rice under abiotic stresses conditions. In our study, analysis of the MYB genes expression helped us to understand the cellular process where they involved, their interaction with other genes and their products.

2 Materials and Methods

2.1 MYB domain identification and Phylogenetic Analysis

Myb domain was retrieved by searching for PFAM ID PF00249 as a query in Rice at TIGR (<http://rice.plantbiology.msu.edu/>). Only the longest one was saved, when more than one alternative splicing sequence was found for the same locus. Phylogenetic tree for MYB proteins were constructed by iTOL (<http://itol.embl.de/>) to know the conserved pattern between rice MYB genes.

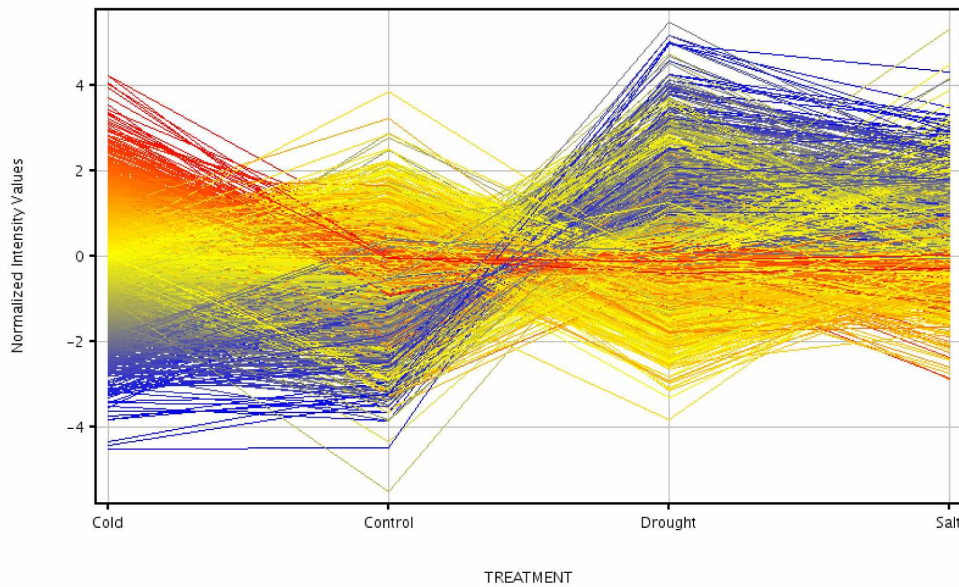


Figure 1. Profile plot for the differentially expressed genes found in our study

2.2 Expression Profiling of *MYB* and cluster analysis

To analyze the genome-wide expression profiles of rice *OsMYB* genes, microarray analysis was carried out using Affymetrix rice whole genome array. Expression data of *MYB* expression under abiotic stress were extracted from result of 12 hybridization experiment GSE6901 retrieved from GEO Database. .CEL files were downloaded and subjected to Genespring GX 10 (Agilent Technologies Inc, Santa Clara CA) and normalized with the PLIER16 algorithm (3) for further analysis. Obtained expression value were log₂ transformed, probes having two fold up - down regulation were taken. Hierarchical clustering was performed by average linkage and Euclidean distance algorithm using GeneSpring GX 10.

3 Results

3.1 Identification of *MYB* genes and Phylogenetic analysis

By a reiterative database exploration 249 potential *OsMYB* domains in rice were retrieved. Non-redundant dataset for *MYB* genes in rice genome were used as input for further analysis. Computational analysis of 249 identified *MYB* has shown the presence of several other functional domains including WD domain, G-beta repeat, response regulator receiver domain, BTB/POZ domain, SWIRM/Zinc

finger domain and many more. Phylogenetic analysis performed with the Maximum Likelihood method using all 249 proteins containing a single or double *MYB* domain, divided the genes into 3 main phylogenetic groups. Other subgroups and smaller clades were identified within each group, based upon bootstrap values.

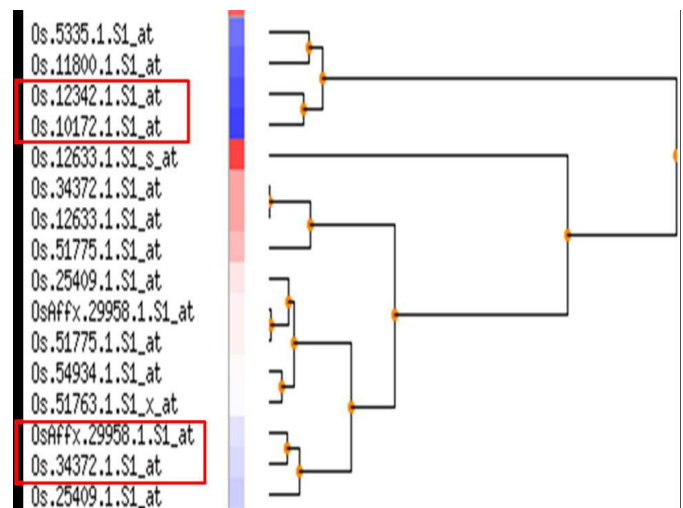


Figure 2. Clustering of upregulated *OsMYB* genes. Red boxes highlight the presence of co-expressed *MYB* gene clusters.

Table 1. Some up regulated *MYB* genes in three abiotic stress conditions.

Probes IDs (Cold stress)	Locus ID	2 Fold Upregulation
Os.57355.1.S1_x_at	LOC_Os01g58550	4.7924976
Os.5335.1.S1_at	LOC_Os04g49450	3.9795542
Os.11800.1.S1_at	LOC_Os01g50100	3.7111404
Os.12342.1.S1_at	LOC_Os12g04204	3.4641871
Os.10172.1.S1_at	LOC_Os02g41510	3.2356162
Probes IDs (Drought stress)	Locus ID	2 Fold Upregulation
Os.12633.1.S1_s_at	LOC_Os11g26790	7.9102745
Os.34372.1.S1_at	LOC_Os06g48300	6.630616
Os.51775.1.S1_at	LOC_Os12g05210	6.3568006
Os.25409.1.S1_at	LOC_Os06g45184	5.8992395
OsAffx.29958.1.S1_at	LOC_Os09g21180	5.7759666
Os.54934.1.S1_at	LOC_Os05g37060	5.6580715
Os.51763.1.S1_x_at	LOC_Os01g12690	5.57427
Probes IDs (Salt stress)	Locus ID	2 Fold Upregulation
Os.12633.1.S1_at	LOC_Os11g26790	6.636118
Os.51775.1.S1_at	LOC_Os12g05210	5.7613506
OsAffx.29958.1.S1_at	LOC_Os09g21180	5.308271
Os.34372.1.S1_at	LOC_Os06g48300	5.186286
Os.25409.1.S1_at	LOC_Os06g45184	5.0420284

3.2 Expression profiling and Clustering analysis result

Microarray data analysis was done by employing .CEL file to GeneSpring. Profile plot of all 2866 up and down regulated genes crossed all statistical test was made (Figure 1).

Differentially expressed genes were analyzed to extract *MYB* genes showing expression in drought, salt and cold stress conditions. We found 158 *MYB* genes out of 249 showing differential expression. Drought, salt and cold stresses upregulated (≥ 2 fold) 102, 72 and 16 *MYB* genes, respectively. Table 1 shows the up-regulated *MYB* genes found in our study. Clustering analysis of the upregulated *MYB* genes identified in this study was performed to pinpoint genes with similar expression profiles between different stress conditions. Understanding of this functional network structure of *MYB* genes, such as gene regulatory and biochemical networks, systems biology is the area that has to be explored

and the area that we believe to be the main stream in biological sciences in this century [4].

4 Conclusions

Our approach has identified co-regulated *MYB* gene networks that have potential role in abiotic stress response of rice. This will contribute to illustrate the functions of gene cooperation pathways not yet identified by classical genetic analyses. We defined the existence of *OsMYB* gene clusters comprising both phylogenetically related and unrelated genes that were significantly co-expressed, suggesting that specific sets of *MYB* genes might act in co-regulatory networks.

5 References

- [1] Daniele Merico, David Gfeller & Gary D Bader (2009) **How to visually interpret biological data using networks**, *nature biotechnology* volume 27 number 10 october 2009.
- [2] H. Kitano, (2002) **Systems biology: A brief overview**, *Science* 295, 1662–1664.
- [3] Hubbell E, Liu WM and Mei R (2002) **Robust estimators for expression analysis**. *Bioinformatics* 18:1585-1592.
- [4] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown and David Botstein, **Cluster analysis and display of genome-wide expression patterns**. Department of Genetics and, Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305
- [5] Michael TP, Mockler TC, Breton G, McEntee C, Byer A, et al. (2008) **Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules**. *PLoS Genet* 4(2): e14.
- [6] Nafees A. Khan and Sarvajeet Singh, I.K. (2008) **Abiotic Stress and Plant Responses** 300 p, 65, 81-89866-95-2,
- [7] Nicholas A. Heard, Christopher C. Holmes, David A. Stephens, David J. Hand and George Dimopoulos, **Bayesian Co-clustering of Anopheles Gene Expression Time Series: A Study of Immune Defense Responses to Multiple Experimental Challenges**.
- [8] Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function**. *Mol Syst Biol* 2007, **3**:88.
- [9] Wille A, Zimmermann P, Vranova E, Furholz A, Laule O, Bleuler S, Hennig L, Prelic A, von Rohr P, Thiele L, et al.: **Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana**. *Genome Biol* 2004, **5**(11):R92.

Simplifying Gene Expression Microarray Comparative Analysis.

Philip Church¹, Andrzej Goscinski¹, Adam Wong¹, and Christophe Lefevre²

¹School of Information Technology, Deakin University, Geelong, VIC, Australia

²Institute of Technology Research and Innovation, Deakin University, Geelong, VIC, Australia

Abstract - *Gene Expression Comparative Analysis allows bioinformatics researchers to discover the conserved or specific functional regulation of genes. This is achieved through comparisons between quantitative gene expression measurements obtained in different species on different platforms to address a particular biological system. Comparisons are made more difficult due to the need to map orthologous genes between species, pre-processing of data (normalization) and post-analysis (statistical and correlation analysis). In this paper we introduce a web-based software package called EXP-PAC which provides on line interfaces for database construction and query of data, and makes use of a high performance computing platform of computer clusters to run gene sequence mapping and normalization methods in parallel. Thus, EXP-PAC facilitates the integration of gene expression data for comparative analysis and the online sharing, retrieval and visualization of complex multi-specific and multi-platform gene expression results.*

Keywords: Gene Expression, Normalization, Clusters, Statistical Algorithms

1 Introduction

Comparative analysis is a fundamental tool in biology due to the influence of evolutionary and selective forces in shaping biological systems. Conservation among species greatly assists the detection and characterization of functional elements because important functional elements tend to be most conserved during evolution, whereas inter-species differences are likely indicators of biological adaptation. Comparative gene expression Analysis allows researchers to investigate the conserved or specific functional regulation of genes. Its basic principle is to group datasets based on gene evolutionary relatedness and isolate the components that behave in similar or different ways. Thus, comparing the regulation of genes in related organisms can assist the investigation of gene function. The microarray approach [1] is the most common method of collecting gene expression data currently being used in bioinformatics. More recently high throughput sequencing methodology is allowing an alternative approach for the estimation of gene expression. Data from microarray or sequencing experiments must be stored digitally using one of the many gene expression file formats before being analyzed

using statistical algorithms and analysis. Normalization is a key part of gene expression microarray analysis since unnatural variations can be introduced during the data collection and digitization process. Thus, this data must typically be corrected, standardized and cross-referenced before being compared and analyzed.

Here, we present a web based package called EXP-PAC using the PHP/MySQL paradigm for the collaborative, integrative and comparative analysis of related gene sequences and gene expression experiments. The implementation also makes use of high performance computing to assist the integration, and analysis, of multiple gene expression datasets with common normalization methods and the inter-specific mapping of reference sequence datasets. Although the mapping of gene sequences between species has been performed and made available for a number of model organisms, for example in the Homologene database (<http://www.ncbi.nlm.nih.gov/homologene>), our package enables the rapid integration of sequence data collected from uncommon animal species, for which orthologous genome maps may not yet be referenced in public databases, and addresses the need of researchers working on a more diverse set of organisms or specific biological systems. For example we have developed an implementation of EXP-PAC dedicated to the integration and comparative analysis of gene expression during lactation in the mammalian lineage, which is accessible through the International Milk Genomics Consortium Web Portal (www.imgconsortium.org).

2 Gene expression comparative analysis

Gene expression comparative analysis is usually performed in the following three steps as illustrated in Fig. 1.

1. Data is collected in a wet-lab using a gene expression platform (cDNA, high throughput sequencing, Microarrays, etc.).
2. Collected data is converted to a digital format and any un-natural variation is removed.
3. Data analysis is used to group together similar datasets to locate components putatively responsible for biological functions.

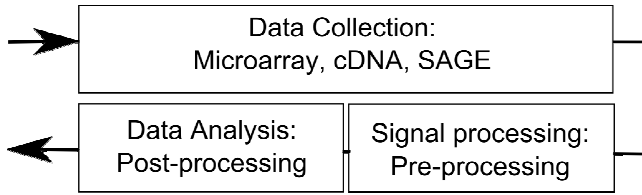


Fig. 1. The three stages of gene expression comparative analysis

2.1 Comparative transcriptome mapping

Before comparative gene expression analysis can proceed, the genes of one organism need to be cross-referenced to the related genes of another species. This is usually done through the identification of similarity by sequence similarity search algorithms such as BLAST [2]. Bi-directional reciprocal best hits often need to be identified and investigated for validation and identification of problematic gene family member assignment. Online access to sequences and maps greatly facilitates analysis of such gene family relationships and correct attribution of orthologous relationships for the construction of inter-genome maps. Although precompiled reference gene mapping data may be already available for a growing list of model organisms (Homologene), researchers working on non-model organism need to address the issue of cross-referencing genes. Our software package is built as an extension of an EST-PAC, a previously described package for the annotation of biological sequences [3]. Among other annotation tools, this package automates the management of

sequence similarity search algorithms and the analysis of results through a web interface using a database and job management system. More recently we have implemented a new version allowing the use of high performance computing platform to optimize the performance of sequence similarity searches which is an important addition for the execution of multiple full genome searches required for the construction of inter-specific gene mapping as the execution time for bi-directional reciprocal mapping growth quadratically with the number of species. Once systematic sequence comparisons have been done, additional scripts can be deployed to compile cross-reference tables. The Unigene (<http://www.ncbi.nlm.nih.gov/unigene>) database maintains representative sequences for the genes of model organisms and, since many commercial gene expression platforms reference Unigene identifiers, we typically use these sequence references when available and build cross-references for other species or gene expression platforms using available cDNA libraries or transcript sequences predicted from related genome sequences.

3 EXP-PAC

EXP-PAC is a web-based system developed for the comparative analysis of gene expression. The EXP-PAC system combines the features of EST-PAC [3] with an on-line tool extension for the storage, analysis and visualization of gene expression data providing interfaces to facilitate SQL query based post-analysis of results (see Fig. 2.). EST-PAC is a sequence analysis framework, which provides online

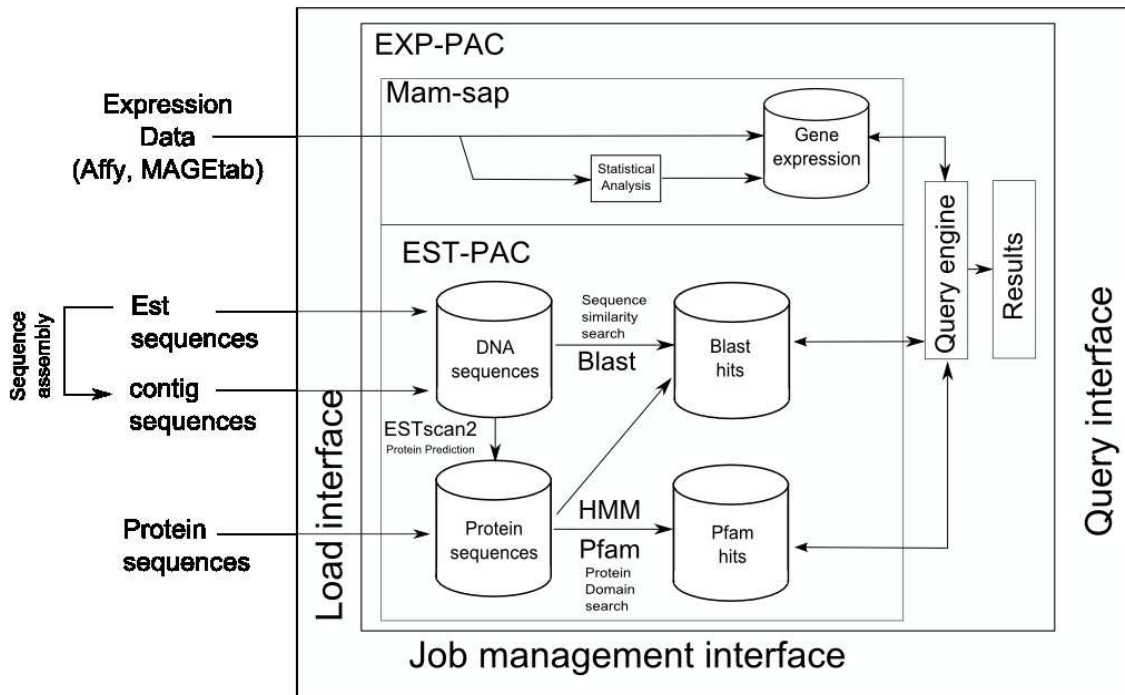


Fig. 2. The structure of the EXP-PAC system

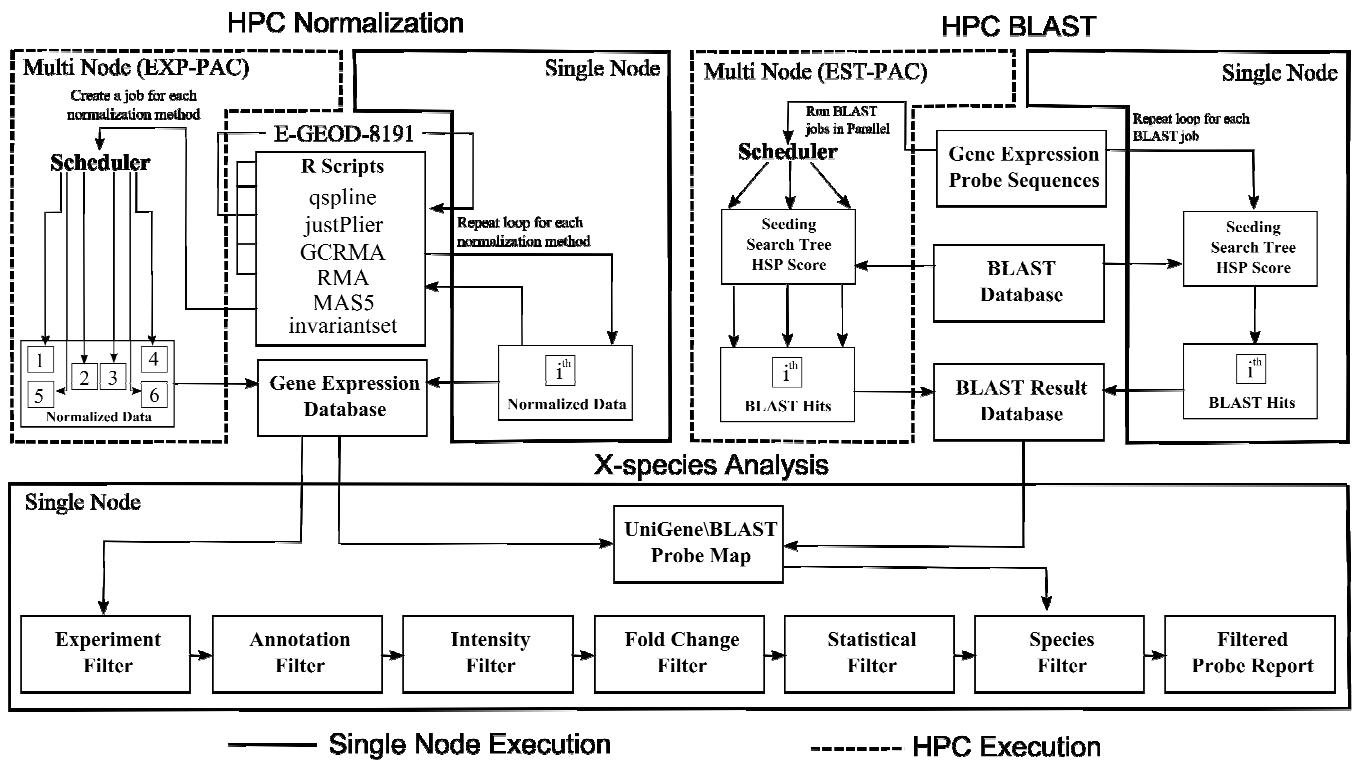


Fig. 3. EXP-PAC gene expression analysis workflow

interfaces for the storage of sequence data, the secure management of sequence annotation programs through embedded tools and, interfaces for the retrieval of sequence annotations. A new embedded high performance computing version of EST-PAC (EST-PAC^{HPC}) allows the cross referencing of transcriptome sequence catalogs through sequence similarity searches with the BAST program (companion paper in BIOCOMP11).

3.1 Annotation, data export and file sharing

The system allows the uploading of extendible gene annotation file formats associated with different gene expression platforms. This is necessary because gene annotation formats vary between gene expression platforms and data depositories. In addition, all data files uploaded in the system are archived and can be retrieved and downloaded from the interface, allowing data sharing and traceability.

3.2 Gene expression data upload and analysis

EXP-PAC provides users with the ability to upload a number of gene expression file formats (raw microarray data, SOFT [4], MAGE-tab [5], etc.) that may be available from download in gene expression databases [4, 6] or generated in the lab. Affymetrix microarray data files (also called CEL files) can be uploaded and automatically normalized with the R statistical scripting language using different established normalization methods for the Bioconductor package;

including RMA [7], MAS5 [8], GCRMA [9] and PLIER [10] (Fig. 3.). EXP-PAC supports normalization through a distributed platform which uses the Sun Grid Engine [11] in order to speed up microarray data management and analysis for this common platform. By specifying the location of a bash script supported scheduler, normalization methods can be distributed over multiple nodes reducing the time taken for the normalization process. Other types of datasets can be normalized independently. Raw and normalized data can be uploaded and compared. Results from statistical analysis, obtained for example in the specialized Bioconductor package for R, can also be uploaded from tab-delimited files. Meta-data can be edited to group samples and adjust graphic display and color. Gene expression, associated gene annotation and statistical data can then be queried using an interface dynamically generated from the uploaded data. In addition, through creation of a sequence to probe ID map, it is possible for a user to perform comparisons on multiple species or experiments, retrieving the expression of likely orthologous genes (identified in the EST-PAC sequence similarity database) throughout a set of experiments in related species.

3.3 Query interface

EXP-PAC provides users with a web interface through which gene expression data can be queried (Fig. 4.). A number of gene expression filtering methods are provided including; fold change, intensity levels, group average, probe ID and

The screenshot displays the EXP-PAC query interface, divided into two main sections:

- Search in data series:**
 - Search term: `Mouse_ROM_KO_GSE16629`
 - Annotation: `GPL1261-3958`
 - Keyword search: find `csn` in: `id`
 - Probe set ID list search: (empty field)
 - Intensity filter: intensity of `any` > `0`
 - Fold change: `GREATEST` / `LEAST` is > than `5` fold
 - Order by: decreasing intensity of `WT_7W_V_GSM417493`
- Species experiment selection:**
 - Uni-Gene Table: `BlastOutput_hit(Local)`
 - Maximum amount of hits (1-99): `5`
 - E-value: `1`
 - Buttons: `Select all`, `Deselect all`, `Toggle select`
 - Species list:
 - Bos taurus
 - Homo sapiens
 - Mouse
 - Mus musculus
 - Rattus norvegicus

Fig. 4. EXP-PAC query interface

annotation. Returned probes can be ordered by selected probe intensity; displayed in descending order. Graphs are also produced to visualize the gene expression levels of each probe. A query builder tool allows users to create more complicated queries through generic interfaces that map to the SQL language. Using this tool, users create a database view by specifying tables and columns from list boxes. Created database views can be filtered using alphabetical and numeric values and operators. The results from created SQL queries can be saved or exported as a comma delimited text file. Visualization tools for investigating the distribution of gene expression data are provided to validate the normalization process. Users may also retrieve gene expression data across different species and experiments using pre-compiled reference maps of related probes and genes.

4 Conclusions

In this paper, we have presented an on-line framework for gene expression research. Compared to available gene expression software packages, EXP-PAC is unique in that it provides a method for the integration of cross-species gene expression experiments allowing comparative analysis and a method to perform high performance computing for reference sequence mapping and some common normalization methods. Most importantly, the EXP-PAC software package provides researchers with a simple way to manage and analyse gene expression and sequence data. SQL based analysis allow users to perform broad searches of stored datasets. In addition it is easy to integrate R scripts into the EXP-PAC system, allowing support for new and specialized methods and algorithms for gene expression or sequence analysis. Thus, EXP-PAC enables the development of analysis strategies integrating multiple experimental platforms in different species and provides an online workbench for comparative gene expression analysis.

5 References

- [1] A. Brazma, *et al.*, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nature Genetics*, vol. 29, pp. 365-371, 2001.
- [2] S. F. Altschul, *et al.*, "Basic Local Alignment Search Tool," *Journal of Molecular Biology* vol. 215, p. 8, 1990.
- [3] Y. Strahm, *et al.*, "EST-PAC a web package for EST annotation and protein sequence prediction," *Source Code for Biology and Medicine*, vol. 1, p. 2, 2006.
- [4] T. Barrett, *et al.*, "NCBI GEO: mining millions of expression profiles--database and tools," *Nucl. Acids Res.*, vol. 33, pp. D562-566, January 1, 2005 2005.
- [5] T. Rayner, *et al.*, "A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB," *BMC Bioinformatics*, vol. 7, p. 489, 2006.
- [6] H. Parkinson, *et al.*, "ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Res*, vol. 37, pp. D868-72, Jan 2009.
- [7] R. A. Irizarry, *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostat*, vol. 4, pp. 249-264, April 1, 2003 2003.
- [8] E. Hubbell, *et al.*, "Robust estimators for expression analysis," *Bioinformatics*, vol. 18, pp. 1585-1592, December 1, 2002 2002.
- [9] Z. Wu, *et al.*, "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays," *Journal of the American Statistical Association*, vol. 99, p. 909, 2004.

- [10] I. Affymetrix. (2005, Technical note: guide to probe logarithmic intensity error (PLIER) estimation. Available: www.affymetrix.com/support/technical/technotes/plier_technote.pdf

- [11] W. Gentsch, "Sun Grid Engine: Towards Creating a Compute Power Grid," presented at the Proceedings of the 1st International Symposium on Cluster Computing and the Grid, 2001.

Comparison of Affymetrix expression array summarization methods for reproducibility and consistency across studies

Xiaoyang Ruan¹, Ourania Kosti², Rado Goldman², Hongfang Liu^{1*}

¹Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN

²Department of Oncology, Lombardi Comprehensive Cancer Center, Washington DC, USA

ABSTRACT

Affymetrix gene expression microarray is a popularly used platform for differential analysis. The analysis pipeline includes five steps: background correction, normalization, PM-only correction, and summarization, and differential analysis. Using publicly available microarray data, we compared the performance of five summarization methods: Median, Mean, Median Polish, Robust Linear Model, Li-Wong. Our evaluation criterion was reproducibility between studies designed to answer same scientific questions. Our analysis shows that mean value summarization gives smaller number of transcripts with inconsistent fold change direction while maintaining reproducibility comparable to competing complex methods. We conclude that after raw data has been preprocessed by the most popularly used pipeline (Robust Multiple Regression (RMA) background correction, quantile normalization, and PM-only correction), mean value summarization may convey a better representation of the true expression levels of target transcripts. The study suggests that the selection of bioinformatics algorithms needs to be application oriented. Sometimes simple initiative approach is probably better.

1 INTRODUCTION

Microarray technology, based on DNA hybridization to measure expression levels of mRNA or to detect Single Nucleotide Polymorphism (SNP) and copy number, has become an invaluable tool in biomedical research since the mid 1990s [1, 2]. One of the popular gene expression microarray platforms is Affymetrix where a target transcript is typically represented by a probe set consisting of 11-16 pairs of short oligos. Each pair consists of a perfect match (PM) and a mismatch (MM) oligo. The PM probe exactly matches the sequence of a particular standard genotype, while the MM differs in a single substitution in the central (13th base), intended to distinguish noise caused by non-specific hybridization. Transcript expression level is a summarization of the signal of individual probes in the corresponding probeset [3].

Data analysis for Affymetrix microarray generally consists of four preprocessing steps: background correction, normalization, PM correction and summarization. Background correction removes noise signals arising from many sources, such as non-specific binding, processing bias in wash stage or optical noise from the scanner. Normalization rescales intensity from

multiple chips to the same level so that gene expression levels on different chips can be comparable. PM correction controls for non-specific binding between probe and non-target sequences. The summarization step estimates the transcript expression level based on intensity measures of probes in the corresponding probe set.

There is a rich source of algorithms available to pre-process raw data from Affymetrix gene expression array. A relatively complete list of currently available preprocessing steps was tabulated by Irizarry R.A. [4] and Harr B. [5]. However, some of these have become obsolete given the accumulating evidence of poor performance. For example, MAS and subtractmm methods for PM correction were shown to consistently yield negative signals, which indicates that use of MM probes for detection of non-specific binding is unreliable [3, 6]. The widely used background correction method, robust multi-array average (RMA), relies solely on PM values [3]. GCRMA [7] was developed to take the effect of GC content on different probes into consideration. Bolstad *et al.* [8] compared several normalization methods and showed that quantile normalization has advantages in both speed and bias. Nowadays, the following pre-processing pipeline, RMA or GCRMA background correction—quantile normalization—pmonly correction—median polish or Li-Wong summarization, has become a standard [4, 5, 9].

The performance of various pre-processing methods is generally evaluated using spike-in and dilution data series [3, 4, 10, 11], MAQC data series [12-15], or based on the classification power of the number of differentially expressed genes obtained [16]. When using spike-in data, the differentially expressed genes are known in advance and assumed to be the true targets. However, this assumption is not safe in biological questions since it is unknown whether a gene expression difference reflects a true biological difference or not. This is especially important in microarray data analysis because of the high background noise and the various sources of variation (including but not limited to differences in probe labeling efficiency, RNA concentration, and hybridization efficiency). Moreover, many known comparison studies are based on a single dataset or specific controlling samples. This is potentially susceptible to the data structure of specific type (or group) of sample, or specific type (or batch) of microarray chip. MAQC [12] project spearheaded by FDA involved multi-platform and cross-lab comparison. However, it is actually based on fixed controlling RNA samples. Several existing publications on MAQC project did not discuss the performance of different summarization

*To whom correspondence should be addressed

methods in true data. It is hard to design a single trail that can take all potential confounding factors into consideration. In this paper, we compared the performance of different summarization pipelines by applying competing algorithms on microarray dataset pairs that are publicly available and can be used to answer the same scientific questions. We aimed to identify the method(s) that yield(s) consistent results between the pairs. In the following, we first present experimental design and evaluation metrics. We then discuss and conclude the study.

2 METHODS

2.1 Experiment Design

The experiment contains three levels of cross validation. The first level is different datasets pairs extracted from research results, which sheds a light on the possible performance difference caused by data structure of specific samples. The second level is different microarray platforms, which helps to avoid platform specific influence. The last level is the use of two different differential analysis algorithms, which takes the possible impact of algorithms specific effect on the competing methods into consideration.

Specifically, we identified three dataset-pairs (six datasets in total) from respective Affymetrix microarray platforms. The preprocessing pipeline is fixed to RMA—quantiles—pmonly, and only different summarization methods were compared. Five summarization algorithms primarily available in the latest Affymetrix built-in processing method [17], including median (Avgdiff), mean, median polish [10], robust linear model (RLM) [18, 19] and Li-Wong (dChip) [20], were compared for reproducibility between datasets extracted to address the same questions. Two other summarization methods, MAS [21] and playerout [22], were not discussed because they are less common these days (Table 1).

Two differential analysis algorithms (significance analysis of microarray (SAM) and CyberT) were implemented to the processed datasets to get the final result. SAM [23] estimates t statistics by adding a small constant s_0 to denominator to minimize coefficient of variation at low expression level. CyberT [24] uses regularized t -test in the Bayesian probabilistic framework. We also utilized GeneGo webtool to investigate the impact of competing methods on consistency of inferred biological pathways.

2.2 Datasets Pairs

Raw data were downloaded from the NCBI Gene Expression Omnibus (GEO) website. Sample annotations were parsed from the sample description files or the description column contained in each GSM sample. The three dataset pairs used in our analysis are summarized in Table 2 and details are presented below.

Pair a - GSE6956 [25] and GSE17356 [26] were designed to investigate biological factors that predispose African American (AA) men to prostate cancer when compared to European American (EA) men. GSE6956 contained 89 samples from prostate tumor tissue samples ($n=69$) and non-tumor tissue samples ($n=20$). We used the array data of 69 tumor samples for our study. Samples in GSE17356 are primary prostate cancer epithelial cell cultures ($n=27$).

Table 1. Summarization methods

Summarization Method	Author	Year	R Package	Discussed in Paper
Mean	-	-	-	yes
Median (Avgdiff)	Affymetrix	1999	expresso [17]	yes
MAS	Affymetrix	2002	expresso [17]	no
Median Polish	Irizarry RA et al	2003	expresso [17]	yes
Li-Wong	Li C, Wong WH	2001	expresso [17]	yes
playerout	Emmanuel. N.Lazaridis	2002	expresso [17]	no
Robust Linear Model (RLM)	Sboner A et al	2009	threestep (affyPLM)	yes

($n=27$). Group1 are prostate cancer samples isolated from AA men. Group2 are samples isolated from EA men. Fifteen genes were shown to be differentially expressed between AA and EA prostate cancer patients in both studies (See Table IV in paper reporting GSE17356 [26]). The common scientific question is “Which genes are differentially expressed between AA and EA men with prostate cancer”.

Pair b - GSE6532 [27] is a series with multiple data sources and platforms. It was designed in an effort to identify a gene classifier for predicting clinical prognosis of Tamoxifen-treated estrogen receptor positive (ER+) breast cancer patients. GSE6532 has a total of 741 samples (Supplementary Table 1). For comparative analysis we used 56 samples tested on U133A platform from the John Radcliffe Hospital (OXFT) and 81 samples from London, United Kingdom, Uppsala University Hospital (KIT). For both datasets, only ER+ breast cancer patients treated with Tamoxifen were used in our analysis. Group1 is defined as individuals with distant metastasis free survival (DMFS) ≤ 3 years, and Group2 are those with DMFS ≥ 5 years. The common scientific question is “In Tamoxifen-treated ER+ breast cancer patients, which genes are differentially expressed between individuals with DMFS ≤ 3 and ≥ 5 years”.

Pair c - GSE5460 [28] was designed to investigate the ability of global gene expression in primary breast tumors to predict receptor status, histological and other characteristics of the tumors. It contains 129 breast cancer samples from PLUS2 platform. GSE2109 is from expression project for oncology (expO) contributed by the International Genomics Consortium (IGC). A total of 2158 samples from roughly 100 tumor tissues are represented, of which 360 samples are from female breast cancer tissue. Since detailed phenotypic information is available for the two studies, we arbitrarily narrowed down sample phenotype to grade III ductal carcinoma to minimize the difference between pairing datasets. In the remaining part, ER+ samples were set as Group1 and estrogen receptor negative (ER-) samples as Group2. The common scientific question is “In grade III ductal carcinoma, which genes are differentially expressed between ER+ and ER- individuals”.

Table 2. Construction of comparing datasets

Datasets Pair	GSE number	Microarray Platform	Probe set Number	Group1 status	Group2 status	Group1 number	Group2 number
Pair a	GSE6956 GSE17356	HG-U133A 2.0	22277	AA ^a	EA ^a	34	35
						14	13
Pair b	GSE6532KIT GSE6532OXFT	HG-U133A	22283	ER+ & TAM DMFS<=3 ^b	ER+ & TAM DMFS>=5 ^b	21	35
						24	57
Pair c	GSE2109 GSE5460	HG-U133PLUS2 ^c	22283	ER-	ER+	65	48
						45	18

^a African American and European American men with prostate cancer

^b Tamoxifen (TAM) treated estrogen positive (ER+) breast cancer with distant metastasis free survival (DMFS) <=3 and >=5 years

^c Plus2 is basically a combination of HG-U133A and HG-U133B. Only HG-U133A probe sets were extracted out from Plus2 for the analysis due to a large number of non-gene targeting probe sets in HG-U133B part.

2.3 Summarization Algorithms

Five summarization algorithms (mean, median, median polish, robust linear model (RLM), and Li-Wong) were compared in the R environment. A complete list of the processing steps is listed in Table 3. Two differential analysis methods (SAM and CyberT) were used to get p value (use default option). FDR was obtained by applying *q-value* [29] function with default options. The relevant software package was downloaded from the BioConductor website.

Median

The median value of probes in a probe set was used to represent summary expression level. The median method gives result same as the result by avgdiff approach provided in affymetrix built-in processing method [17].

Mean

The mean value of probes in a probe set was used to represent summary expression level.

Median Polish

The model of median polish can be written as $T(PM_{ij}) = e_i + a_j + \varepsilon_{ij}$, where $T(PM_{ij})$ represents the measure after background correction, normalization, and log2 transformation of the PM intensity, e_i represents the log2 scale expression value found on array i , a_j represents the log scale affinity effects for probes j , and ε_{ij} represents random error. Implementation of median polish method is available in *expresso* function of R package *affy*.

Li-Wong

Li-Wong method has the following model: $MM_{ij} = v_j + \theta_i \alpha_j + \varepsilon_{ij}$, and $PM_{ij} = v_j + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon_{ij}$. Here PM_{ij} and MM_{ij} denote the PM and MM intensity values for array i and probe pair j for this gene, v_j is the baseline response of probe pair j due to nonspecific hybridization, θ_i is expression index for the gene in array i , α_j is the rate of increase of the MM response of probe pair j , ϕ_j is the additional rate of increase in the corresponding PM response, and ε_{ij} represents random error. The rates of increase are assumed to be nonnegative. The model for individual probe responses can be written as $y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}$. In the case of PM-only correction, $PM_{ij} - MM_{ij}$ is simply replaced by PM_{ij} . Implementation of Li-Wong method is available in *expresso* function of R package *affy*.

Robust Linear Model

The RLM method was developed by Hampel F.R.[19]. Use of RLM as summarization method was provided in *threestep* function of *affyPLM* R package (an extension of the base *affy* package).

2.4 Performance Metrics

Assume datasets A and B have N_a and N_b probe sets differentially expressed at significance level p_x . They share N_{both} probe sets in common. Among N_{both} probe sets, N_{diff} values have different fold change (FC) direction (i.e., the probe set is up-regulated in one dataset and down-regulated in another), and N_{same} have the same direction. The inconsistent FC proportion (IFP) and reproducibility are defined as following

Table 3. Processing flow for raw CEL file

	Background correction	Normalization	PM correction	Summarization	Differential Analysis Tool
Analysis Method	RMA ^a	Quantiles	PM-only	Median (avgdiff)	SAM CyberT
				Mean	
				Median Polish	
				Robust Linear Model (RLM)	
				Li-Wong (lw)	

^a For RLM method, RMA2 background correction method is used (RMA is not available in *threestep* function and it is not easy to reproduce).

Table 4. Number of consistent pathways with $p < 0.05$

Dataset pair	Platform	SAM					CyberT				
		median	mean	mp ^a	RLM ^b	lw ^c	median	mean	mp ^a	RLM ^b	lw ^c
GSE6956 VS. GSE17356	HG-U133A2	54	84	50	67	100	60	65	68	66	67
GSE6532OXFT VS. GSE6532KIT	HG-U133A	52	103	69	57	86	51	96	52	48	70
GSE2109 VS. GSE5460	HG-U133PLUS2 ^d	96	77	90	55	68	90	78	82	48	65

^aMedian Polish ^bRobust Linear Model ^cLi-Wong ^dOnly probe sets from HG-U133A used

$$IFP = \frac{N_{diff}}{N_{both}} \quad \text{Reproducibility} = \frac{N_{same}}{N_a + N_b - N_{same}}$$

2.5 Pathway Consistency and TAP-k Score Ranking

We fetched the top 1000 significant probe sets from each dataset and conducting pathway analysis using the GeneGo web tool (GeneGo Inc.). Pathways with P value less than 0.05 from the two comparing datasets were used for pathway consistency analysis.

Threshold Average Precision (TAP-k) [30], a metric used in bioinformatics area for comparing retrieval efficacy of different search engines, is used to measure pathway level consistency. To use TAP-k, a reference pathway database was constructed to represent the “true” pathways. In our study, a reference pathway is defined as those appeared ≥ 3 times among the pathway consistency analysis results by using the five summarization methods. TAP-k score is used to rank summarization method based on concordance rate with the reference pathways.

3 RESULTS

3.1 Reproducibility and Inconsistent Fold Change Direction Proportion

Figure 1 shows the comparison result of the five summarization methods using two differential analysis tools. There are five plots for each pair to show the trend: IFP vs. N_{same} (the number of consistent probe sets), Reproducibility vs. N_{same} , IFP vs. p-value, Reproducibility vs. p-value, and N_{same} vs. p-value. In datasets pairs a and b, where HG-U133A2 and HG-U133A were respectively used, mean value summarization showed a constantly lower inconsistent fold change proportion (IFP) than competing methods (red line in Figure 1 a-1, a-3, b-1, b-3). The same tendency is observed when using either SAM (solid red line) or CyberT (dashed red line) as differential analysis tool. The reproducibility of mean strategy is comparable to other methods at different significance levels (Figure 1 a-4, b-4). Li-Wong summarization method produced more consistent probe sets when SAM is used (cyan line in Figure 1 a-5, b-5), but at the cost of high IFP (cyan line in Figure 1 a-3, b-3) and hence poor performance in the plot of IFP versus N_{same} (cyan line in Figure 1 a-1, b-1). Moreover, the performance of Li-Wong method is more sensitive to the two differential analysis strategies currently used. As indicated in Figures 1 a-5 and b-5 (cyan color), Li-Wong identified more consistent probe

sets when SAM (solid line) is used, but this is not reproduced when applying CyberT (dashed line) method. Median summarization strategy performs worse in all the three dataset pairs we considered here.

Pair c has an overall low IFP (near zero when $p < 0.05$) and high reproducibility. In Figure 1 c-5, RLM (Blue) and Li-Wong (Cyan) methods identified more consistent probe sets than other methods when same p value cutoff standard is used. However, when plotting reproducibility vs. N_{same} , we see slightly better performance of median polish (Green) and mean (Red) methods (Figure 1 c-2). All summarization methods have IFP near to zero when N_{same} is less than 1000 (Figure 1 c-1).

3.2 GeneGo Pathway Consistency Analysis

The reference pathways constructed in the TAP-k score ranking test of each dataset pairs were provided as supplementary materials. The performance of five summarization methods ranked by TAP-k score is illustrated in Figure 2 a, b, c.

GeneGo pathway consistency analysis showed largely variable performance of competing methods depending on both the comparing dataset and differential analysis method used. In general, our analysis shows mean and Li-Wong methods have better performance in identifying more consistent pathways on pairs a and b. In pair c, median and median polish has the best performance (Table 4).

Mean method (Red) ranked first or second in three dataset pairs and its performance is more stable than competing methods. RLM (Blue) and Li-Wong (Cyan) have high TAP-k score in pair a, but the performance is not reproduced in pair c. No obvious alteration in ranking was observed between SAM and CyberT.

4 DISCUSSION

The comparison study of Kerby Shedden [16] based on one ovary tumor dataset and one colon tumor dataset (both used HG-U133A platform) showed that Trimmed mean and Li-Wong methods are more sensitive---detect more genes at a given FDR level. However, the number of significantly differentially expressed genes detected at given FDR level highly depends on the differential analysis algorithm used. Li-Wong strategy by SAM returned nearly double number of probe sets at a given significance level (same when FDR is used) than by CyberT (Supplementary Table 2). Moreover, certain truncation (in the manner recommended by the developers of each method) was implemented in Kerby Shedden's

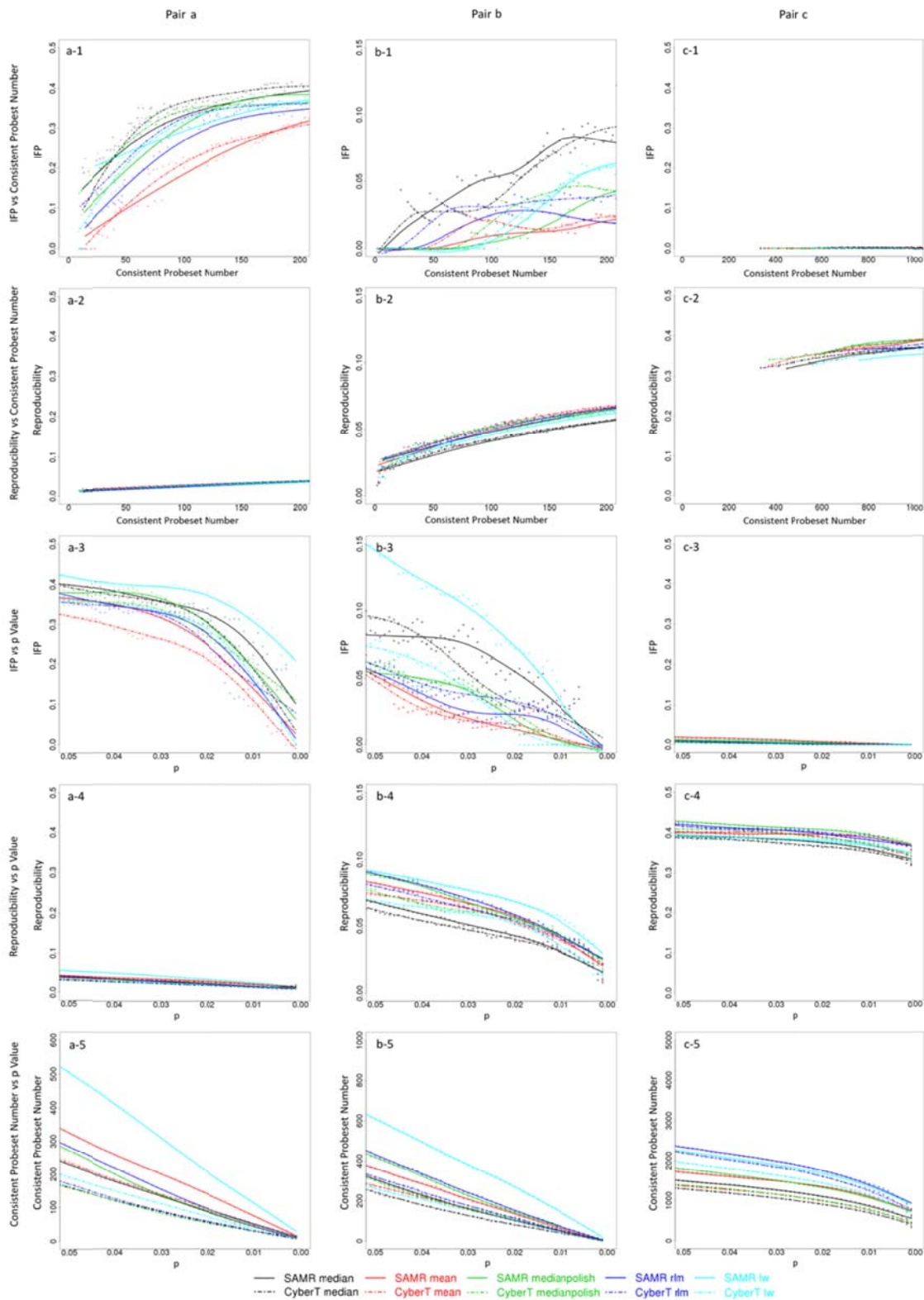


Fig.1. Performance plots for competing summarization methods on three dataset pairs: a) GSE6956 and GSE17356, b) GSE6532KIT and GSE6532OXFT, c) GSE2109 and GSE4560. Solid and dashed lines are results from SAM and CyberT algorithms respectively. Black, red, green, blue, cyan colors are results from median, mean, median polish, RLM, Li-Wong summarization methods respectively.

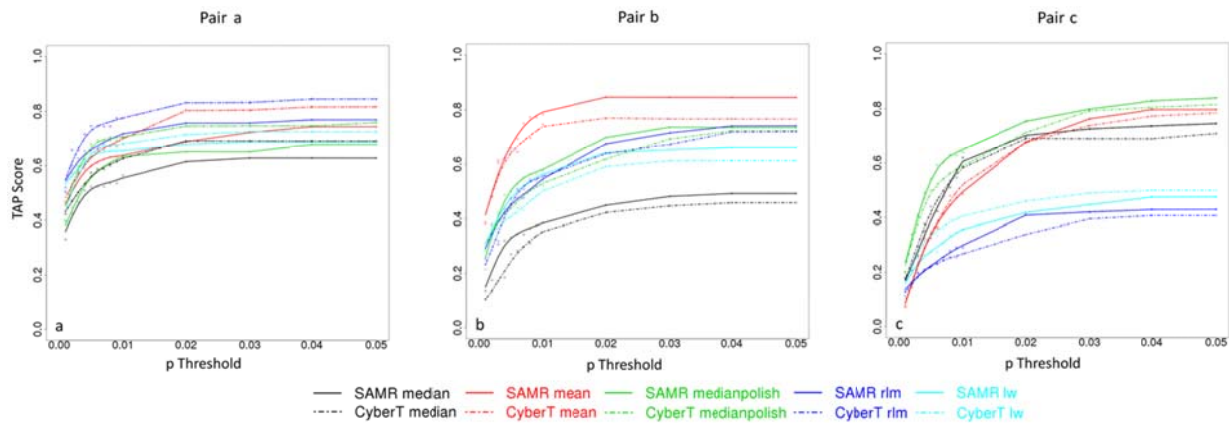


Fig.2. TAP-k score ranking of five summarization methods at different p value threshold: a) GSE6956 and GSE17356, b) GSE6532KIT and GSE6532OXFT, c) GSE2109 and GSE4560. Solid and dashed lines are results from SAM and CyberT algorithms respectively. Black, red, green, blue, cyan colors are results from median, mean, median polish, RLM, Li-Wong summarization methods respectively.

comparison. It is unknown how the truncation may affect the outcome. Rafael A. Irizarry [4] studied the performance of a panel of pre-processing algorithms for Affymetrix using spike-in data. However, it is not clear how those algorithms perform for real biological question. With these potential pitfalls in view, we based the comparison on microarray dataset pairs from research results, and evaluate the performance by checking IFP and reproducibility. One advantage of our approach is that the confounding factors causing a transcript to be falsely called significant in one dataset are unlikely to entirely reappear on the same transcript at another dataset. Partial reappearance of the confounding factors might not be strong enough to constitute a false positive call. Accordingly, results that can be verified in two (or more) datasets with the same study aims are more likely to be true positive. We consider that algorithms generate better consistency among real datasets may convey a better representation of the true expression level. On the other hand, we did not truncate any data in the whole preprocessing flow to avoid unwanted bias towards specific methods. After taking into account these sources of variation, the present study is more likely to reflect the true performance of the competing methods.

Among the five methods studied here, the median method performs the worst in reproducibility. There is no clear winner with respect to reproducibility. The Li-Wong method indeed shows slightly higher reproducibility in pairs a and b (Figure 1 a-4 and b-4) than others under the same p cutoff value. It is comparable to other methods in the plot of reproducibility versus consistent probe set number. This means in one dataset, in order to get an equal number of probe sets reproduced as other competing methods, Li-Wong method needs to call the same amount of probe sets significant in the pairing dataset. Moreover, the Li-Wong method is among the highest inconsistent fold change under the same p cutoff value. The consequence is that among a large number of calls with significance, only a small proportion can be

reproduced with consistent FC direction. This on the other hand indicates that evaluating the performance of one method by merely checking the number of significant probe sets calls is not appropriate. Interestingly, different from our initial speculation that mean value may suffer from high IFP due to its vulnerability to outliers, it outperformed competing methods mainly by lowering IFP on datasets pairs a and b. It is possible that RMA background correction and quantile normalization steps have already excluded potential outliers, and thus further outlier-oriented adjustment is not necessary. Implication is that certain probe sets with the potential to give inconsistent FC direction were not called significant under this strategy. At present, we cannot safely say this controlled FDR without biological evidence that they were not significantly altered by disease status. This strategy, however, indeed renders researches with same study aim more consistency. As indicated in the GeneGo pathway consistency analysis, the mean strategy gives a consistent pathway number ranged first or second in pairs a and b. Its stable performance in TAP-k score ranking indicates its potential to give estimated pathways closer to the reference set. Interestingly, when we compare these five strategies on spike-in dataset proposed by Leslie M. Cope [31] on HG-U133A platform, the performance of median polish and RLM ranked 1st while mean and Li-Wong ranked 15 and 19, which is not in agreement with their performance in pathway consistency analysis. This implies that spike in study may not provide an accurate view of how methods may perform in reality. It might be affected by sources of systematic variation and it is not clear how this might affect evaluation of different data extraction methods.

Plus2 is a combination of the probe sets from HG-U133A and HG-U133B and have probe sets number about twice the size of single A or B platforms. Considering the fact that HG-U133B has nearly eight thousands probe sets with no corresponding gene target, and that a considerable number of the remaining part

target the same genes as platform HG-U133A, we only used the probe sets that were covered by HG-U133A when analyzing pair c. This also helps to make the comparison with the other two pairs consistent. The datasets in pair c showed excellent performance in both lower IFP and higher reproducibility than the other two pairs. We also observed much more number of consistent probe sets in pair c. This might be resulted from the large biological difference between ER- and ER+ breast cancers [28]. Thus the differentially expressed transcripts are more easily identifiable. Mean strategy only has slightly better performance (comparable to median polish) in plot of reproducibility versus N_{same} . Its performance in other plot and pathway consistency analysis is not superior to competing methods. A possible explanation is that in situations where obvious biological differences exist, the consistency is less affected by the summarization methods used.

It is intriguing that mean summarization, a remarkably simple algorithm with the lowest time complexity, outperform (dataset pairs a and b) or comparable to (pair c) several competing algorithms. Similar argument can be found in the 70-gene signature for breast cancer prognosis classification developed by Van't Veer *et al.* [32]. The group sorted the differentially expressed genes between relapsed and relapse free breast cancer patients by p value and picked the top 70 most significant genes, and used the mean expression levels of these 70 genes in relapse free group as the signature. This simple strategy has not yet been outperformed by other more sophisticated strategies [33]. A possible explanation is that complex algorithms with too specific kinds of adjustment result in "fit to noise" under circumstances where high background noise exists. Methods such as Li-Wong iteratively fit a model to the probe data from multiple microarrays to exclude outliers. These iterations may cause signal distortion. It might help to increase the reproducibility of "disease-caused" differentially expressed transcripts, but at the cost of high proportion of inconsistent results.

Note that we used p value rather than FDR as cutoff standard because different datasets generate very different number of probe sets at the same FDR level. Pair b has actually no common probe sets when set FDR to <0.1 . Additionally, the p values obtained from SAM and CyberT are based on regularized t test (by using adjusted variance). We thus use p value to do the comparison while similar results were obtained when using FDR as cutoff standard (obtained by *qvalue* algorithm [29]).

5 CONCLUSION

In the present work, we compared the performance of five summarization algorithms on their ability to lower IFP and improve reproducibility. While maintaining comparable reproducibility, mean summarization strategy gives smaller proportion of probe sets with inconsistent FC direction in two datasets pairs than several currently widely used summarization approaches. Its performance

is weakened in the paired datasets where high biological difference may exist between comparison groups.

ACKNOWLEDGEMENT

The research is supported by the following grant: NSF 0845523, NIH R01LM009959A1, and DOD PC081609.

Reference

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
2. E. M. Southern, *Methods Mol. Biol.* **170**, 1 (2001).
3. R. A. Irizarry *et al.*, *Biostatistics*. **4**, 249 (2003).
4. R. A. Irizarry, Z. Wu, H. A. Jaffee, *Bioinformatics*. **22**, 789 (2006).
5. B. Harr, C. Schlotterer, *Nucleic Acids Res.* **34**, e8 (2006).
6. F. Naef, C. R. Hacker, N. Patil, M. Magnasco, *Genome Biol.* **3**, RESEARCH0018 (2002).
7. Z. Wu, R. A. Irizarry, *Nat. Biotechnol.* **22**, 656 (2004).
8. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics*. **19**, 185 (2003).
9. N. Jiang *et al.*, *BMC. Bioinformatics*. **9**, 284 (2008).
10. R. A. Irizarry *et al.*, *Nucleic Acids Res.* **31**, e15 (2003).
11. D. Rajagopalan, *Bioinformatics*. **19**, 1469 (2003).
12. *Nat. Biotechnol.* **24**, 1039 (2006).
13. R. D. Canales *et al.*, *Nat. Biotechnol.* **24**, 1115 (2006).
14. L. Shi *et al.*, *Nat. Biotechnol.* **24**, 1151 (2006).
15. R. Shippy *et al.*, *Nat. Biotechnol.* **24**, 1123 (2006).
16. K. Shedden *et al.*, *BMC. Bioinformatics*. **6**, 26 (2005).
17. Ben Bolstad, *Affymetrix* (2010).
18. A. Sboner *et al.*, *J. Proteome. Res.* **8**, 5451 (2009).
19. F. R. Hampel, E. M. Onchetti, P. J. Rousseeuw, W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions* (John Wiley and Sons, New York, NY, 1986).
20. C. Li, W. H. Wong, *Proc. Natl. Acad. Sci. U. S. A* **98**, 31 (2001).
21. Affymetrix, *Technical report, Affymetrix* (2002).
22. E. N. Lazaridis, D. Sinibaldi, G. Bloom, S. Mane, R. Jove, *Math. Biosci.* **176**, 53 (2002).
23. V. G. Tusher, R. Tibshirani, G. Chu, *Proc. Natl. Acad. Sci. U. S. A* **98**, 5116 (2001).
24. P. Baldi, A. D. Long, *Bioinformatics*. **17**, 509 (2001).
25. T. A. Wallace *et al.*, *Cancer Res.* **68**, 927 (2008).
26. O. A. Timofeeva *et al.*, *Int. J. Oncol.* **35**, 751 (2009).
27. S. Loi *et al.*, *BMC. Genomics* **9**, 239 (2008).
28. X. Lu *et al.*, *Breast Cancer Res. Treat.* **108**, 191 (2008).
29. J. D. Storey, R. Tibshirani, *Proc. Natl. Acad. Sci. U. S. A* **100**, 9440 (2003).
30. H. D. Carroll, M. G. Kann, S. L. Sheetlin, J. L. Spouge, *Bioinformatics*. **26**, 1708 (2010).
31. L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, T. P. Speed, *Bioinformatics*. **20**, 323 (2004).
32. Van, V *et al.*, *Nature* **415**, 530 (2002).
33. B. Haibe-Kains, C. Desmedt, C. Sotiriou, G. Bontempi, *Bioinformatics*. **24**, 2200 (2008).

Comprehensive Comparison of Gene Set Analysis Tools

Zheng Liu¹, Xuejun Li², Yate-Ching Yuan¹, and Xiwei Wu^{*1}

¹Bioinformatics Core, Department of Molecular Medicine, Beckman Research Institute, City of Hope National Medical Center, 1500 Duarte Rd, Duarte, CA 91010, USA

²Division of Biostatistics, Department of Information Sciences, Beckman Research Institute, City of Hope National Medical Center, 1500 Duarte Rd, Duarte, CA 91010, USA

*To whom correspondence should be addressed.

Abstract - *Gene set analysis has enhanced the microarray data analysis field with biological insights. The first introduced and widely used Over-representation analysis (ORA) method, has the limitation of the requirement of a predetermined differentially expressed genes list. To overcome this limitation, distribution based analysis (DBA) methods were developed with different analysis steps and null hypothesis. To understand the advantages and limitations of these methods, we present a comprehensive survey and evaluate the performance for nine commonly used gene set analysis tools. Methods testing self-contained hypothesis generally have better sensitivity and specificity than methods testing competitive hypothesis. But most of the methods have bias towards larger gene sets with self-contained methods more severe. Therefore, better sensitivity and specificity is obtained at the tradeoff of bigger bias in self-contained methods, and vice versa in competitive methods. We propose a combined performance plot to compare these methods, among which GSA demonstrated superiority over others.*

Keywords: Pathway analysis, microarray, gene set analysis.

1 Introduction

In the last decade, microarray technology largely expedited the biological discovery in basic, clinical and translational research. Initially, the analysis of microarray data was focused on differential expression analysis, where a list of genes that show statistically significant expression difference between conditions can be identified. However, biologists still face difficulties in correlating the target genes with biological significance, e.g. identification of signaling pathways that were differentially activated or repressed is often more interesting than a list of gene names. A gene set contains multiple genes sharing similar biological properties, e.g. gene ontology terms, signaling pathway, and chromosome location. The advantage of analyzing genes as a set is that it can detect coordinate changes that are usually moderate or weak at single gene level. To achieve a biologically relevant interpretation, the target gene list is usually compared to a reference gene list, which is typically all the genes on the microarray, for enrichment of certain gene ontology terms or biological pathways. We refer this method as over representation analysis (ORA). Because of the arbitrary selection of cutoff at the gene list identification step, important findings might be missed and the results are not

stable. A number of cut-off free gene set analysis methods, which provide statistical methods to analyze multiple genes, were introduced later on to prevent any arbitrary cutoff. These tools are often denoted as distribution based analysis (DBA).

Recently, Nam et al. [1] thoroughly summarized and classified 26 gene analysis tools based on their null hypothesis and statistical methods. But the advantages and limitations of these methods are not completely understood. Tian et al. [2] suggested that tests based on both null hypotheses should be considered equally. Goeman et al. [3] further classified gene set analysis methods into three categories, self-contained, competitive and mixed. Dinu et al. [4] recently compared three self-contained analysis tools, SAM-GS [5], global test [6] and ANCOVA global test, and concluded that SAM-GS has slightly higher power. But none of them has conducted thorough performance comparison. We evaluated these tools, and systemically compared their performance using statistical simulation.

2 Methods

2.1 Analysis Tools

In the current study, we have compared GSEA [7] (both gene permutation and phenotype permutation), Tian/sigPathway (both gene permutation and phenotype permutation), ErmineJ [8] ORA, ErmineJ GSR, GSA [9], SAM-GS, SAFE [10], global test and PAGE [11]. Within these tools, there are 4 tools (Global_Test, SAFE, SAME-GS, and Tian_Pheno) testing the self-contained hypothesis, 5 tools (ErmineJ_GSR, ErmineJ_ORA, GSEA_Gene, PAGE, and Tian_Gene) testing the competitive hypothesis, and 2 tools (GSA, GSEA_Pheno) are mixed.

2.2 Simulation Method

Given the diversity of methods implemented in each tool, it is very interesting to examine whether their performance is also different. We developed a testing framework to systemically compare the performance using statistical simulation. We collected 464 signaling and metabolic pathways from KEGG and BioCarta, which are two commonly used canonical pathway databases. For testing purpose, we created 50 pseudo-pathways, each consisting of 20 pseudo-genes, which are differentially expressed between conditions. The major reason to include real pathways is to

generate some false positives so that we can assess the performance of each tool.

The simulation data were generated as a 20,000 x 20 matrix (20,000 genes, 10 normal and 10 treated samples) that follows a standard normal distribution. Differentially expressed (DE) genes were simulated by adding a small constant to the 10 treated samples. The magnitude of increase and the number of DE genes were carefully selected to mimic different scenarios in real experiments, as addressed in more detail in section 3. To prevent any biased results due to any particular simulated data set, one hundred independent simulation data sets were created for each scenario. These simulated data sets were then analyzed by using different analysis tools. Default or recommended parameters of each tool were used whenever possible. Receiver Operating Characteristics (ROC) curve was used to assess the performance of the tools based on the gene set ranks produced by each analysis tools. The mean and standard deviation of the AUCs from 100 simulations were obtained to represent the performance of each tool.

3 Results

3.1 Effects of number of DE gene

To examine the performance of the tools under different levels of differential expression in the gene sets, we generated 10%, 20%, 30%, 40% and 50% DE genes in each of the 50 pseudo-pathways. We also wanted to simulate the phenomenon in real microarray experiment that not all the DE genes belong to any gene sets, which likely to introduce additional level of noise to the data. Therefore, besides the DE genes within the pseudo-pathways, additional DE genes were created in each simulation data set to fix the number of DE genes at 2000. To determine how big the constant should be used to create DE genes, we tested 0.5, 1, and 2.5. Changes with 1 and 2.5 were so strong that all the tested analysis tools were able to achieve an AUC of almost 1. Therefore, we decided to use 0.5 as the constant and all the subsequent results were generated using this constant.

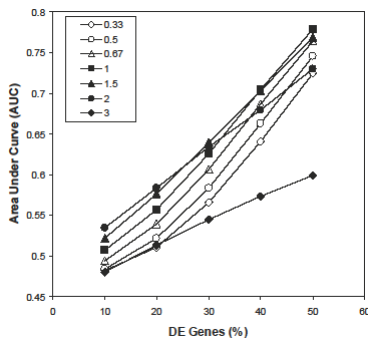


Figure 1. Performance of ORA method under different cutoffs

We found that the performance of ORA method was highly dependent on the selected cutoff. The performance of ORA method increases as the percentage of DE genes increases (Figure 1). Totally 5 cutoff values (0.33, 0.67, 1, 1.5, 2 and 3) were tested. The AUC values range from 0.55 with 10% DE genes to 0.85 with 50% DE genes with cutoff

value of 1. More importantly, different cutoff values result in quite different performance. Cutoff of 1 has the largest AUC, followed by 0.67. Cutoff of 3 gives the lowest AUC, while cutoff of 0.33 and 2 resides in the middle. This result is expected because the theoretical t-statistics for DE genes in the simulated data set is close to 1.5.

3.1.1 Comparison of Gene Set Analysis Methods

To compare the ability to detect enriched gene sets for each analysis methods, 100 simulated data sets were generated and analyzed by each of the tools. To accurately estimate the false positive and false negative rate, gene sets reported as positively and negatively associated with the treatment phenotype were combined in all of the tools. The average AUCs across the 100 simulated data sets are shown in Figure 2 A-E. All the tools perform better with more DE genes in the gene sets. Global_Test and SAM_GS perform the best when the percentage of DE is low, and Sigpath_pheno and GSA perform the best when the percentage of DE is high. Note that even we used the best cutoff for ORA, it is almost the worst method and only better than PAGE and Tian_Gene. PAGE and Tian_Gene have almost identical performance. More importantly, we observed a general trend that phenotype resampling methods are better than gene set resampling methods. As a mixed hypothesis testing method, GSA seems to have consistently performance across different percentage of DE gene.

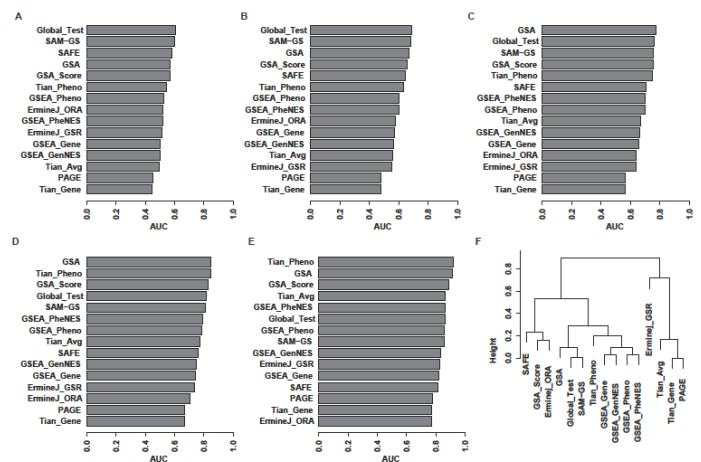


Figure 2. Performance comparison of gene set analysis tools. Mean and standard deviation of Area Under Curve (AUC) from 100 simulated data were calculated for each tool. Hierarchical Clustering of gene set analysis tools based on average ranks of gene sets in 100 simulated data. A-E shows the AUC with 10%, 20%, 30%, 40%, and 50% DE genes respectively. F. The plot shows the clustering result using 20% portion of DEG in a gene set. The color scale represents the similarity between each tool.

We next looked at how similar these tools are relative to each other. The similarity was determined by Euclidean distance between the average ranks of the gene sets across the 100 simulated data sets. We observed very similar results using different percentage of DE genes, and only the data with 30% DE genes are shown in Figure 2F. These tools can be classified into four groups using hierarchical clustering method. Global_Test, SAM-GS, GSA, Tian_Pheno and

GSEA form the biggest group. SAFE and ErmineJ_ORA the second group, while PAGE and Tian_Gene form the third. ErmineJ_GSR is the most distinct from all other methods. GSEA phenotype resampling and gene set resampling method form a subgroup, possibly due to its unique random walking algorithm. We also noted that the distance between PAGE and Tian_Gene is almost 0, which is not surprising because standardization based on large number of gene set resampling within Tian_Gene is equivalent to the standardization used in PAGE. It is unexpected though that ErmineJ_GSR is different from all other methods, because it is theoretically the same method as Tian_Gene. Its performance is also somewhat better than Tian_Gene (Fig. 2F).

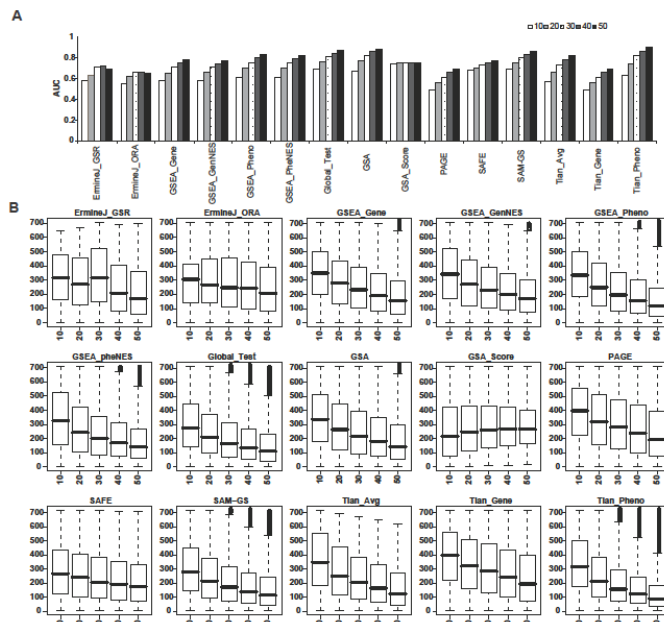


Figure 3. Effect of gene set size. A. The average AUC for different size of gene sets are plotted for each tools. B. The ranks for gene sets with different sizes. The x-axis is the size of gene sets, and the y-axis is the rank of the gene sets based on p-value.

3.1.2 Effects of Gene Set Size

To evaluate the effects of gene set size, we generated simulation data sets with gene set sizes of 10, 20, 30, 40 and 50 separately. Figure 3A shows the average AUCs of each tool across different gene set sizes with 100 simulations. Although to different extent, all of the methods have better performance to detect larger gene sets.

To further examine whether the analysis tools are biased to larger gene set size, we created 5 groups of pseudo gene sets, with size equal to 10, 20, 30, 40 and 50 respectively for each group. Each group contains 50 pseudo-gene sets. Therefore, there are 714 gene sets in the simulated data set, including 250 pseudo-gene sets and 464 real gene sets from KEGG and BioCarta. We created DE genes in the 10 treated samples, in 30% of genes for each of the 250 pseudo gene sets as well as randomly adding 0.5 in the genes not belonging to any gene sets to keep the overall number of DE genes being 2000 out of 20,000 total genes. The average ranks of gene sets with different sizes from 100 simulations

were obtained. If the gene set size has no effects, the average ranks should be similar for gene sets with different sizes. However, as shown in Figure 3B, the gene sets with larger sizes rank better than those with smaller sizes, regardless of what tools are used. The bias is more severe in methods that included a standardization step based on the null distribution of ES, such as Tian/sigPathway and GSEA.

3.2 Performance plot

After the above simulation study, we conclude that both AUC and gene set size are critical factors to evaluate the performance of gene set analysis tools. Therefore, we present the AUC and gene set size effects together on the same plot so that the performance of each tool can be easily compared. The x-axis is the average AUC of simulated gene sets with 10% DE gene background, and y-axis is the slope of ranks among different sizes of gene sets. The best tool should have high AUC, which means better sensitivity and specificity, and low absolute slope, which means less bias to large gene sets. Therefore, the best tools should reside at the upper right corner of the plot. As shown in Figure 4, tools testing competitive hypothesis generally have less bias to gene set sizes, but also have lower AUC. In contrary, tools testing self-contained or mixed hypothesis have more severe bias to gene set sizes, but have better AUC.

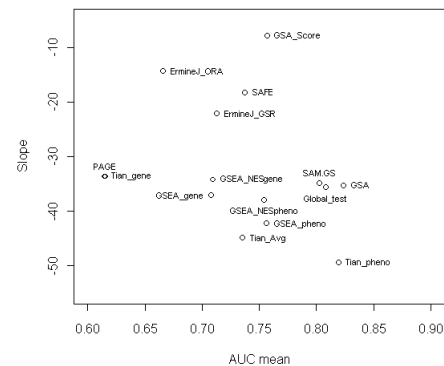


Figure 4. Performance evaluation of gene set analysis tools.

4 Methods performance on experimental data set

To confirm our simulation result with real-world scenario, we further compared the performance of the 11 gene set analysis methods by testing them on the p53 expression data on cancer cell lines. The dataset consisted of the transcriptional profiles from 17 p53+ and 33 p53 mutant cancer cell lines and was downloaded from the GSEA website. We utilize three p53 related pathways to measure the performance of pathway methods. More specifically, we roughly utilized the sum of the rank of the three p53 related pathways based on p-values or normalized enrichment scores assigned by each method to test whether these pathways appear as the top significant pathways.

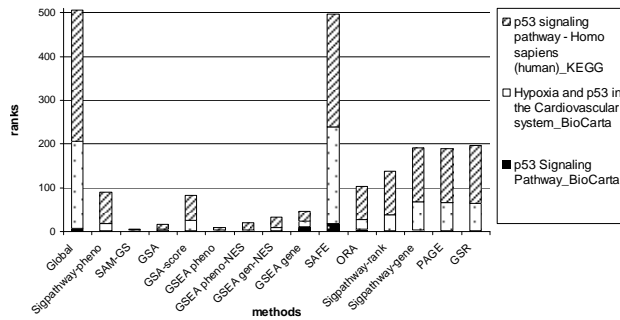


Figure 5. Performance of gene set analysis tools on p53 dataset.

In general, the phenotype-permutation based methods Sigpathway_Pheno, SAM-GS, GSA and GSEA identified the three pathways as relatively top pathways compared to the gene-permutation based methods Sigpathway_Gene, PAGE, and EmineJ_GSR (Figure 5). GSEA gene-set permutation retains its performance mainly due to the unique random-walk strategy.

5 Discussions

In this study, we have systemically compared 11 gene set analysis methods. To our knowledge, this is by far the most comprehensive comparison study. We confirmed that ORA method is highly sensitive to the selected cutoffs, which is likely to create very biased conclusion that is difficult to reproduce. Even when the best cutoff was used, methods based on ORA still have almost the worst sensitivity and specificity when compared to other analysis methods. The strength of ORA methods is that they have less bias to large gene sets. To some extent, we can consider that ORA methods are similar to gene set resampling methods, except that the latter is non-parametric.

We observed that the methods that are self-contained or mixed have better sensitivity and specificity than the methods that are purely competitive. A possible explanation is that gene resampling ignores the correlation structure in the gene sets, which might overestimate the variance in the null distribution of ES. This is also due to the fact that there are 10% DE genes in our simulated data sets, and this portion of genes results in a higher null ES value in gene resampling methods than in phenotype resampling methods. We feel that the chosen 10% DE genes is critical in the evaluation because it is quite common in real microarray experiments that there are significant portion of DE genes not belonging to any tested gene sets. Omitting the 10% DE genes in the simulated data sets will result in very similar performance between self-contained and competitive methods.

It is quite interesting to observe the bias towards large gene set size in most of the tools. This bias still exists even in the tools implementing a standardization step. The good performance of GSA scores suggests that a better scoring system without phenotype resampling can possibly overcome this limitation. As pointed out by Nam D and Kim SY, there are other factors, such as user friendly interface and species support, need to be considered when selecting the best analysis tools.

In summary, we have conducted systemic comparison of popular gene set analysis tools. Our results provide valuable information for researchers to understand the advantages and limitations of these tools.

6 References

- [1] Nam D, Kim SY. "Gene-set approach for expression pattern analysis". Briefings in Bioinformatics Jan 17 2008.
- [2] Tian L, Greenberg SA, Kong SW, et al. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci USA* 2005;102:13544-9.
- [3] Goeman JJ., Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007;23:980-7.
- [4] Dinu I, Liu Q, Potter JD, et al. A biological evaluation of six gene set analysis methods for identification of differentially expressed pathways in microarray data. *Cancer Informatics* 2008;6:357-68.
- [5] Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* 2007;8:242.
- [6] Goeman JJ, van de Geer SA, de Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;20:93-9.
- [7] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102:15545-50.
- [8] Lee HK, Braynen W, Keshav K, et al. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* 2005;6:269.
- [9] Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat* 2007;1:107-29.
- [10] Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 2005;21:1943-9.
- [11] Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 2005;6:144.

High Performance Grid Computation of the Scattered Field Formulation for Nth Order Debye Modeling of the General Dispersive Media

Haythem H. Abdullah¹, Hala A. Elsadek¹, Hesham Eldeeb¹, and Nader Bagherzadeh²

¹ Electronics Research Institute (ERI), Cairo, EGYPT

² Electrical Engineering and Computer Science Department, Henri Samuli school of Engineering, University of California, Irvine (UCI), Irvine, CA, USA

Abstract - *With the advent of the current wireless communication revolution and the increase in its applications, the electromagnetic researchers in conjunction with the medical physicians take the initiative of studying the price of this beneficial technical revolution. It is the health hazards due to the electromagnetic radiation on the human body. The most cost efficient mean of studying these effects is the numerical simulation using the popular FDTD numerical technique. The FDTD simulates the existence of the human body tissues using many fitting models. One of the most famous one is the Debye model. In this paper, the Nth-order Debye model for modelling the human tissues is represented using the scattered field formulation for deducing the FDTD update equation. The scattered field formulation is an accurate method of implementing the different excitation mechanisms for the waves that impinging on the human dispersive tissues media. Although the FDTD is an efficient method, it is a heavy computational one. It may take time of several hours or even days to simulate a single run of a specific problem. In this paper, a parallel scheme is utilized to speed up the running time of the 3D FDTD that includes The Nth-order Debye model. The parallel computations are running on VO (Virtual organization) of the EUMED grid platform. The speed up and behaviour over different number of processors is monitored.*

Keywords: FDTD, Dispersive media, Grid Computation, parallel processing.

1 Introduction

The FDTD method has been applied successfully to a wide variety of problems including complex interaction of electromagnetic fields within materials. These materials include biological tissues, optical materials and ferrite [1] all of which exhibits dispersive behaviour. Three approaches based on the Auxiliary Differential Equation method, (ADE) are developed [2- 5]. The first one is the direct time integration method in which, generalized synchronized scattered field formulation for both Debye and Lorentz models are presented [2- 5]. The second approach is the Nth order ADE which is based on the state equations [4] and is applicable for both Debye and Lorentz media. The third

approach is called the polarization current algorithm for ADE [5] which is generalized to update the scattered electric field [6].

In this paper, the scattered field formulation for the Nth order Debye model [8] in conjunction with other five FDTD update equations are paralyzed using the Massive parallel interface (MPI) library.

Our parallel code is executed on the EUMED grid. ERI has been participated as a partner in the EUMED (European Mediterranean) Grid project entitled (“Empowering E-science across the Mediterranean”) that is a project co-funded by the European Union. It was built using the GLite middleware. Network Time Protocol (NTP) with a time server is used for node synchronization.

2 Problem formulation

Based on the infinite impulse response (IIR) filter design, Sullivan [7] uses the Z-transform technique to calculate the electric field from the electric flux density. In this paper, the analysis is extended to include the N-order Debye model, with the optimum usage of the memory requirement. Then, the scattered field formulation is obtained [8]. The dispersive media update equation in conjunction with the free space; the Uni-axial Perfect Matched Layer UPML, perfect conductor, lossy media, and the thin wire approximation for wire antenna are paralleled as a single patch. The analysis of the dispersive media has the focus in this paper because it is the more general update equation in the FDTD algorithm. Let's start with the displacement vector which is defined as;

$$\bar{D}(\omega) = \varepsilon_0 \varepsilon(\omega) \bar{E}(\omega) \quad (1)$$

Where, $\varepsilon(\omega)$ is the relative permittivity function of the media.

Let's start with the Nth order Debye model with the permittivity of the medium described as follows

$$\varepsilon(\omega) = \varepsilon_\infty + \sum_{p=1}^N \frac{\Delta \varepsilon_p}{1 + j\omega \tau_p} \quad (2)$$

ε_∞ is the relative permittivity at infinite frequency, $\Delta\varepsilon_p$ is the change in relative permittivity due to the Debye pole, N is the number of poles, and τ_p is the pole relaxation time.

Using the Infinite Impulse Response (IIR) filter design [7], the permittivity function can be represented in the Z-domain as follows

$$\varepsilon(z) = \frac{\varepsilon_\infty}{\Delta t} + \sum_{p=1}^N \frac{\Delta\varepsilon_p / \tau_p}{1 - e^{-(\Delta t / \tau_p)} Z^{-1}} \quad (3)$$

Transforming equation (2) into time domain results in

$$\bar{D}(t) = \varepsilon_0 \int_{-\infty}^{\infty} \varepsilon(\tau) \bar{E}(t - \tau) d\tau \quad (4)$$

The convolution integral of equation (4) is converted to a multiplication in the Z-domain, and a factor of Δt , the time increment, is included as follows

$$\bar{D}(z) = \varepsilon_0 \varepsilon(z) \bar{E}(z) \Delta t \quad (5)$$

Assuming N auxiliary variables $\bar{I}_p(z)$ each one corresponds to one Debye pole of the permittivity function $\varepsilon(z)$ such that equation (5) can be rewritten as

$$\bar{D}(z) = \varepsilon_0 \varepsilon_\infty \bar{E}(z) + \sum_{p=1}^N \bar{I}_p(z) \quad (6)$$

Where

$$\bar{I}_p(z) = \frac{\varepsilon_0 \Delta\varepsilon_p \Delta t / \tau_p}{1 - e^{-(\Delta t / \tau_p)} Z^{-1}} \bar{E}(z) \quad (7)$$

Rewriting equation (7) in a more convenient form results

$$\left[1 - e^{-(\Delta t / \tau_p)} Z^{-1}\right] \bar{I}_p(z) = \varepsilon_0 \Delta\varepsilon_p (\Delta t / \tau_p) \bar{E}(z) \quad (8)$$

Now, equation (8) can be transformed into discrete time domain simply by shifting each field component multiplied by Z^{-1} in the Z-domain, one time step later in the discrete time domain giving

$$\bar{I}_p^{n+1} = e^{-(\Delta t / \tau_p)} \bar{I}_p^n + \varepsilon_0 \Delta\varepsilon_p (\Delta t / \tau_p) \bar{E}^{n+1} \quad (9)$$

Thus the auxiliary variable \bar{I}_p^{n+1} can be calculated from its previous value and the present electric field sample. Now, proceeding to get the electric field updating equation by transforming equation (6) into discrete time domain giving

$$\bar{D}^{n+1} = \varepsilon_0 \varepsilon_\infty \bar{E}^{n+1} + \sum_{p=1}^N \bar{I}_p^{n+1} \quad (10)$$

Equation (10) reveals that the electric \bar{E}^{n+1} cannot be updated from \bar{I}_p^{n+1} , since it is calculated from the electric field at the same time step, so it is useful to substitute for \bar{I}_p^{n+1} by its value given in equation (9) and proceeding again to get the electric fields in terms of the previous value of \bar{I}_p^n giving

$$\bar{E}^{n+1} = (1/C_e) \bar{D}^{n+1} - \sum_{p=1}^N (C_{xp}/C_e) \bar{I}_p^n \quad (11)$$

Where \bar{D} is the electric flux density

$$C_e = \varepsilon_0 \left[\varepsilon_\infty + \sum_{p=1}^N \Delta\varepsilon_p (\Delta t / \tau_p) \right] \quad (12-a)$$

$$C_{xp} = e^{-(\Delta t / \tau_p)} \quad (12-b)$$

Now we can summarize the programming sequence and summary of the updating equations within the time increment loop for the scattered field formulation by assigning the summation of the incident and the scattered fields instead of the total fields.

$$\frac{\partial \bar{D}(t)}{\partial t} = \nabla \times \bar{H}(t) + \varepsilon_0 \frac{\partial \bar{E}_i(t)}{\partial t}$$

$$\bar{E}_s^{n+1} = (1/C_e) \bar{D}^{n+1} - \sum_{p=1}^N (C_{xp}/C_e) \bar{I}_p^n - \bar{E}_i^{n+1}$$

$$\bar{I}_p^{n+1} = e^{-(\Delta t / \tau_p)} \bar{I}_p^n - \varepsilon_0 \Delta\varepsilon_p (\Delta t / \tau_p) (\bar{E}_s^{n+1} + \bar{E}_i^{n+1})$$

$$\frac{\partial \bar{H}_s(t)}{\partial t} = -\frac{1}{\mu} \nabla \times \bar{E}_s(t)$$

Where $\bar{H}_s(t)$ is the magnetic field intensity.

The previous sequence is the general sequence for both the scattered field formulation and the total field formulation. The total field formulation can be easily set by considering the incident fields equal zero as the case of studying the effect of the scattering from mobile wireless devices on the human (user) tissues, so that the total field will be equal to the scattered field. Finally, one can say that the scattered field formulation may be considered as the general case.

3 Analysis of the parallel formulation

In this paper, the FDTD algorithm treats six types of media; the perfect conductor [9], the thin wire approximation [10] with infinitesimal gap [11], free space, general lossy media, Uni-axial perfect matched layer [9], and then the Nth order

Debye model for modeling general dispersive media [6]. In our serial code, six functions are assigned to compute the three electric field components and the three magnetic field components at each cell location. Each function has a selection between the update equations for the six media. From the above discussion, it is worth mentioning that, in each function there is only a selection of one group of update equations for only one medium at a specific location. Figure 1 shows the locations of the field components in each cell. This ensures no Amstrong complexity in the analysis. Amstrong complexity occurs when the data structures are subjected to a sequence of instructions rather than one set of instruction. In this sequence, one instruction may perform certain modifications that have an impact on other instructions in the sequence at the run time. By analyzing the time axis, the electric and magnetic fields at each time step are evaluated from the neighborhood fields in the previous time step as shown in Figure 1. So, this axis cannot be distributed between processors. By analyzing the spatial axes x, y, or the z axis, the electric field or its corresponding displacement vector D

or the auxiliary variables \bar{I}_p^n are calculated from the neighborhood magnetic fields. From this fact, one can divide one axis from the three spatial axes between processors to calculate the fields as illustrated in Figure 2.

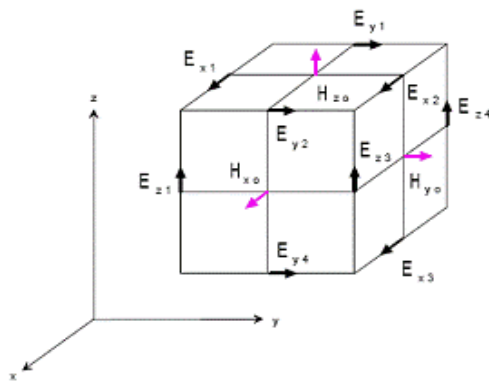


Fig.1 A unit cell from the discretized domain with fields' components' positions [9].

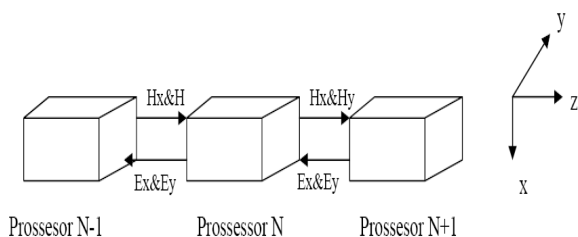


Fig.2 Data Dependencies [13]

4 Results and Discussions

Let us calculate the reflection coefficient at the interface between the air and the muscle tissue. The permittivity of the 2/3 muscles is assumed because the average permittivity of human body tissues is close to that for 2/3 muscles. The associated parameters are: $\epsilon_\infty = 19$, $\epsilon_{s1} = 10019$, $\epsilon_{s2} = 42$, $\tau_1 = 0.71 \times 10^{-6} / 2\pi$, and $\tau_2 = 0.75 \times 10^{-10} / 2\pi$. The one dimensional problem assumes problem space of 1000 cells, 500 of which are used to represent the air, and the remaining 500 cells are used to represent the 2/3 muscle equivalent material. The cell size is taken $37.5 \mu m$ and time step $\Delta t = \Delta x / 2c$. The assumed incident pulse takes the form $E(t) = 1000e^{-(t-t_0)^2/T^2}$ where $t_0 = 400\Delta t$, and $T = 152\Delta t$. Figure 3 illustrates the reflection coefficient of the numerical FDTD solution compared to the analytical one that is given by

$$|R(\omega)| = \left| \frac{\sqrt{\epsilon_0} - \sqrt{\epsilon(\omega)}}{\sqrt{\epsilon_0} + \sqrt{\epsilon(\omega)}} \right| \tag{13}$$

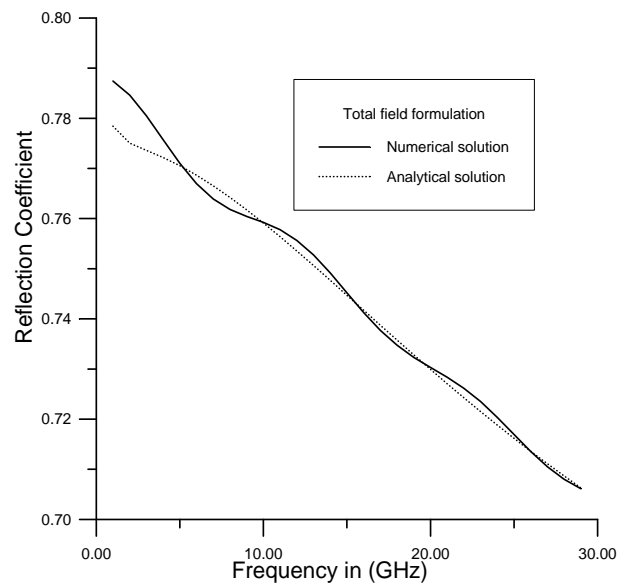


Fig.3 The reflection coefficient at the interface between the air and the 2/3 muscle tissue material.

The parallel 3D FDTD is applied to calculate the effect of rectangular microstrip antenna, that is most used now in mobile phones, [12]. The antenna is 5 cm away from the head side. The space domain enclosing the human head and the antenna is equal to $60 \times 60 \times 90$ cells. The microstrip antenna has a substrate material of relative permittivity 2.2, a substrate

thickness of 6.73 mm and rectangular patch size of 134.6 mm in the x direction and 111 mm in the z direction. The feed is performed via a microstrip line of 67.3 mm length in the x direction and 37 mm width in the z direction. A noticeable time reduction is observed up to 25 processor over the network "ce0.m3pec.u-bordeaux1.fr:2119/jobmanager-pbs-eumed" in the EUMED grid using the MPI library. The serial code over the same network takes time of 9520 second. Figure 4 shows the time reduction by applying the same problem on EURO-EMD Grid parallel computation environment. From Figure 4, one can notice a linear reduction of the time up to 14 processor while still exist a reduction even for 25 processors. The calculation time for the 3D FDTD is reduced by 38 times (2.6% only of the serial execution time) over 25 processors which illustrates the merits of applying parallel computation to the algorithm. If the number of processors increase than 25, the reduction in the execution time approximately stopped due to the effect of communication time which at this point becomes comparable or even greater to the execution time.

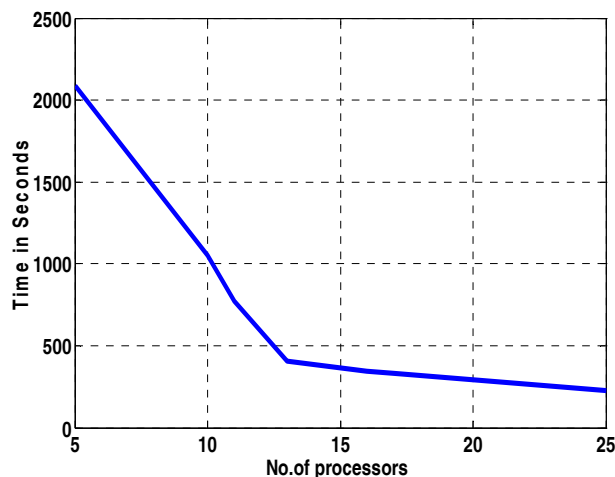


Fig.4 The run time of the parallel 3D FDTD over the EUMED grid

5 Conclusions

In this paper, a parallel processing algorithm is developed for the FDTD code including six update equations. The focus of this study is in illustrating the parallel strategy of including the dispersive properties of the human tissues in a parallel code. The derivation of the scattered field formulation for including the dispersive properties of the human media is by using Nth order Debye model. Good accuracy is observed when calculating the reflection coefficient from a two pole Debye tissue. The parallel 3D FDTD code is applied on the scattering from the human head due to the excitation by

microstrip antenna. The reduction of the execution time is observed up to 25 processor by about 96%. The parallel code is executed on the virtual organization of EUMED grid.

6 References

- [1] Ahmed Attiya and Haythem H. Abdullah, "Shift Operator Finite Difference Time Domain: An Efficient Unified Approach for simulating Wave Propagation in Different Dispersive media," 1st Middle East Conference on Antennas and Propagation October 20-22, 2010, Cairo, Egypt.
- [2] R. M. Josef, S. C. Hagness, and A. Taflove, "Direct time integration of Maxwell's equations in linear dispersive media with absorption for scattering and propagation of femtosecond electromagnetic pulses," *Optics Lett.*, Vol. 16, 1991, PP. 1412-1414.
- [3] Om. P. Gandhi, B. Gao, and J. Chen, "A Frequency Dependent Finite-Difference Time Domain Formulation for General Dispersive media," *IEEE Trans. Microwave Theory and Techniques*, Vol. 41, 1993, PP. 658-665.
- [4] J. L. Young, "Propagation in linear dispersive media: finite-difference time-Domain Methodologies," *IEEE Trans. Antennas Propagat.*, Vol. 43, No. 4, 1995, PP. 422-426.
- [5] M. Okoniewski, M. Mrozowski, M. Stuchly, "Simple treatment of multi-term dispersion in FDTD," *IEEE Trans. Microwave and Guided Wave Letters*, Vol. 7, No. 5, 1997, PP.121-123.
- [6] H. H. Abdullah, F. M. EL-Hefnawi, E. A. Hashish, and A. Z. Elsherbeni, "Scattered field FDTD formulations for Debye dispersive media," *URSI2001, International Conference on Electromagnetic in Advanced Applications ICEAA01*, Sept. 10-14, Turin, Italy, 2001, PP. 3-6.
- [7] D.M. Sullivan, "The Z-Transform Theory and the FDTD Method," *IEEE Trans. Antennas and Propagat.*, Vol. 44, No. 1, 1996, PP. 28-34.
- [8] H. Beggs, "validation and demonstration of frequency approximation methods for modeling dispersive media in FDTD," *ACES Journal*, Vol. 14, No. 2, 1999, PP. 52-58.
- [9] A. Taflove, and S. Hagness, *Computational Electrodynamics: The Finite Difference Time Domain Method*, Third Edition, Artech House, INC., 2005.
- [10] F. Edelvik, "A new technique for accurate and stable modeling of arbitrarily oriented thin wires in the FDTD," *IEEE Trans. Electromagn.Compat.*, Vol. 45, No. 5, 2003, PP. 416-423.
- [11] S. Watanabe, and M. Taki, "An improved FDTD Model for the Feeding Gap of a thin-Wire Antenna," *IEEE Trans. Microwave and Guided wave letters*, Vol. 8, No. 4, 1998, PP. 152-154.
- [12] Hesham Eldeeb, Hala Elsadek, Maha Dessokey, Haytham H. Abdallah, and Nader Bagherzadeh, "High Performance Parallel Computing for FDTD Numerical Technique in Electromagnetic Calculations for SAR

Distribution inside Human Head”, 14th WSEAS international conference on computer, 23-25 July 2010, corfu, Greece.

- [13]H. Eldeeb, H. Elsadek, H. Abdalh, M. Desouky and . Bagerhzadeh, "FDTD accelerator for SAR distribution in human head due to radiation from wireless devices", Conference proceeding of EMTS 2007, pp: 2310-2314.

Criteria for Annihilation of HIV-1 During HAART Therapy

Frank Nani and Mingxian Jin

Department of Mathematics and Computer Science
Fayetteville State University, Fayetteville, NC 28301, USA

Abstract - HAART therapy of HIV-1 induced AIDS is modeled by a system of non-linear deterministic differential equations. The clinically plausible patho-physiological equations depict the dynamics of uninfected $CD4^+$ T cells (x_1), HIV-1 infected $CD4^+$ T cells (x_2), HIV-1 virions in the blood plasma (x_3), HIV-1 specific $CD8^+$ T cells (x_4), and the concentration of HAART drug molecules (x_5). The criteria for the existence of clinically desirable therapeutic outcomes are presented. In particular, the necessary and sufficient conditions for the annihilation of HIV-1 virions are clearly exhibited in terms of the model physiological parameters. Computer simulations are presented illustrating patho-physiodynamics of HIV-1 induced AIDS. In this paper, HAART protocols with constant continuous or periodic transdermal and intravenous drug infusions are used in our mathematical model.

Keywords: HIV-1 patho-physiodynamics, mathematical modeling, HAART therapy, AIDS cure criteria, Michaelis-Menten kinetics

1 Introduction

HIV-1 virions induce AIDS by orchestrating an irreversible destruction of the $CD4^+$ T cells which then paralyze the immune system of the HIV-1 positive person. As a result of these physiological events, a host of opportunistic bacterial and viral infections overwhelm the HIV-1 positive person [12]. Highly active anti-retroviral therapy (HAART) protocols have been approved as an efficacious treatment of HIV-1 induced AIDS. This protocol consists of nucleoside reverse transcriptase inhibitors, non-nucleoside reverse transcriptase inhibitors, protease inhibitors, anti-fungals /anti-bacterials and in future, integrase inhibitors. The reverse transcriptase inhibitors prevent reverse transcription of HIV-1 specific DNA. The protease inhibitors are antagonistic to maturation and formation of new HIV-1 virions. The possible role of integrase inhibitors is to prevent the integration of HIV-1 viral DNA into the patients' DNA [14].

HAART therapy is responsible for the reduction of viral load in $CD4^+$ T cells and production of measurable reconstitution of the patients' immune system [17]. However, HAART protocols have limited therapeutic efficacy due to extreme toxicity, non-compliance, intermittent scheduling, biochemical/clinical drug resistance, short drug half-life and low bio-availability.

Many mathematical models of HAART therapy have been developed in an attempt to demonstrate the existence of efficacious and optimal therapies that will minimize side effects [8, 9, 10, 11, 13, 14, 15, 16]. Zaric et al. in 1998 presented a model which was focused on the simulation of protease inhibitors and role of drug resistant HIV-1 virions [18]. Stengel in [14] presented a mathematical model of HIV-1 infection and HAART which demonstrated the efficacy of a mathematically optimal therapy. Using the LQR, Scheme, Caetano and Yoneyama in [2] constructed a HAART model which incorporated the roles of latently infected $CD4^+$ T cells, and discussed how the reverse transcriptase and protease inhibitors affected HIV-1 dynamics during HAART.

In this paper, an elaborate mathematical model will be constructed which will incorporate physiologically plausible effects such as Michaelis-Menten kinetics, role of HIV-1 latent viral reservoirs, continuous transdermal drug delivery, and the implicit lymphocyte proliferation induction by the $CD4^+$ T cells. The activation and proliferation is accomplished by a paracrine and autocrine processes which are mediated by the cytokine interleukin-2, secreted by the $CD4^+$ T cells. Several authors investigated the consequences of structured long-term and short-term treatment interruptions during HAART [1, 2, 4, 8]. The current model will discuss these consequences by means of simulations.

The current paper will be divided into five sections. The first section gives the introduction into HAART therapy and provides the basis for current research. This is followed by presentation and discussion of the model parameters in Section 2. In Section 3 the mathematical model of HAART therapy will be constructed. Also the necessary and sufficient criteria for annihilation of HIV-1 virions during HAART will be presented in this section. In Section 4, clinically plausible computer simulations will be exhibited. Section 5 will be the summary and discussion of the basic results of the paper.

2 Parameters

The model parameters, constants, and variables are listed as follows.

- x_1 : the number density of non-HIV-1-infected $CD4^+$ helper T-lymphocytes per unit volume
- x_2 : the number density of HIV-1 infected $CD4^+$ helper T-lymphocytes per unit volume

x_3 : the number density of HIV-1 virions in the blood plasma per unit volume

x_4 : the number density of HIV-1 specific CD8⁺ cytotoxic T-lymphocytes per unit volume

x_5 : the concentration of drug molecules of the HAART treatment protocol

S_1 : rate of supply of un-infected CD4⁺ T₄-lymphocytes

S_2 : rate of supply of latently infected CD4⁺ T₄-lymphocytes

S_3 : rate of supply of HIV-1 virions from macrophage, monocytes, microglial cells and other lymphoid tissue different from T₄-lymphocytes

S_4 : rate of supply of CD8⁺ T₈ lymphocytes from the thymus

D : rate of HAART drug infusion by transdermal delivery

a_i, b_i : constant associated with activation of lymphocytes by cytokine interleukin-2 (IL-2) ($i = 1, 2, 3, 4$)

c : rate of HAART drug degradation and excretion

α_i : constant associated with HIV-1 infection of CD4⁺ T₄ helper cells ($i = 1, 2, 3$)

β_1 : the number of HIV-1 virions produced per day by replication and budding in CD4⁺ T₄ helper cells

β_2 : rate constant associated with replication and "budding" of HIV-1 in syncytia CD4⁺ T₄ helper cells per day per microliter (μ l) and released into the blood plasma

β_3 : the number of HIV-1 virions produced per day by replication and "budding" in non-syncytia CD4⁺ T₄ helper cells and released into the blood plasma

η_i : constant depicting the rate of which HIV-1 virions incapacitate the CD8⁺ T₈ cytotoxic cells ($i = 1, 2$)

(σ_0, λ_0) : Michaelis-Menten metabolic rate constants associated with HAART drug elimination

(σ_i, λ_i) : Michaelis-Menten metabolic rate constants associated with HAART drug pharmacokinetics ($i = 2, 3$)

ξ_i : cytotoxic coefficient where $0 \leq \xi_i \leq 1$ ($i = 2, 3$)

q_i : constant depicting competition between infected and un-infected CD4⁺ T₄ helper cells ($i = 1, 2$)

k_i : constant depicting degradation, loss of clonogenicity or "death" ($i = 1, 2, 3, 4$)

e_{i0} : constant depicting death or degradation or removal by apoptosis (programmed cell death) ($i = 1, 2, 3, 4$)

K_i : constant associated with the killing rate of infected CD4⁺ T₄ cells by CD8⁺ T₈ cytotoxic lymphocytes ($i = 1, 2$)

All the parameters are positive.

$$\left\{ \begin{array}{l} \dot{x}_1 = S_1 + a_1 x_1^2 e^{-b_1 x_1} - \alpha_1 x_1 x_3 - q_1 x_1 x_2 - k_1 x_1 - e_{10} \\ \dot{x}_2 = S_2 + a_2 x_1 x_2 e^{-b_2 x_1} + \alpha_2 x_1 x_3 - q_2 x_1 x_2 - k_2 x_2 - \beta_1 x_3 \\ \quad - K_1 x_2 x_4 - e_{20} - \frac{\xi_2 \sigma_2 x_2 x_5}{\lambda_2 + x_5} \\ \dot{x}_3 = S_3 + \beta_2 x_2 x_3 + \beta_3 x_3 - \alpha_3 x_1 x_3 - \eta_1 x_3 x_4 - k_3 x_3 - e_{30} \\ \quad - \frac{\xi_3 \sigma_3 x_3 x_5}{\lambda_3 + x_5} \\ \dot{x}_4 = S_4 + a_4 x_1 x_4 e^{-b_4 x_1} - K_2 x_2 x_4 - \eta_2 x_3 x_4 - k_4 x_4 - e_{40} \\ \dot{x}_5 = Df(t) - \frac{\sigma_0 x_5}{\lambda_0 + x_5} - \frac{\sigma_2 x_2 x_5}{\lambda_2 + x_5} - \frac{\sigma_3 x_3 x_5}{\lambda_3 + x_5} \\ f(t) = \begin{cases} 1 & \text{for constant continuous input} \\ \lceil \lceil \sin mt \rceil \rceil & \text{for periodic input} \end{cases} \\ x_i(t_0) = x_{i0} \quad \text{for } i = \{1, 2, 3, 4, 5\} \end{array} \right. \quad (3.1)$$

The model includes the following clinical improvement:

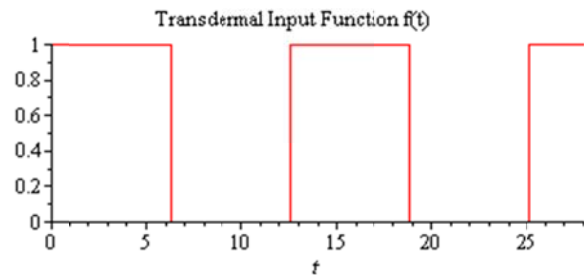
(i) The drug delivery uses transdermal, stealth-liposome encapsulated drug delivery, instead of the matrix tablet form because of improved therapeutic efficacy and reduced gastro-intestinal toxicity [6]. It is also assumed that elastic liposomes are formulated and selectively targeted such as to reduce toxicity to non-HIV-1-infected CD4⁺ T cells (x_1) and CD8⁺ cytotoxic T cells (x_4).

(ii) The HAART drug is such that each renal excretion and body clearance rate follows Michaelis-Menten kinetics.

(iii) $g(x_1, x_j) = a_j x_1 x_j e^{-b_j x_1}$ for $j = (1, 2, 4)$

This function depicts the process of lymphocyte activation which is mediated by x_1 (CD 4⁺) T helper cells. These cells secrete a cytokine called interleukin-2.

(iv) The periodic input function $f(t) = \lceil \lceil \sin(5t) \rceil \rceil$ can be depicted by the following plot:



3 Model Equations and Analyses

3.1 Model equations

The HIV-1 patho-physiological dynamics during HAART therapy can be modeled using the following system of non-linear ordinary differential equations:

3.2 Criteria for Annihilation of HIV-1 Virions

In this subsection, the necessary and sufficient criteria will be presented for the cure of an AIDS patient through

annihilation of the HIV-1 infected CD4⁺ T cells and plasma viremia. This criterion is derived for the scenario for which $f(t) \equiv 1$, which corresponds to constant continuous application of HAART drug either by transdermal delivery or intravenous infusion.

The desired physiological steady states during HAART therapy, are $E_1 = [\hat{x}_1, 0, 0, 0, \hat{x}_5]$ and $E_2 = [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$. In each of these, the HIV-1 infected CD4⁺ T cells (x_2) and plasma HIV-1 virions (x_3) are annihilated. In particular, $E_2 = [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$ is plausibly physiologically easily attainable in an AIDS patient since some HIV-1 specific CD8⁺ (cytotoxic T) cells usually persist during HAART as memory T cells.

Thus the criteria for annihilation of HIV-1 virions will be derived using $E_2 = [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$ as a target steady state. In $R_+^{x_1 x_4 x_5} = [x_1, x_4, x_5 \mid x_1 \geq 0, x_4 \geq 0, x_5 \geq 0]$, the model equations reduce to (3.2).

$$\begin{cases} \dot{x}_1 = S_1 + a_1 x_1^2 e^{-b_1 x_1} - k_1 x_1 - e_{10} \\ \dot{x}_4 = S_4 + a_4 x_1 x_4 e^{-b_4 x_1} - k_4 x_4 - e_{40} \\ \dot{x}_5 = Df(t) - \frac{\sigma_0 x_5}{\lambda_0 + x_5} \\ f(t) = 1 \\ x_i(t_0) = x_{i0} \quad \text{for } i = \{1, 4, 5\} \end{cases} \quad (3.2)$$

Consider the Liapnnov functional:

$$V := \sum \frac{1}{2} c_i (x_i - \bar{x}_i)^2 \quad (3.3)$$

where $i = \{1, 4, 5\}$ and $c_i \in R_+ = (0, \infty)$

The derivative of V along the solution curves of the model equations yields the result:

$$\dot{V} = c_1 (x_1 - \bar{x}_1) \dot{x}_1 + c_4 (x_4 - \bar{x}_4) \dot{x}_4 + c_5 (x_5 - \bar{x}_5) \dot{x}_5$$

At a steady state $f(t) = 1$, the following equations hold.

$$\begin{cases} S_1 - e_{10} = k_1 \bar{x}_1 - a_1 \bar{x}_1^2 e^{-b_1 \bar{x}_1} \\ S_4 - e_{40} = k_4 \bar{x}_4 - a_4 \bar{x}_1 \bar{x}_4 e^{-b_4 \bar{x}_1} \\ D = \frac{\sigma_0 \bar{x}_5}{\lambda_0 + \bar{x}_5} \end{cases} \quad (3.4)$$

Thus

$$\begin{aligned} \dot{V} &= c_1 k_1 (x_1 - \bar{x}_1)(\bar{x}_1 - x_1) \\ &\quad + a_1 c_1 (x_1 - \bar{x}_1)[G(\bar{x}_1) - G(x_1)] \\ &\quad + c_4 k_4 (x_4 - \bar{x}_4)(\bar{x}_4 - x_4) \\ &\quad + a_4 c_4 (x_4 - \bar{x}_4)[F(\bar{x}_1, \bar{x}_4) - F(x_1, x_4)] \\ &\quad + c_5 \sigma_0 [L(\bar{x}_5) - L(x_5)] \end{aligned}$$

where

$$\begin{aligned} G(x_1) &= x_1^2 e^{-b_1 x_1} \\ F(x_1, x_4) &= x_1 x_4 e^{-b_4 x_1} \\ L(x_5) &= \frac{x_5}{\lambda_0 + x_5} \end{aligned} \quad (3.5)$$

such that G, F, L are continuous, differentiable, and have bounded, derivatives.

Let

$$\begin{aligned} u_1 &= x_1 - \bar{x}_1 \\ u_2 &= x_4 - \bar{x}_4 \\ u_3 &= x_5 - \bar{x}_5 \end{aligned} \quad (3.6)$$

and let

$$X = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \in R_+^3 \quad (3.7)$$

such that X^T denotes the transpose of X . Define $A \in M_{mn}(R)$ such that

$$A = \begin{bmatrix} a_{11} & \frac{1}{2} a_{12} & \frac{1}{2} a_{13} \\ \frac{1}{2} a_{12} & a_{22} & \frac{1}{2} a_{23} \\ \frac{1}{2} a_{13} & \frac{1}{2} a_{23} & \frac{1}{2} a_{33} \end{bmatrix} \quad (3.8)$$

Now

$$\begin{aligned} \dot{V} &:= a_{11} u_1^2 + \frac{1}{2} a_{12} u_1 u_2 + \frac{1}{2} a_{13} u_1 u_3 \\ &\quad + \frac{1}{2} a_{12} u_2 u_1 + a_{22} u_2^2 + \frac{1}{2} a_{23} u_2 u_3 \\ &\quad + \frac{1}{2} a_{13} u_3 u_1 + \frac{1}{2} a_{23} u_3 u_2 + a_{33} u_3^2 \\ &= -c_1 k_1 (x_1 - \bar{x}_1)^2 - a_1 c_1 (x_1 - \bar{x}_1)[G(x_1) - G(\bar{x}_1)] \\ &\quad - c_4 k_4 (x_4 - \bar{x}_4)^2 - a_4 c_4 (x_4 - \bar{x}_4)[F(x_1, x_4) - F(\bar{x}_1, \bar{x}_4)] \\ &\quad - c_5 \sigma_0 (x_5 - \bar{x}_5)[L(x_5) - L(\bar{x}_5)] \end{aligned} \quad (3.9)$$

In particular, the $[a_{ij}]_{3 \times 3}$ are defined as follows:

$$\begin{cases} a_{11} := -[c_1 k_1 + a_1 c_1 (\frac{G(x_1) - G(\bar{x}_1)}{x_1 - \bar{x}_1})] \\ a_{12} := -a_4 c_4 [\frac{F(x_1, x_4) - F(\bar{x}_1, \bar{x}_4)}{x_4 - \bar{x}_4}] = a_{21} \\ a_{13} = a_{31} = 0 \\ a_{22} = -c_4 k_4 \\ a_{23} = a_{32} = 0 \\ a_{33} := -c_5 \sigma_0 [\frac{L(x_5) - L(\bar{x}_5)}{x_5 - \bar{x}_5}] \end{cases} \quad (3.10)$$

As the flow associated with the model equations approaches $E_2 = [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$, the matrix entries $[a_{ij}]_{3 \times 3}$ have the following form:

$$\begin{aligned} a_{11} &\rightarrow -[c_1 k_1 + a_1 c_1 G'(\bar{x}_1)] \\ a_{12} &\rightarrow -a_4 c_4 \frac{\partial}{\partial x_1} [F(\bar{x}_1, \bar{x}_4)] \\ a_{22} &= -c_4 k_4 \\ a_{33} &\rightarrow -c_5 \sigma_0 L'(\bar{x}_5) \end{aligned} \quad (3.11)$$

In particular, it can be shown that

$$\begin{aligned} \frac{\partial}{\partial x_1} F(\bar{x}_1, \bar{x}_4) \Big|_{x_1 = \bar{x}_1} &= \bar{x}_4 e^{-b_4 \bar{x}_1} (1 - \bar{x}_1 b_4) \\ G'(\bar{x}_1) &= \bar{x}_1 e^{-b_1 \bar{x}_1} (2 - \bar{x}_1 b_1) \\ L'(\bar{x}_5) &= \frac{\lambda_0}{(\lambda_0 + \bar{x}_5)^2} > 0 \end{aligned} \quad (3.12)$$

Thus

$$F_{x_1}(\bar{x}_1, \bar{x}_4) = \begin{cases} > 0 \text{ if } \bar{x}_1 < \frac{1}{b_4} \\ = 0 \text{ if } \bar{x}_1 = \frac{1}{b_4} \\ < 0 \text{ if } \bar{x}_1 > \frac{1}{b_4} \end{cases} \quad (3.13)$$

Similarly,

$$G'(\bar{x}_1) = \begin{cases} > 0 \text{ if } \bar{x}_1 < \frac{2}{b_1} \\ = 0 \text{ if } \bar{x}_1 = \frac{2}{b_1} \\ < 0 \text{ if } \bar{x}_1 > \frac{2}{b_1} \end{cases} \quad (3.14)$$

Since $F(x_1, x_4)$, $G(x_1)$, and $L(x_5)$ are continuous and differentiable functions in each variable, the matrix entries a_{11} , a_{12} , a_{22} , a_{33} exist and remain bounded in the space: $R_{+}^{x_1 x_4 x_5}$, as $[x_1, 0, 0, x_4, x_5] \rightarrow [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$.

Hence, \bar{V} can be written in the form

$$\begin{aligned} \bar{V} &= X^T A X \\ \text{where} \\ A &= \begin{bmatrix} a_{11} & \frac{1}{2} a_{12} & 0 \\ \frac{1}{2} a_{12} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \end{aligned} \quad (3.15)$$

The matrix A is negative definite if the following criteria hold:

$$\begin{aligned} A_1 &= \det a_{11} < 0 \quad \text{or} \quad a_{11} < 0 \\ A_2 &= \det \begin{vmatrix} a_{11} & \frac{1}{2} a_{12} \\ \frac{1}{2} a_{12} & a_{22} \end{vmatrix} > 0 \quad \text{or} \quad a_{11} a_{22} - \frac{1}{4} (a_{12})^2 > 0 \\ A_3 &= \det \begin{vmatrix} a_{11} & \frac{1}{2} a_{12} & 0 \\ \frac{1}{2} a_{12} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{vmatrix} < 0 \quad \text{or} \quad a_{33} [a_{11} a_{22} - \frac{1}{4} (a_{12})^2] < 0 \end{aligned} \quad (3.16)$$

Theorem 3.1: Suppose

- (i) $G'(\bar{x}_1) > 0$
- (ii) $c_4 k_4 [c_1 k_1 + a_1 G'(\bar{x}_1)] < \frac{1}{4} [a_4 c_4 F_{x_1}(\bar{x}_1, \bar{x}_4)]^2$

Then the physiological steady state $E_2 = [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$ is globally asymptotically stable, and hence the HIV-1 virions are annihilated in the CD4⁺ T cells and the blood plasma. Thus $E_2 = [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$ is a global attractor.

Proof. The result follows immediately from the negative definite criteria on A_1 , A_2 , and A_3 . It is noted that $a_{33} < 0$, and $a_{11} < 0$ if $G'(\bar{x}_1) > 0$. If a_{11} and a_{33} are both negative, then the restriction on A_3 is satisfied if condition (ii) of the theorem holds. Thus \bar{V} is negative definite and the theorem holds. □

4 Simulation results and discussion

In this section, the simulation results are presented. The computer programming code for the simulations was written in C++. Transdermal delivery was simulated as a rectangular periodic function $f(t)$ such that $f := |\text{ceil}(\sin(5t))|$.

In particular, the drug input is continuous for 6 months and off for another 6 months until HAART is discontinued. Figure 1 presents an unsuccessful HAART therapy for a hypothetical AIDS patient with the patho-physiological parametric configuration P_1 in Table 1. In this HAART scenario, the plasma HIV-1 virions (x_3) completely overwhelmed the non-infected CD4⁺ T helper cells (x_1) and the HIV-1 specific CD8⁺ cytotoxic T cells (x_4). The HIV-1 infected CD4⁺ T cells (x_2) exhibit periodic dynamics and the prognosis for the hypothetical patient is unwholesome.

Figure 2 depicts a successful HAART outcome in which the plasma HIV-1 virions (x_3) are annihilated using the hypothetical patient parameter configuration P_2 in Table 2. Also the HIV-1 infected $CD4^+$ T helper cells (x_2) are

drastically reduced to below 100 cells/ μ l. The non-infected $CD4^+$ T helper cells (x_1) are repopulated in this simulation. This outcome has been clinically observed and discussed by Ye et al. [17].

TABLE 1. Hypothetical AIDS Patient Parametric Configuration P_1

$S_1 = 400$ /day/ μ l $a_1 = 0.09$ /day/cell/ μ l $b_1 = 0.01$ /cell/ μ l $\alpha_1 = 0.5$ /day/virion/ μ l $k_1 = 0.0005$ /day/ μ l $q_1 = 0.00045$ /day/ μ l/cell $e_{10} = 0.0025$ cells/day/ μ l $x_{10} = 800$ cells/ μ l	$S_2 = 800$ /day/ μ l $a_2 = 0.03$ /day/cell/ μ l $b_2 = 0.004$ /cell/ μ l $\alpha_2 = 0.5$ /day/virion/ μ l $k_2 = 0.005$ /day/ μ l $q_2 = 0.00001$ /day/ μ l/cell $\beta_1 = 1.5$ virions/ $CD4^+$ /day $K_1 = 0.0001$ /day/ μ l $e_{20} = 0.0005$ cells/day/ μ l $\xi_2 = 0.85$ $x_{20} = 400$ cells/ μ l	$S_3 = 10$ /day/ μ l $\beta_2 = 0.0015$ virions/ $CD4^+$ /day/ μ l $\alpha_3 = 1.05$ virions/ $CD4^+$ /day $k_3 = 0.0001$ /day $e_{30} = 0.0001$ /day $\eta_1 = 0.25$ $\xi_3 = 0.001$ $x_{30} = 500$ cells/ μ l	$S_4 = 10$ /day/ μ l $a_4 = 0.35$ /day/cell/ μ l $b_4 = 0.01$ /cell/ μ l $K_2 = 0.0024$ /day/ μ l $k_4 = 0.08$ /day/ μ l $e_{40} = 0.0002$ cells/day/ μ l $\eta_2 = 0.45$ $x_{40} = 730$ cells/ μ l	$D = 3000$ units $\sigma_0 = 0.5$ mg/day $\sigma_2 = 30$ mg/day $\sigma_3 = 5$ mg/day $\lambda_0 = 5$ mg/L $\lambda_2 = 10$ mg/L $\lambda_3 = 0.025$ mg/L $x_{50} = 1500$ cells/ μ l $n = 5$
---	--	--	--	--

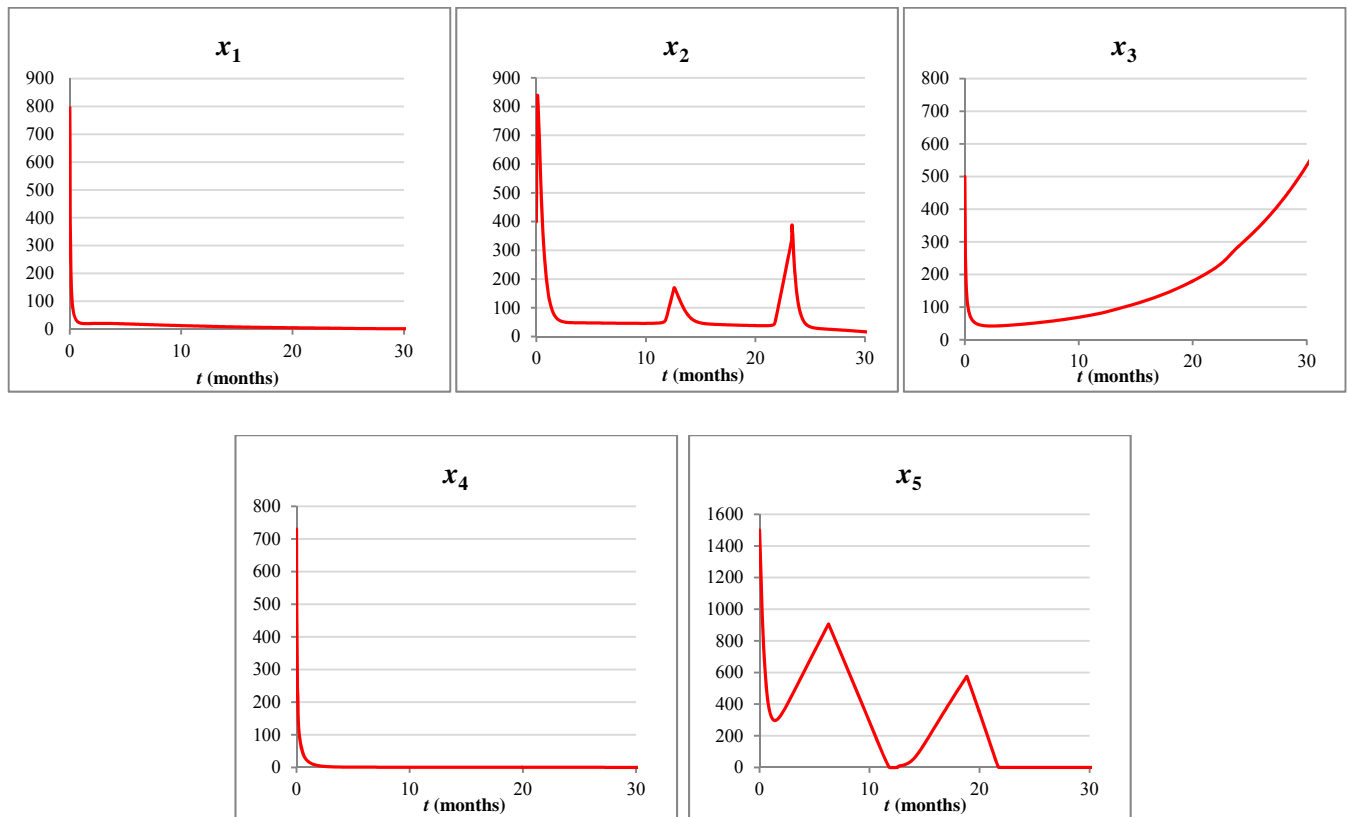
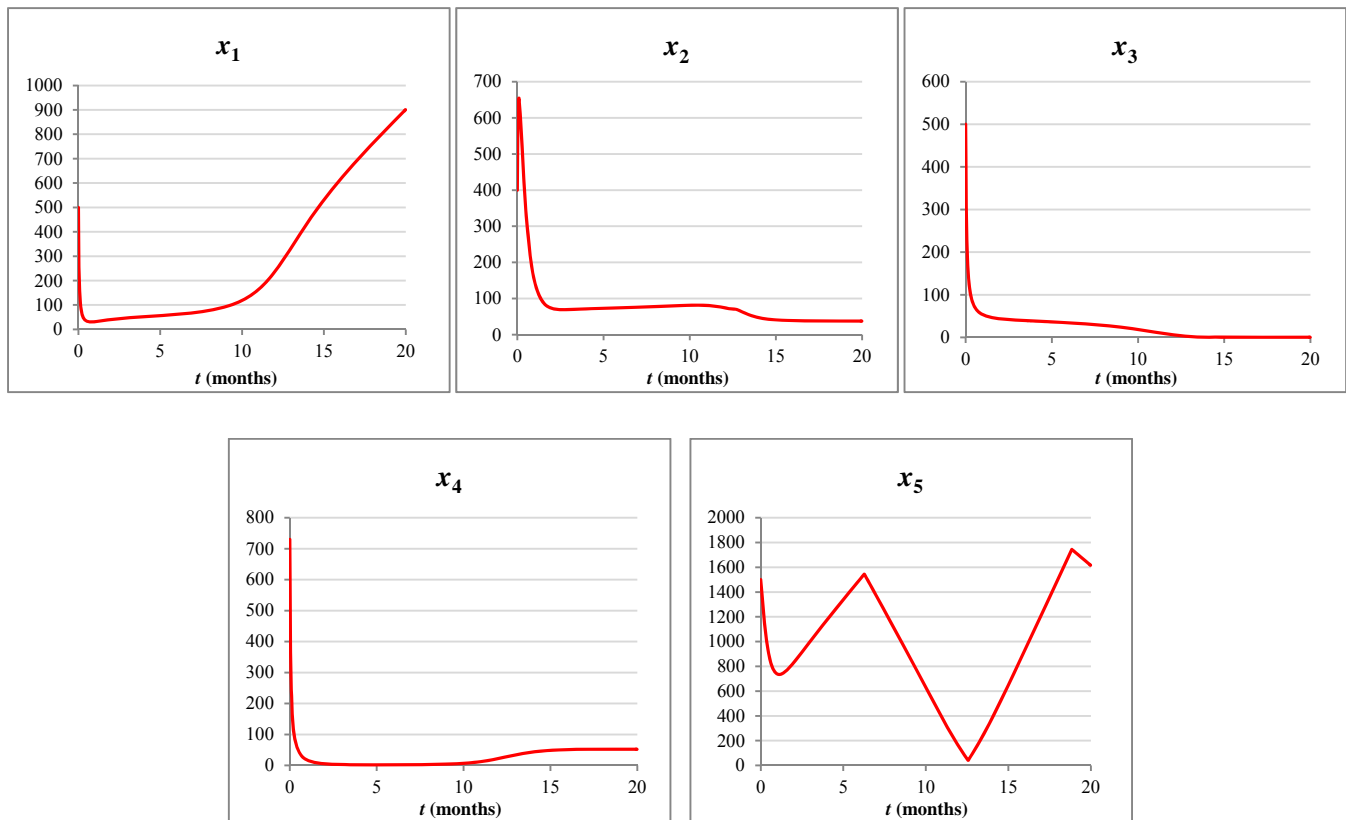


Figure 1 Simulation results using parametric configuration P_1

TABLE 2. Hypothetical AIDS Patient Parametric Configuration P_2

$S_1 = 800 \text{ /day/}\mu\text{l}$ $a_1 = 0.15 \text{ /day/cell/}\mu\text{l}$ $b_1 = 0.01 \text{ /cell/}\mu\text{l}$ $\alpha_1 = 0.5 \text{ /day/virion/}\mu\text{l}$ $k_1 = 0.0005 \text{ /day/}\mu\text{l}$ $q_1 = 0.00045 \text{ /day/}\mu\text{l/cell}$ $e_{10} = 0.0025 \text{ cells/day/}\mu\text{l}$ $x_{10} = 500 \text{ cells/}\mu\text{l}$	$S_2 = 800 \text{ /day/}\mu\text{l}$ $a_2 = 0.03 \text{ /day/cell/}\mu\text{l}$ $b_2 = 0.004 \text{ /cell/}\mu\text{l}$ $\alpha_2 = 0.5 \text{ /day/virion/}\mu\text{l}$ $k_2 = 0.005 \text{ /day/}\mu\text{l}$ $q_2 = 0.00001 \text{ /day/}\mu\text{l/cell}$ $\beta_1 = 1.5 \text{ virions/CD4}^+ \text{ /day}$ $K_1 = 0.0001 \text{ /day/}\mu\text{l}$ $e_{20} = 0.0005 \text{ cells/day/}\mu\text{l}$ $\xi_2 = 0.85$ $x_{20} = 400 \text{ cells/}\mu\text{l}$	$S_3 = 10 \text{ /day/}\mu\text{l}$ $\beta_2 = 0.0015$ $\text{virions/CD4}^+ \text{ /day/}\mu\text{l}$ $\beta_3 = 1.05 \text{ virions/CD4}^+ \text{ /day}$ $\alpha_3 = 0.027 \text{ /day/virion/}\mu\text{l}$ $k_3 = 0.0001 \text{ /day}$ $e_{30} = 0.0001 \text{ /day}$ $\eta_1 = 0.25$ $\xi_3 = 0.001$ $x_{30} = 500 \text{ cells/}\mu\text{l}$	$S_4 = 10 \text{ /day/}\mu\text{l}$ $a_4 = 0.35 \text{ /day/cell/}\mu\text{l}$ $b_4 = 0.01 \text{ /cell/}\mu\text{l}$ $K_2 = 0.0024 \text{ /day/}\mu\text{l}$ $k_4 = 0.08 \text{ /day/}\mu\text{l}$ $e_{40} = 0.0002 \text{ cells/day/}\mu\text{l}$ $\eta_2 = 0.45$ $x_{40} = 730 \text{ cells/}\mu\text{l}$	$D = 4000 \text{ units}$ $\sigma_0 = 0.5 \text{ mg/day}$ $\sigma_2 = 30 \text{ mg/day}$ $\sigma_3 = 5 \text{ mg/day}$ $\lambda_0 = 5 \text{ mg/L}$ $\lambda_2 = 10 \text{ mg/L}$ $\lambda_3 = 0.025 \text{ mg/L}$ $x_{50} = 1500 \text{ cells/}\mu\text{l}$ $n = 5$
---	--	--	---	---

Figure 2 Simulation results using parametric configuration P_2

5 Summarizing remarks

In this research, we presented a mathematical model which describes the patho-physiological dynamics of HIV-1 induced AIDS during HAART therapy. This model incorporates several physiological aspects of HIV-1 patho-physiology. These include the recruitment of virions from latent HIV-1 reservoirs (S_3) such as macrophages, microglial cells and lymphoid tissues. The model also includes autocrine and paracrine activation of $CD4^+$ and

$CD8^+$ T cells. Michaelis-Menten pharmacokinetics is used to describe the dynamics of the HAART drug in the AIDS patient. The simulations used a blend of estimated and literature based [3, 10, 13] hypothetical patient parametric configurations. The simulation results depict respectively scenarios for efficacious and non-efficacious HAART therapeutic outcomes.

The necessary conditions for existence of a plausible physiological outcome $E_2 = [\bar{x}_1, 0, 0, \bar{x}_4, \bar{x}_5]$ are

$$\begin{cases} S_1 + a_1 \bar{x}_1^2 e^{-b_1 \bar{x}_1} - k_1 \bar{x}_1 - e_{10} = 0 \\ S_2 - e_{20} = 0 \\ S_4 + a_4 \bar{x}_1 \bar{x}_4 e^{-b_4 \bar{x}_1} - k_4 \bar{x}_4 - e_{40} = 0 \\ D - \frac{\sigma_0 \bar{x}_5}{\lambda_0 + \bar{x}_5} = 0 \end{cases} \quad (5.1)$$

The sufficient conditions for successful HAART therapy and cure of AIDS are

$$\begin{cases} G'(\bar{x}_1) > 0 \\ c_4 k_4 [c_1 k_1 + a_1 G'(\bar{x}_1)] < \frac{1}{4} [a_4 c_4 F_{x_1}(\bar{x}_1, \bar{x}_4)]^2 \end{cases} \quad (5.2)$$

It is possible to refine (5.2) to

$$\bar{x}_1 = K_m^{CD8+} < 2K_m^{CD4+}$$

where K_m denotes the Michaelis-Menten constant. In a future publication, more necessary and sufficient criteria for the cure of HIV-1 induced AIDS will be presented and discussed.

6 References

- [1] S.H. Bajaria, G. Webb, D.E. Kirschner, "Predicting differential responses to structured treatment interruptions during HAART", *Bulletin of Mathematical Biology*, 66, pp. 1093–1118, 2004
- [2] M.A.L. Caetano, T. Yoneyama, "Short and long period optimization of drug doses in the treatment of AIDS", *Anais de Academia Brasileira de Ciências*, September año/vol 74, número 003, pp. 379-392, 2002
- [3] M.S. Ciupe, B.L. Bivort, D.M. Bortz, P.W. Nelson, "Estimating kinetics parameters from HIV primary infection data through the eyes of three different mathematical models", *Mathematical Biosciences* 200, 1–27, 2006
- [4] L.K. Doepel, "International HIV/AIDS trial finds continuous antiretroviral therapy superior to episodic Therapy", *NIH News*, National Institute of Health, at <http://www.nih.gov/news/pr/jan2006/niald-18.htm>, Jan 18, 2006
- [5] C. Hess et al., "HIV-1 specific CD8+ T cells with an effector phenotype and control of viral replication", *Lancet* 362, pp. 863-866, 2004
- [6] S. Jain, A. K. Tiwary, N. K. Jain, "Transdermal delivery of an anti-HIV agent using elastic liposomes: mechanism of action", *Current Drug Delivery*, vol. 3(2), pp. 57-166, 2006
- [7] X. Jin et al., "An antigenic threshold for maintaining human immunodeficiency virus type 1-specific cytotoxic T lymphocytes", *Mol. Med.* 6, pp.803-809, 2000
- [8] J. Lisziewicz and F. Lori, "Structured treatment interruptions in HIV/AIDS therapy", *Microbes and Infection*, 4, pp.207-214, 2002
- [9] S. H. Lowe, J.M. Prins, J.M. Lange, "Anti-retroviral therapy in previously untreated adults infected with the human immunodeficiency virus type 1: established and potential determinants of virological outcome", *Neth. J. Med.*, 62, pp.424-440, 2004
- [10] F. Nani and M. Jin, "Mathematical modeling and simulation of latency phase HIV-1 dynamics", *Int'l Conf. Bioinformatics and Computational Biology (BIOCAMP'10)*, vol. II, pp. 428-434, July 2010
- [11] M.A. Nowak, S. Bonhoeffer, G. M. Shaw, R.M. May, "Anti-viral drug treatment: dynamics of resistance of free virus and infected cell population", *J. Theor. Biol.* 184, pp. 203-217, 1997
- [12] G. Pantaleo, A.S. Fauci, "New concepts in the immunopathogenesis of HIV infection" *Annual Review of Immunology*, vol. 13, pp. 487-512, 1995
- [13] A. S. Perelson et al., "Decay characteristics of HIV-1 infected compartments during combination therapy", *Nature*, vol. 387, pp. 188-191, 1997
- [14] R. F. Stengel, "Mutation and control of the human immunodeficiency virus", *Mathematical Biosciences*, vol. 231, pp. 93-102, 2008
- [15] W.Y. Tan, Z. Xiang, "Some state space models of HIV pathogenesis under treatment by anti-viral drugs in HIV-infected individuals", *Mathematical Biosciences*, 156, pp.69-94, 1999
- [16] D. Wodarz, M.A. Nowak, "Specific therapy regimes could lead to long-term immunological control of HIV", *Proc. National Acad. Sci. USA*, 96, pp. 14464-14469, 1999
- [17] P. Ye, A. P. Kourtis, and D. E. Kirschner, "Reconstitution of thymic function in HIV-1 patients treated with highly active antiretroviral therapy", *Clinical Immunology*, vol. 106, pp. 95-105, 2003
- [18] G.S. Zaric, A. M. Bayoumi, ML Brandean, and DK Owens, "Effects of protease inhibitors on the spread of HIV strains, A simulation study", *Simulation*, pp.262-275, 1998

Stochastic Modelling of Tumour Immune Interactions

K.S.S. Iyer¹, Swaminathan Sankaran², and Rahul Athale³

¹International Institute of Information Technology,
Pune, India 411057.

E-mail: kss_ayer@hotmail.com

²Paul J. Hill School of Business,
University of Regina,
Regina, Canada.

E-mail: swaminathansan@gmail.com

³International Institute of Information Technology,
Pune, India 411057.

E-mail: athale.rahul@gmail.com

Abstract—*Tumour immune interaction is modelled to evaluate the tumour cell size as a stochastic time dependent model. The life of a tumour cell is assumed to be in hypothetical phases of independently distributed time duration. The analysis uses generating functions to obtain the first few moments of the tumour cell size analyzed. The first few moments are expressed as a function of time and cell proliferation kinetics including the tumour cell escape rate from immune surveillance. Numerical results are obtained and are found to be consistent with the current theory.*

Keywords: Generating functions; Immune response system; Laplace transforms; Proliferation kinetics; Tumour size modelling.

1. Introduction

The stochastic models on carcinogenesis have received considerable interest and quite a few papers have been published [[1], [2], [3], [4]]. It is important to develop stochastic models of tumour growth that include a representation of immune response. In a recent paper [5] used Monte Carlo Simulation to evaluate the tumour cell size in the presence of immune response. The resort to simulation, it seems, was mainly due to the fact that no explicit analytic solution is possible when the proliferation rates are time or age dependent. However, if the research concern lies mainly with the determination of first few moments, the problem becomes tractable. The major contribution of this paper lies in addressing this important issue, demonstrating the possibility of obtaining explicit expressions for the first few moments of tumour cell population in the presence of an active immune system. The tumour cell life time is treated here as evolving in phases. The life time from precancerous stage to dormant or dead state is divided into three phases. In fact the method of phases has already been employed in cavity radiation problems [[6]]. The layout of the paper is as follows. Section 2 describes the formulation of the model and Section 3 derives the equations satisfied by generating

functions. Section 4 derives the equations satisfied by the first two moments and their solutions. Section 5 provides numerical results under selected values of the parameters in Phases 1, 2 and 3, and explores the behaviour of tumour size over time. The last section concludes with a discussion and summary.

2. Formulation of the Model

It is well known that the immune system guards against the development of tumours and it also attempts to detect and eliminate cancerous or precancerous cells. Hence, tumour size is to be considered as a function of time and in terms of proliferation kinetics including the interaction of the immune response system with the cancerous cells. According to [7] and [8] “the tumour development can be eliminated by tumour infiltrating cytotoxic lymphocytes (TICL’s) during the avascular stage.” TICL’s interact with tumour cells and disable them from developing into proliferating malignant cells. As a result, the tumour cells either die or escape the immune surveillance and leave the primary tumour site and attempt to form tumours elsewhere. We consider the evolution of the tumour cell population according to the process of birth, (nascent tumour cell capable of proliferation), death (immune cell) and emigration (escape of tumour cell) [[9]]. Thus the life span of any tumour cell can be divided into three phases. In the first phase, the newly born tumour cell is passive and waiting to become mature enough for proliferation. $\lambda(t)\Delta t$ is the probability for the cell to pass into Phase 2 in the time interval $(t, t + dt)$. In the second phase the tumour cell is active in proliferation and the probability of a single cell to proliferate into two cells is $\eta(t)\Delta t$ in $(t, t + dt)$. In both the first and second phases the immune system can detect and form TICLs with probability $\mu_1(t)\Delta t$ in $(t, t + dt)$. In the second phase the tumour cell has a probability $\mu_2(t)\Delta t$ to pass into Phase 3, there to die or be dormant. In the third phase the tumour cell is incapable

of proliferation. The first two phases have independent time spans and that of the third phase is indefinite. The tumour cells generated in different phases are also independent and evolve with respect to time. We assume that each tumour cell necessarily goes through the three phases. At the outset we observe that it is sufficient to deal with the tumour cell population generated by one cell each in each of the three phases. This is justified by the independence of the birth and death process of each of the cells. We also assume that the tumour cell which escapes surveillance starts the cycle as an independent cell in phase 1 or phase 2 at a secondary site. It is assumed that cancer has already set in and the immune therapy is triggered by number of immune response cells namely TICL's.

3. Generating Functions of the Tumour Cell Population.

Let $X(t)$, $Y(t)$ and $Z(t)$ represent the number of tumour cells in Phases 1, 2 and 3 respectively. The population generated by the tumour cell is of the branching type when there is no escape possible for that cell from immune surveillance. However, when such escape is possible, the population generated by the escaped cells is also independent. Thus, we define two generating functions:

$$g_i(z_1, z_2, z_3, t) = E \left[z_1^{X(t)} z_2^{Y(t)} z_3^{Z(t)} / X(0) = 2 - i, Y(0) = 1 - i, v = 0 \right] \quad (1)$$

where, $i = 1, 2$

$$G(z_1, z_2, z_3, t) = E \left[z_1^{X(t)} z_2^{Y(t)} z_3^{Z(t)} / X(0) = Y(0) = Z(0) = 0, v \neq 0 \right] \quad (2)$$

where v represents the escape rate from immune surveillance and E is the expectation operator.

3.1 Relation between $G(z_1, z_2, z_3, t)$ and $g_i(z_1, z_2, z_3, t)$.

$G(z_1, z_2, z_3, t)$ is the generating function of the population generated by the escaped tumour cell. We assume the time of the first tumour cell that escapes is exponentially distributed with parameter v and that the population thus generated is independent of other cells.

$$G(z_1, z_2, z_3, t) = e^{-vt} + v \int_0^t e^{-vu} G(z_1, z_2, z_3, t - u) [g_1(z_1, z_2, z_3, t - u) + g_2(z_1, z_2, z_3, t - u)] du \quad (3)$$

The first term represents the probability that the cell does not escape in $(0, t)$. The second term represents the

probability of a tumour cell escaping immune response in $(u, u + du)$ with probability $e^{-vu}vdu$. Assuming the cell is in phase 1 or 2 it generates a population during $t - u$.

The integral Equation (3) can be solved and we obtain,

$$G(z_1, z_2, z_3, t) = exp \left[-v \int_0^t \left\{ 1 - \sum_{i=1}^2 g_i(z_1, z_2, z_3, u) \right\} du \right] \quad (4)$$

3.2 Derivation of equations governing $g_i(z_1, z_2, z_3, t)$.

We now go into deriving equations for g_1, g_2 and g_3

We obtain the differential equation satisfied by g_1 by analysing in the time interval $(0, \Delta t)$ for Phase 1 [[10]]. At $t = 0$, we have a newly born tumour cell and it can:

- 1) Move into proliferation Phase 2 in $(0, \Delta t)$ with probability $\lambda \Delta t$;
- 2) Be detected by immune response and move to Phase 3 with probability $(\mu_1 + \mu_2) \Delta t$;
- 3) Remain as it is in the same state with probability $[1 - (\lambda + \mu_1 + \mu_2)] \Delta t$.

Combining these events we can write, with $\Delta t \rightarrow 0$

$$\partial g_1(z_1, z_2, z_3, t) / \partial t = -(\lambda + \mu_1 + \mu_2)g_1 + \lambda g_2 + (\mu_1 + \mu_2)g_3 \quad (5)$$

In the case of Phase 2, at $t = 0$ we assume that there is a tumour cell which can actually proliferate. The following events can then happen in Phase 2 in the time $(0, \Delta t)$. It can:

- 1) Move straight into Phase 3 with probability $\mu_2 \Delta t$
- 2) Move into proliferation and can split into 2 tumour cells with probability $\eta \Delta t$
- 3) Be detected by immune response and move into Phase 3 with probability $\mu_1 \Delta t$
- 4) Remain as it is in the same state with probability $[1 - (\mu_1 + \mu_2 + \eta)] \Delta t$

Combining these events we can write, with $\Delta t \rightarrow 0$

$$\frac{\partial g_2}{\partial t} = -(\mu_1 + \mu_2 + \eta)g_2 + 2\eta g_1 + (\mu_1 + \mu_2)g_3 \quad (6)$$

In view of our assumption that cells in Phase 3 have zero proliferation rates the generating function g_3 is independent of z_1 and z_2 and can be evaluated explicitly as:

$$g_3(z_1, z_2, z_3, t) = 1 + (z_3 - 1)e^{-(\mu_1 + \mu_2)t} \quad (7)$$

It is rather difficult to solve for g_1 and g_2 explicitly. However, the moments of $X(t), Y(t)$ and $Z(t)$ can be evaluated

4. Moments of the tumour cell population.

We introduce the first two moments of the tumour cell population by $N_k^i(t), M_k^{i,j}(t), N^i(t), M^{i,j}(t)$ where $N_k^i(t)$ and $N^i(t)$ are the first moments of the cell population

considering cell escape rates $v = 0$ and $v \neq 0$ respectively, and $M_k^{i,j}(t)$ and $M^{i,j}(t)$ are the corresponding second moments. It is known

$$N_k^i(t) = \left. \frac{\partial g_k}{\partial z_i} \right|_{z_1=z_2=z_3=1} \tag{8}$$

$$N^i(t) = \left. \frac{\partial G}{\partial z_i} \right|_{z_1=z_2=z_3=1} \tag{9}$$

$$M_k^{i,j}(t) = \left. \frac{\partial^2 g_k}{\partial z_i \partial z_j} \right|_{z_1=z_2=z_3=1} \tag{10}$$

$$M^{i,j}(t) = \left. \frac{\partial^2 G}{\partial z_i \partial z_j} \right|_{z_1=z_2=z_3=1} \tag{11}$$

We first connect $N^i(t)$ and $N_k^i(t)$ From Equation (2) differentiating both sides

$$N^i(t) = v \int_0^t \sum_{k=1,2} N_k^i(u) du \tag{12}$$

Also by differentiating twice Equation (2) we get

$$M^{i,j}(t) = v \int_0^t \sum_{k=1}^2 M_k^{i,j}(u) du + N^i(t)N^j(t) \tag{13}$$

We now differentiate Equation (3) and Equation (4) to obtain

$$\frac{\partial N_1^i(t)}{\partial t} = -(\lambda + \mu_1 + \mu_2)N_1^i + \lambda N_2^i |_{i=1,2} \tag{14}$$

$$\frac{\partial N_1^3(t)}{\partial t} = -(\lambda + \mu_1 + \mu_2)N_1^3 + \lambda N_2^3 + (\mu_1 + \mu_2)e^{-(\mu_1 + \mu_2)t} \tag{15}$$

$$\frac{\partial N_2^i(t)}{\partial t} = -(\mu_1 + \mu_2 + \eta)N_2^i + 2\eta N_2^i |_{i=1,2} \tag{16}$$

$$\frac{\partial N_2^3(t)}{\partial t} = -(\mu_1 + \mu_2 + \eta)N_2^3 + 2\eta N_1^3 + (\mu_1 + \mu_2)e^{-(\mu_1 + \mu_2)t} \tag{17}$$

With initial conditions

$$N_1^1(0) = N_2^2(0) = N_3^3(0) = 1 \tag{19}$$

$$N_1^2(0) = N_1^3(0) = N_2^1(0) = N_2^3(0) = 0. \tag{20}$$

Solving the system of Equations (14–17) using Laplace

transforms, we get

$$N_1^1(t) = \frac{\alpha + a}{\alpha - \beta} e^{\alpha t} + \frac{\beta + a}{\beta - \alpha} e^{\beta t} \tag{21}$$

$$N_1^2(t) = \frac{\lambda}{(\alpha - \beta)} [e^{\alpha t} - e^{\beta t}] \tag{22}$$

$$N_1^3(t) = \frac{(a - c)\mu_1 + c\lambda}{(\alpha + c)(\beta + c)} e^{-ct} + \frac{(a + \alpha)\mu_1 + c\lambda}{(\alpha + c)(\alpha - \beta)} e^{\alpha t} + \tag{23}$$

$$\frac{(a + \beta)\mu_1 + c\lambda}{(\beta + c)(\beta - \alpha)} e^{\beta t}$$

$$N_2^1(t) = \frac{2\eta}{(\alpha - \beta)} [e^{\alpha t} - e^{\beta t}] \tag{24}$$

$$N_2^2(t) = \frac{\alpha + b}{(\alpha - \beta)} e^{\alpha t} + \frac{\beta + b}{(\beta - \alpha)} e^{\beta t} \tag{25}$$

$$N_2^3(t) = \frac{(b - c)c + 2\eta\mu_1}{(c + \alpha)(c + \beta)} e^{-ct} + \frac{(\alpha + b)c + 2\eta\mu_1}{(\alpha + c)(\alpha - \beta)} e^{\alpha t} + \tag{26}$$

$$\frac{(\beta + b)c + 2\eta\mu_1}{(\beta + c)(\beta - \alpha)} e^{\beta t}$$

Where $a = \mu_1 + \mu_2 + \eta$, $b = \lambda + \mu_1 + \mu_2$, $c = \mu_1 + \mu_2$ and α and β are the roots of the equation

$$S^2 + S(a + b) + ab - 2\eta\lambda = 0 \tag{27}$$

Size of the tumour at time t when the tumour cell escape rate v is zero is given by,

$$T(t) = \sum_{i=1}^3 \int_0^t [N_1^i(t') + N_2^i(t') + N_3^i(t')] dt' \tag{28}$$

Size of tumour at time t when the tumour cell escape rate v is not zero is given by,

$$T_e(t) = \sum_{i=1}^3 N^i(t) = v \sum \int_0^t N_1^i(u) du \tag{29}$$

$T_e(t)$ is the size of tumour when a single tumour cell escapes the primary site and develops elsewhere.

The second moments can be obtained by differentiating Equations (5–7) successfully, we get,

$$\frac{\partial M_1^{i,j}(t)}{\partial t} = -(\lambda + \mu_1 + \mu_2)M_1^{i,j} + \lambda M_2^{i,j} + (\mu_1 + \mu_2)M_3^{i,j} \tag{30}$$

$$\frac{\partial M_2^{i,j}(t)}{\partial t} = -(\eta + \mu_1 + \mu_2)M_2^{i,j} + 2\eta M_1^{i,j} + (\mu_1 + \mu_2)M_3^{i,j} \tag{31}$$

$$\frac{\partial M_3^{i,j}(t)}{\partial t} = -(\mu_1 + \mu_2)M_3^{i,j}(t) \tag{32}$$

Since we are interested in the size of the tumour, we refrain from giving the solutions though the above equations

can be solved by Laplace transform. However as $t \rightarrow \infty$ we can obtain the steady state expression for $N_j^i(t)$.

$$N_1^1(\infty) = \frac{\mu_1 + \mu_2 + \eta}{ab - 2\eta\lambda} \quad (33)$$

$$N_1^2(\infty) = \frac{\lambda}{ab - 2\eta\lambda} \quad (34)$$

$$N_1^3(\infty) = \frac{\lambda + \eta + \mu_1 + \mu_2}{ab - 2\eta\lambda} \quad (35)$$

$$N_2^1(\infty) = \frac{2\eta}{ab - 2\eta\lambda} \quad (36)$$

$$N_2^2(\infty) = \frac{\lambda + \mu_1 + \mu_2}{ab - 2\eta\lambda} \quad (37)$$

$$N_2^3(\infty) = \frac{\lambda + \eta + \mu_1 + \mu_2}{ab - 2\eta\lambda} \quad (38)$$

5. Exploratory Numerical Results.

We proceed to evaluate the tumour size numerically for different values of $\lambda, \eta, \mu_1,$ and μ_2 . The tumour cell population or size is shown as two variables: $T(t)$ and $T_e(t)$. $T(t)$ is the size at the primary host site and $T_e(t)$ is the size of the population generated by the escaped cell at another secondary site.

In a series of graphs we plot $T(t)$ and $T_e(t)$ for an exploratory set of values of $\lambda, \eta, \mu_1,$ and μ_2 gathered from prior results in the literature. We note that they follow a piecewise Gompertz curve pattern as found by Boondreck et. al. (2006) through Montecarlo simulation. The graphs presented here in Figure 1, Figure 2, and Figure 3 show the smoothed Gompertz curve fit for these results. Next we also calculated and tabulated the number of proliferated cells $P(t)$ over time as the parameters are varied¹. Finally, to facilitate comparison with the simulated results in [5], using non-linear regression, we fitted Gompertz curves for the values obtained by us analytically using the exploratory set of parameter values. These graphs are shown in Figure 4, Figure 5 and Figure 6. Our analytical results from this extended model, permitting both attachment by the immune response system to incapacitate the cancer cells and also escape from that system to another site to proliferate, confirm the growth pattern of cells and tumour size over time, derived by them through simulation. In the next and concluding section we discuss these results.

6. Summary and Discussion.

First it is interesting to note that for fixed values of the probabilities of the cell moving directly from either Phase 1 or 2 to Phase 3, i.e., for fixed μ_1 and μ_2 to become inactive and for fixed values of the infection rate or probability λ , as the proliferation rate η increases the analytical results from the model show that the number of proliferated cells at the

¹Copies of these tables can be obtained as Excel files by e-mailing the request to either one of the authors.

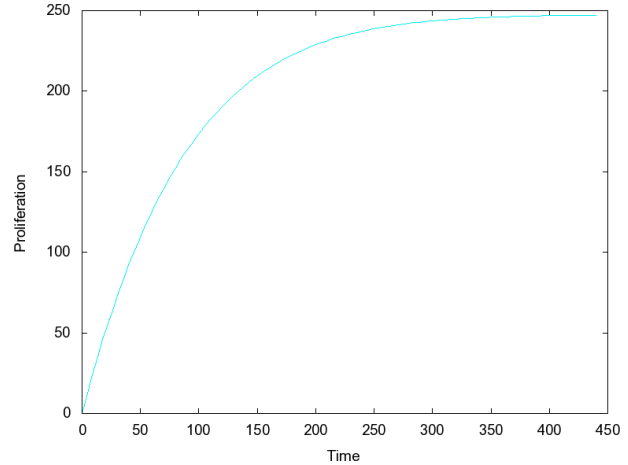


Fig. 1: Number of cells at primary site when escape rate $v = 0, \lambda = 0.5, \eta = 0.4, \mu_1 = 0.15, \mu_2 = 0.05$

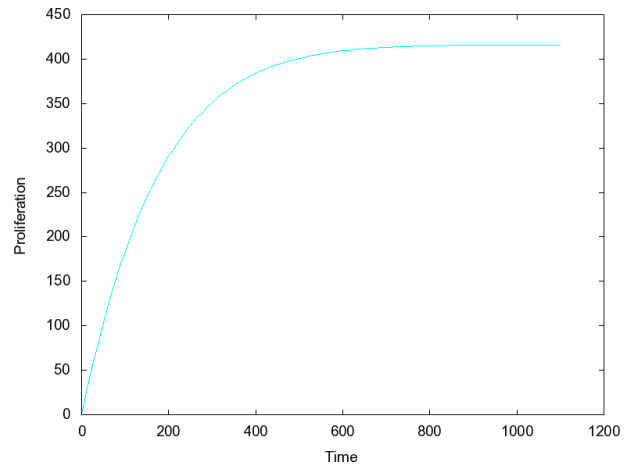


Fig. 2: Number of cells at primary site when escape rate $v = 0, \lambda = 0.5, \eta = 0.7, \mu_1 = 0.2, \mu_2 = 0.08$

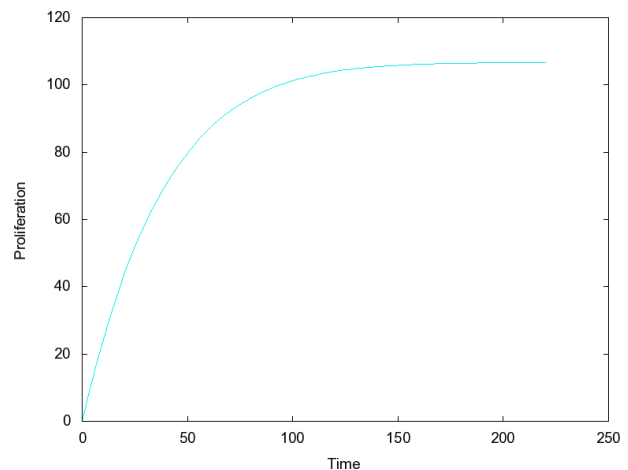


Fig. 3: Number of cells at primary site when escape rate $v = 0, \lambda = 0.1, \eta = 0.1, \mu_1 = 0.02, \mu_2 = 0.03$

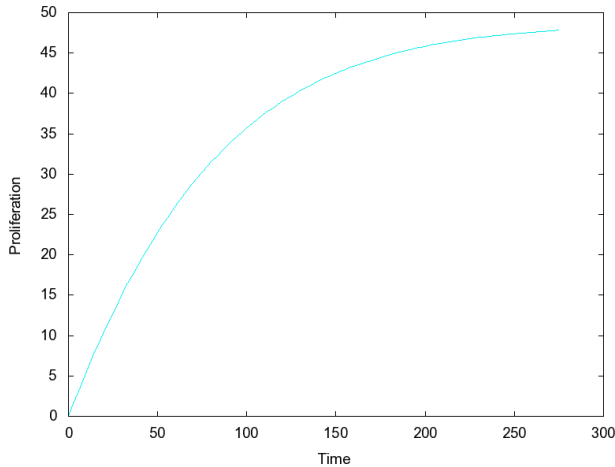


Fig. 4: Number of cells at secondary site when escape rate $\nu = 0, 1, \lambda = 0.5, \eta = 0.4, \mu_1 = 0.15, \mu_2 = 0.05$

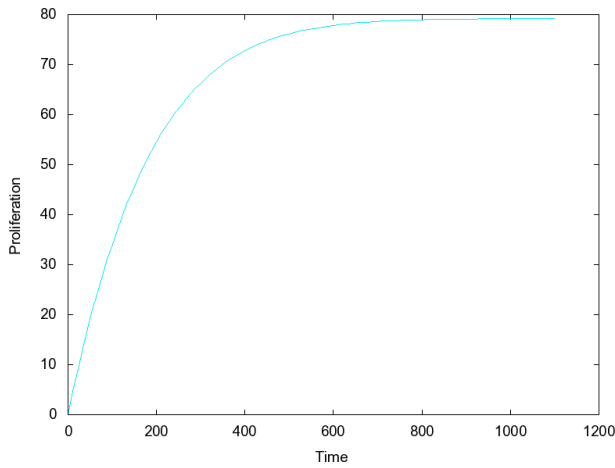


Fig. 5: Number of cells at primary site when escape rate $\nu = 0.1, \lambda = 0.5, \eta = 0.7, \mu_1 = 0.2, \mu_2 = 0.08$

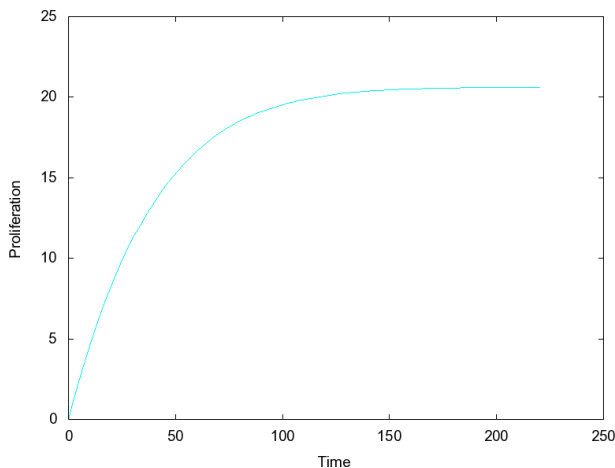


Fig. 6: Number of cells at secondary site when escape rate $\nu = 0.1, \lambda = 0.1, \eta = 0.1, \mu_1 = 0.02, \mu_2 = 0.03$

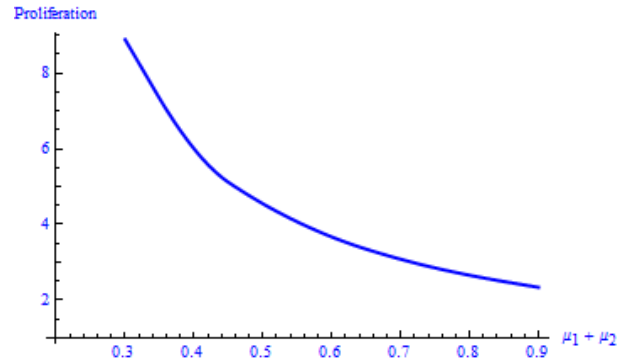


Fig. 7: Number of cells at secondary site when escape rate $\lambda = 0.2, \eta = 0.3, \mu_1 + \mu_2$ changing from 0.3 to 0.9

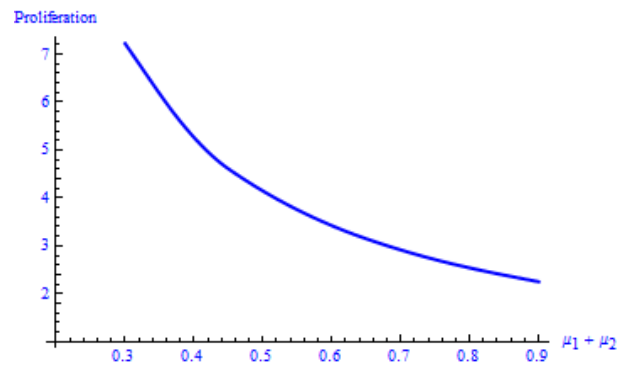


Fig. 8: Number of cells at secondary site when escape rate $\lambda = 0.05, \eta = 0.3, \mu_1 + \mu_2$ changing from 0.3 to 0.9

secondary site increases but at a decreasing rate. The rate of decrease increases as $\mu_1 + \mu_2$ increases. Furthermore, the number of these proliferated cells converges to an asymptotic limit after the expiration of a period of time. Again this convergence time is not uniform. It is reached rapidly over time for larger values of $\mu_1 + \mu_2$, and slowly for smaller values. We see a similar pattern for the total tumour size $T(t)$ when the escape rate of an infected cell is zero and for the tumour size $T_e(t)$ when that escape rate ν is positive. One would normally expect when ν increases both $T(t)$ and $T_e(t)$ would increase all else being held constant, and $T_e(t)$ would increase more rapidly. All these are borne out by the graphs in the respective figures when these values are obtained purely from our analytical results for exploratory values of the rate parameters.

Following [5] we investigated the shape of the various curves for $P(t), T(t)$ and $T_e(t)$ as functions of time to see if they fit the Gompertz curve shape obtained with their simulated data. The figures show that a Gompertz curve fit obtained through non-linear regression from SAS fit them remarkably well. Further analysis shows that fixed values of other rate parameters, the time required for the doubling of the number of proliferated cells increases at an increasing

rate, i.e., at greater speed as the proliferation rate η increases.

We also checked the paper [11] and collected values of μ_1 and μ_2 . The graphs in Figure 7 and 8 represent the proliferation of tumour cell in the absence of TICL's and for different rates of disabling of malignant cells by TICL. It can be seen from the graphs as the rate increases the proliferation decreases. This could help in deciding the level of therapy for controlling the malignant cells. Work is in progress to prepare a table for practitioners to make use of the table.

These results have some practical implications. The major one is that any treatment that can either directly reduce, or provide more time for the body's immune response system to attack and slow down, the proliferation rate would be beneficial to the patient and slow down the spread of cancer. The same technique can be used to find the latent cell population in HIV.

7. Acknowledgements

The authors gratefully acknowledge administrative support from the International Institute of Information Technology, Pune, research assistance from Mr. Pranav Nagpurkar, Graduate Student in the School of Management Technology.

References

- [1] D. G. Kendall, "Birth and death process and the theory of carcinogenesis," *Biometrika*, vol. 47, pp. 13–21, 1960.
- [2] P. Armitage and R. Doll, "The age distribution of cancer and a multistage theory of carcinogenesis," *British Journal of Cancer*, vol. 8, no. 1, pp. 1–12, 1954.
- [3] W. Y. Tan, *Stochastic Models of Carcinogenesis*. New York: Marcel Dekker, 1991.
- [4] K. S. S. Iyer and V. N. Saxena, "A stochastic model for the growth of cells in cancer," *Biometrics*, vol. 26, no. 3, pp. 401–410, 1970.
- [5] A. Boondrek, Y. Lenbury, J. Wong-Ekkabut, I. M. Tang, and P. Picha, "A stochastic model of cancer growth with immune response," *Journal of the Korean Physical Society*, vol. 49, no. 4, pp. 1652–66, 2006.
- [6] S. K. Srinivasan, "Age dependent model of cavity radiation and detection," *Pramana, Journal of Physics*, vol. 27, pp. 13–21, 1986.
- [7] A. Matzavinos and J. Chaplain, "Mathematical modelling of the spatio temporal response of cytotoxic t-lymphocytes to a solid tumour," *Math.med. and Biology*, vol. 21, no. 1, pp. 1–12, 2004.
- [8] —, "Travelling wave analysis of a model of immune response," *Comptes Rendus Biologies*, vol. 327, p. 995, 2004.
- [9] D. G. Kendall, "Stochastic processes and population growth," *J. R. Statist. Soc.*, vol. B11, pp. 230–63, 1949.
- [10] M. S. Bartlett, *Stochastic Process*. Cambridge: Cambridge University Press, 1966.
- [11] F. M. S. and W. T. M., "Modeling the effects of a simple immune system and immunodeficiency on the dynamics of conjointly growing tumor and normal cells," *Int J Biol Sci*, vol. 7, pp. 700–707, 2011.

A Gaussian Packing Model For Phasing in Macromolecular Crystallography

Yan Yan¹ and Gregory S. Chirikjian¹

¹The Department of Mechanical Engineering, Johns Hopkins University, Baltimore, MD, USA

Abstract— *Molecular replacement (MR) is a computational method that is frequently used to obtain phase information for a unit cell packed with a macromolecule of unknown structure. The goal of MR searches is to place a homologous/similar molecule in the unit cell so as to maximize the correlation with x-ray diffraction data. MR software packages typically perform rotation and translation searches separately. This works quite well for single-domain proteins. However, for multi-domain structures and complexes, computational requirements can become prohibitive and the desired peaks can become hidden in a noisy landscape. The main contribution of our approach is that computationally expensive MR searches in continuous configuration space are replaced by a search on a relatively small discrete set of candidate packing arrangements of a multi-rigid-body model. These candidate arrangements are generated by minimizing a Gaussian-based potential function that forces the model conformations to separate from each other and not overlap within the unit cell. This is done before computing Patterson correlations rather than only performing collision checks when evaluating the feasibility of peaks. The list of feasible arrangements is short because collision-free packing requirement together with unit cell symmetry and geometry impose strong constraints. After computing Patterson correlations of the candidate arrangements, an even shorter list can be obtained using 10 candidates with highest correlations. In numerical trials, we found that a candidate from the feasible set is usually similar to the arrangement of the target structure within the unit cell. To further improve the accuracy, a Rapidly-exploring Random Tree (RRT) can be applied in the neighborhood of this packing arrangement. Our approach is demonstrated with multi-domain models in silico for 2D, with ellipses (ellipsoids in 2D) representing both the domains of the model and target structures. Configurations are defined by sets of angles between the ellipses. Our results show that an approximate configuration can be found with the mean absolute error less than 3 degrees.*

Keywords: X-ray crystallography, molecular replacement, multi-domain system, packing model, Gaussian function

1. Introduction

The field of structural biology is concerned with characterizing the shape, composition, flexibility, and motion

of biological macromolecules and the complexes that they form. An ultimate goal of this field is to link these properties with the function of macromolecular structures, in the hope of better understanding biological phenomena and designing new drugs.

Here we review some of the issues involved in translating experimental data into 3D structures in the context of protein crystallography. Macromolecular X-ray crystallography (MX) has been the most used method for determining protein structures and associated complexes. It works very well for simple proteins that can be described as single rigid-bodies (called domains). This is because information about the shape of 75,000 previously solved structures in the Protein Data Bank (many of which are single-domain structures) can be used to augment new MX experimental information to gain a complete picture.

However, a challenge to MX arises in interpreting X-ray diffraction patterns for crystals composed of multi-domain systems. This is because even when a multi-domain structure has been solved previously, its overall shape may vary widely from a new version of the structure with, for example, a bound drug. In this case, a widely used computational method called the molecular replacement method (MR), which has been highly successful for single-domain proteins, becomes combinatorially intractable due to the large number of degrees of freedom in multi-domain systems. We present a new method for phasing based on geometric packing that can serve as an alternative to MR. Decades ago, the concept of building models of crystallographic unit cells to phase crystallographic data was explored in the context of small molecules [1], [2], [3]. But to our knowledge, this approach has not been pursued and is virtually unknown in the context of multi-domain macromolecular crystallography, and “phasing by packing” therefore represents a very different way of approaching the problem than MR.

The remainder of this paper is structured as follows. The mathematical aspects of the MR method for single-domain proteins is reviewed first. Then the multi-domain phase problem is formulated. Finally, we present our initial findings that diffraction patterns for multi-domain systems can be phased using our new “phasing by packing” method.

2. Essentials of Macromolecular X-Ray Crystallography (MX)

A biological macromolecule is a large collection of atomic nuclei that are stabilized through a combination of covalent bonds, hydrogen bonds, and hydrophobicity. A traditional goal in structural biology is to obtain the Cartesian coordinates of all atoms in a rigid single-domain protein.

Let $\mathbf{x}_i = (x_i, y_i, z_i)$ denote the Cartesian coordinates of the i^{th} of n atoms in a single-domain protein structure, and let $\rho_i(\mathbf{x})$ be the electron density of that atom in a reference frame centered on it. Due to thermal motions, the electron density of each of these atomic nuclei can be treated as a Gaussian distribution. The density of the whole structure is then of the form

$$f(\mathbf{x}) = \sum_{i=1}^n \rho_i(\mathbf{x} - \mathbf{x}_i). \quad (1)$$

The coordinates $\{\mathbf{x}_i\}$ are typically given either in a reference frame attached to a crystallographic unit cell, or to the center of mass of the protein.

MX does not provide $f(\mathbf{x})$ directly. Rather, it provides partial information about $f(\mathbf{x})$. The goal is then to computationally obtain $f(\mathbf{x})$ and fit an atomic model to it, thereby extracting the coordinates $\{\mathbf{x}_i\}$. A macromolecular crystal is composed of *unit cells* that have a discrete symmetry group, Γ . This symmetry group divides \mathbb{R}^3 into unit cells, $U \cong \Gamma \backslash \mathbb{R}^3$, and also describes how copies of the density $f(\mathbf{x})$ are located within the unit cell. The whole group Γ can be generated by translating unit cells and moving within the unit cell using generators $\{\gamma_1, \dots, \gamma_m\}$. These form a subgroup of Γ , which is in turn a subgroup of the group of rigid-body motions, $SE(3)$, which will be denoted here as G .

The result of an MX experiment is a diffraction pattern. This is the magnitude of the Fourier transform of the full contents of the crystallographic unit cell. Mathematically, this is written for a single-domain protein as

$$\hat{P}(g; \mathbf{k}) = \left| \mathcal{F} \left(\sum_{j=0}^{m-1} f((\gamma_j \circ g)^{-1} \cdot \mathbf{x}) \right) \right|, \quad (2)$$

where $|\cdot|$ denotes the modulus of a complex number, $c = a + ib = |c|e^{i\phi}$. Our reason for using the notation $\hat{P}(g; \mathbf{k})$ will be explained shortly. Here $g \in G$ is the unknown pose of the protein that is sought, and \circ is the group operation for both G and Γ . In particular, it is well-known in robotics that each rigid-body motion consists of a rotation-translation pair $g = (R, \mathbf{t})$, and the composition of any two rigid-body motions g_1 and g_2 defines the operation \circ :

$$g_1 \circ g_2 = (R_1, \mathbf{t}_1) \circ (R_2, \mathbf{t}_2) = (R_1 R_2, R_1 \mathbf{t}_2 + \mathbf{t}_1). \quad (3)$$

Given that $g = (R, \mathbf{t}) \in G$ is a rotation-translation pair, its action on \mathbb{R}^3 is defined by

$$g \cdot \mathbf{x} = R\mathbf{x} + \mathbf{t}. \quad (4)$$

Then the density of a collection of single-domain proteins in the unit cell for $j = 0, \dots, m-1$ will be $\sum_{i=0}^{m-1} f((\gamma_i \circ g)^{-1} \cdot \mathbf{x})$.

The difficulty in extracting $f(\mathbf{x})$ from the MX data is that this measurement folds in both information about $f(\mathbf{x})$ and the symmetry group Γ , and kills the phase information, $\phi(\mathbf{k})$, without which $f(\mathbf{x})$ cannot be recovered by inverse Fourier transform. Moreover, there is an unknown $g \in G$ that describes how each symmetry-related copy of $f(\mathbf{x})$ sits in the unit cell. Single-domain MR is mostly about finding the unknown g , and most commonly this is done by dividing the search into rotational and translational parts.

The number of proteins in a unit cell, the crystallographic space group, Γ , and aspect ratios of the unit cell can be taken as known inputs in MR computations, since they are all provided by experimental observation. And from homology modeling, it is often possible to have reliable estimates of the shape of each domain in a multi-domain protein. What remains unknown are the relative positions and orientations of these domains and the overall position and orientation of the symmetry-related copies of the proteins within the unit cell.

Once these are known, a model of the unit cell can be constructed and used as an initial phasing model that can be combined with the X-ray diffraction data. This is, in essence, the molecular replacement approach that is now more than half a century old [4], [5]. Many powerful software packages for molecular replacement include those described in [6], [7]. Typically these perform rotation searches first, followed by translation searches.

3. The Multi-Domain Molecular Replacement Method (NMR)

The molecular replacement (MR) method, originally developed in the 1960s [4], [10], [11], [12] is a computational method for phasing X-ray diffraction data for biomolecular structures. It has been integrated into crystallographic structure determination codes [6], [14]. Though MR has been wildly successful for single-domain proteins, significant issues arise when using MR for multi-domain proteins and complexes.

Currently two major computational paradigms exist for phasing of X-ray diffraction patterns of multi-domain proteins: (1) use existing software packages to obtain candidate peaks in the rotation function for individual domains separately, then solve for the translation function [13]; (2) attempt to morph multi-domain candidate models that contain their full “6N” degrees of freedom and iteratively refine those models [8]. Both methods suffer from different aspects of the “curse of dimensionality,” which we seek to circumvent using a combination of our initial results reported in [9] and new approaches based on advanced mathematical concepts that are new to the crystallography community.

Consider a multi-domain protein or complex consisting of N rigid bodies. If $f_i(\mathbf{x})$ denotes the density of the i^{th} body, then the density of the whole complex will be of the form $f(\mathbf{x}) = \sum_{i=1}^N f_i(g_i^{-1} \cdot \mathbf{x})$ where $g_i = (R_i, \mathbf{t}_i)$ is a rigid-body motion consisting of a rotation-translation pair and $g_i^{-1} \cdot \mathbf{x} = R_i^T(\mathbf{x} - \mathbf{t}_i)$. These motions are the unknowns in our problem.

If m identical copies of this complex are arranged symmetrically in a unit cell by symmetry operators $\gamma_j = (Q_j, \mathbf{a}_j) \in \Gamma$ (which is the group consisting of n discrete rigid-body motions that are known a priori from the crystal symmetry and geometry), an X-ray diffraction experiment provides the magnitude (without phase) of the Fourier transform of $\sum_{j=0}^{m-1} f(\gamma_j^{-1} \cdot \mathbf{x})$. In contrast, the model density for a single domain and its symmetry mates is $\sum_{j=0}^{m-1} f_i(h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x})$ where h_i is the candidate position and orientation. In traditional MR, the Fourier transform of the Patterson functions, $\hat{P}(g_1, \dots, g_N; \mathbf{k}) = \mathcal{F}[P(g_1, \dots, g_N; \mathbf{x})]$ and $\hat{p}_i(h_i; \mathbf{k}) = \mathcal{F}[p_i(h_i; \mathbf{x})]$, that correspond to these densities and their correlation are respectively

$$\hat{P}(g_1, \dots, g_N; \mathbf{k}) = \left| \sum_{j=0}^{m-1} \mathcal{F}[f(\gamma_j^{-1} \cdot \mathbf{x})] \right|, \quad (5)$$

$$\hat{p}_i(h_i; \mathbf{k}) = \left| \sum_{j=0}^{m-1} \mathcal{F}[f_i(h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x})] \right|, \quad (6)$$

$$c(h_i) = \int_{\mathbf{x} \in \mathcal{C}} P(g_1, \dots, g_N; \mathbf{x}) p_i(h_i; \mathbf{x}) d\mathbf{x} \quad (7)$$

where the Fourier transform \mathcal{F} converts a function of spatial position, \mathbf{x} , into a function of spatial frequency, \mathbf{k} . The real-space Pattersons themselves are obtained by applying the inverse Fourier transform. Of the quantities in (5)-(7), $\hat{P}(g_1, \dots, g_N; \mathbf{k})$ comes from the experiment (this is the multi-domain version of (2)), and $\hat{p}_i(h_i; \mathbf{k})$ and $c(h_i)$ are computed. Here \mathcal{C} is the unit cell and in MR searches the hope is that peaks in the function $c(\cdot)$ correspond to $h_i = g_i$. The difficulty is that, unlike the single domain case, in the multi-domain case P depends on many g_j 's that all interact with each other. Therefore, peaks in this rotational correlation function do not necessarily correspond to good overall matches.

4. PHASING BY PACKING

Instead of running traditional MR searches on domain orientation or full conformation, we propose to construct packing models for the multi-domain systems of interest. This will generate candidate sets of motions $\{h_1, \dots, h_N\}$ that can then be used to construct a model of $P(h_1, \dots, h_N; \mathbf{x})$ rather than $p_i(h_i; \mathbf{x})$. If $P(h_1, \dots, h_N; \mathbf{x})$ and $P(g_1, \dots, g_N; \mathbf{x})$ match well to each other, then that is a much stronger indication that $h_i = g_i$ than having high correlations between $p_i(h_i; \mathbf{x})$ and $P(g_1, \dots, g_N; \mathbf{x})$.

In this approach, an ellipsoid or a combination of several ellipsoids are used to approximate the convex hull of each

domain of protein structures. A multi-ellipsoid-shaped model is built for a multi-domain structure and packed in space to detect feasible packing arrangements. The most important crystal packing constraint is that protein macromolecules do not collide with (or insert into) each other. With high protein-water volume ratio in crystals, they usually have to "smartly" close packed. Since the allowable motion is severely restricted, we can find a discrete candidate set to represent all the feasible packing arrangements. Noticing Gaussian functions have infinite tails, a Gaussian-based cost function (GCF) is constructed to evaluate the level of overlapping (or closeness) among ellipsoids with each ellipsoid represented by a Gaussian function or a mixture of Gaussian functions. The candidate packing arrangements can be obtained by minimizing the GCF to force the packing model to separate from each other and not overlap within the unit cell.

The shape of an ellipsoid can be captured by equidensity contours of a Gaussian function with the mean located at the ellipsoid center and the covariance matrix related to its semi-axis lengths. An arbitrarily oriented ellipsoid in \mathbb{R}^n can be described as

$$(\mathbf{x} - \mu)^T R^T A R (\mathbf{x} - \mu) = 1, \quad (8)$$

where R is the rotation matrix, and $A = \text{diag}[1/a_1^2, 1/a_2^2, \dots, 1/a_n^2]$, with a_i denoting the semi-axis length of the ellipsoid. Compared with a Gaussian function in \mathbb{R}^n ,

$$\rho(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right), \quad (9)$$

we can see that when $\Sigma^{-1} = R^T A R$, the equidensity contours of the Gaussian function are ellipsoids with semi-axis lengths $k \cdot a_1, k \cdot a_2, \dots, k \cdot a_n$, where $k \in \mathbb{R}_{\geq 0}$. To more accurately capture the shape of the ellipsoid with semi-axis lengths a_1, a_2, \dots, a_n , we want the Gaussian function to have high and steady value inside the ellipsoid region and a quick drop outside it. We note that it is not necessary to eliminate the tail outside the ellipsoid since the interaction among the tails can help push the ellipsoids away from each other. We use a Gaussian mixture function $\psi(\mathbf{x}; \mathbf{a}, \mathbf{b})$, i.e.,

$$\psi(\mathbf{x}; \mathbf{a}, \mathbf{b}) = \sum_{i=1}^n \frac{a_i}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{b_i}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right), \quad (10)$$

instead of a single Gaussian $\rho(\mathbf{x})$ to approximate an ellipsoid. In the 1D case in Fig. 1, with both variances $\sigma = 1$, we can see that compared to the single Gaussian $\frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, the Gaussian mixture function with $\mathbf{a} = 0.44 \cdot [3, -1]$ and $\mathbf{b} = 1.16 \cdot [1, 3]$, i.e.,

$$\psi(\mathbf{x}; \mathbf{a}, \mathbf{b}) = \frac{1.32}{\sqrt{2\pi}} \exp(-0.58x^2) - \frac{0.44}{\sqrt{2\pi}} \exp(-1.73x^2), \quad (11)$$

has a flatter top and faster decay tails.

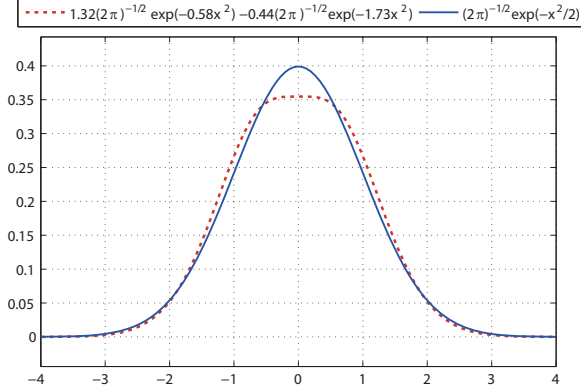


Fig. 1

THE COMPARISON BETWEEN A SINGLE GAUSSIAN WITH A MIXTURE OF GAUSSIANS.

The ellipsoid model of i^{th} domain in a multi-domain structure under a symmetry group Γ can be approximated by $\psi((h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x}); \mathbf{a}, \mathbf{b})$, where h_i is rigid-body operation of the i^{th} domain and γ_j is the symmetry operator in the symmetry group Γ . Therefore we define the GCF as

$$\text{GCF}(h_1, \dots, h_N) \triangleq \int_{\mathbb{R}^n} \left[\sum_{j=0}^{m-1} \sum_{i=1}^N \psi((h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x}), \mathbf{a}, \mathbf{b}) \right]^2 d\mathbf{x}. \quad (12)$$

An advantage of Gaussian functions is that the integration of quadratic terms over \mathbb{R}^n has a closed-form expression. We derived it as follows,

$$\begin{aligned} & \int_{\mathbb{R}^n} \rho_1(\mathbf{x}; \mu_1, \Sigma_1) \rho_2(\mathbf{x}; \mu_2; \Sigma_2) d\mathbf{x} \quad (13) \\ &= \int_{\mathbb{R}^n} (2\pi)^{-n/2} (\det \Sigma_1)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma_1^{-1} (\mathbf{x} - \mu_1)\right) \\ & \quad (2\pi)^{-n/2} (\det \Sigma_2)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_2)^T \Sigma_2^{-1} (\mathbf{x} - \mu_2)\right) d\mathbf{x} \\ &= (2\pi)^{-n} (\det \Sigma_1 \det \Sigma_2)^{-1/2} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma_1^{-1} (\mathbf{x} - \mu_1) \right. \\ & \quad \left. - \frac{1}{2}(\mathbf{x} - \mu_2)^T \Sigma_2^{-1} (\mathbf{x} - \mu_2)\right) d\mathbf{x}. \\ & \text{Since } \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\mathbf{x}^T M \mathbf{x} - m^T \mathbf{x} - C\right) \quad (14) \\ & \quad = (2\pi)^{n/2} (\det M)^{-1/2} \exp\left(\frac{1}{2}m^T M^{-1} m - C\right), \end{aligned}$$

(13) can be rewritten in a closed-form as

$$\begin{aligned} & \int_{\mathbb{R}^n} \rho_1(\mathbf{x}; \mu_1, \Sigma_1) \rho_2(\mathbf{x}; \mu_2; \Sigma_2) d\mathbf{x} \quad (15) \\ &= (2\pi)^{-n} (\det \Sigma_1 \det \Sigma_2 \det(\Sigma_1^{-1} + \Sigma_2^{-1}))^{-1/2} \\ & \quad \exp\left(\frac{1}{2}(\mu_1^T \Sigma_1^{-1} + \mu_2^T \Sigma_2^{-1})(\Sigma_1^{-1} + \Sigma_2^{-1})(\Sigma_1^{-T} \mu_1 + \Sigma_2^{-T} \mu_2) \right. \\ & \quad \left. - \frac{1}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 + \mu_2^T \Sigma_2^{-1} \mu_2)\right). \end{aligned}$$

The closed-form expression of the GCF can be easily derived from (15).

The main procedures of generating candidate phasing models by packing can be described by a flowchart in Fig. 2. In the first step, we discretize the configuration space by a coarse grid, and find the configuration with the smallest GCF value inside each “configuration cell” defined by the grid. The collision-free ones of these configurations form the candidate set of packing arrangements. This discrete candidate set reduces the whole configuration space to a much shorter list. We note that with a closed-form expression, minimizing the GCF is less computationally expensive compared to calculating $c(h_i)$ in traditional MR searches (see (7)).

In the next step, we use a Fourier-based cost function (FCF), where

$$\begin{aligned} & \text{FCF}(h_1, \dots, h_N) \quad (16) \\ &= \left[\int_{\mathbf{k} \in \Omega} (\hat{P}(g_1, \dots, g_N; \mathbf{k}) - \hat{P}(h_1, \dots, h_N; \mathbf{k}))^2 d\mathbf{k} \right]^{\frac{1}{2}}, \end{aligned}$$

to sort these collision-free configurations from low to high. In our simulation, the function $f_i(\mathbf{x})$ defined in Sec. 3 are chosen to be the set indicator function for the ellipsoid representing body i . Then $\hat{P}(g_1, \dots, g_N; \mathbf{k})$ and $\hat{P}(h_1, \dots, h_N; \mathbf{k})$ are defined in (5) and (6), respectively.

Minimizing $\text{FCF}(h_1, \dots, h_N)$ is similar to finding peaks in $c(h_i)$ except that we use a multi-domain model rather than a single-domain one. After the sorting, we keep 10 configurations with lowest FCF as a candidate list. These candidates indicate high correlations with the target structure. The FCF has the rugged surface of the configuration space, so to further improve the accuracy, a stochastic sampling method—Rapidly-exploring random tree (RRT) algorithm [15] is used to minimize the FCF around the “best candidate”. The best candidate can be first chosen as the one with the lowest FCF in the set. If its FCF cannot be reduced below a threshold value C after running the RRT, we switch the best candidate to the one with the next lowest FCF.

5. EXPERIMENTAL EXAMPLE

In this section, the approach to phasing by using packing models is demonstrated in a 2D planar case, with ellipses representing both the domains of the model and target structures. All the angular parameters of the target structure are treated as being unknown, and the only priori information that we have is the magnitude of the Fourier transform of the electron density function $\hat{P}(g_1, \dots, g_N; \mathbf{k})$. Our goal is to find the closest model configuration $\{h_1, \dots, h_N\}$ with respect to the target structure $\{g_1, \dots, g_N\}$. To illustrate our approach, a multi-ellipse-shaped “rabbit” with one “face” and two “ears” is constructed as a packing model for a 3-domain structure in P1 symmetry. Since translations have no impact on the packing result in P1 symmetry, the rabbit

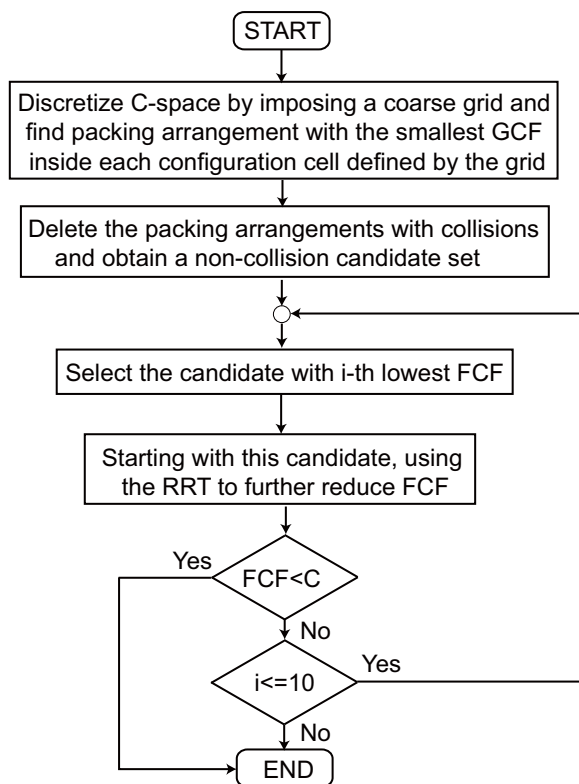


Fig. 2

FLOWCHART OF FINDING CANDIDATE PHASING MODELS BY PACKING.

model has 3 DOF— the rotations of the face, θ_1 and two ears, θ_2 and θ_3 (see the dimensions and ranges of motion in Table 1).

For the Gaussian mixture function in this 2D planar case, we use the same ratios of a_1 , a_2 and b_1 , b_2 as the 1D case in (11), i.e., $a = m_a \cdot [3, -1]$ and $b = m_b \cdot [1, 3]$. m_b^* —the optimal value of m_b , is chosen to “stretch or shrink” the Gaussian mixture function so that it can “best” represent the defined ellipse. After that, m_a^* is calculated to normalize the Gaussian mixture function with m_b^* . More specifically, we define m_b^* as

$$m_b^* = \arg \max_{m_b} |S_{\text{cand.}}(m_b)|, \quad (17)$$

where $S_{\text{cand.}}$ represents the non-collision candidate set, generated by obtaining the packing arrangements with the smallest GCF value inside each configuration box defined by the grid, and deleting the collision ones afterwards. $|S_{\text{cand.}}(m_b)|$ denotes the number of non-collision candidates in this set. With the optimal m_b , the GCF forces the packing models to separate from each other to the greatest extent, and the size of the non-collision candidate set is therefore maximized. Fig. 3 shows the size of the non-collision candidate set $|S_{\text{cand.}}(m_b)|$ with different m_b values under 3 different defined grids (in 30-, 40- and 60-degree

increments). We can see when $m_b = 0.2$, $|S_{\text{cand.}}(m_b)|$ has the highest value, and the peak is independent of how we define the grid. In the experiment, we use the 30-degree grid, and 48 non-collision candidates can be found. With $m_b = 0.2$, we compare the contours of the Gaussian mixture function with the rabbit shape in Fig. 4, and we can see that it fits the shape of the rabbit model well. Also in Fig. 5, we compare collision checking results with GCF values in the θ_1 - θ_2 plane with fixed $\theta_3 = -90$ degrees. It is shown that all non-collision configurations are located in the low GCF value regions, which demonstrates that by minimizing the GCF, the ellipses are less likely to have overlapping.

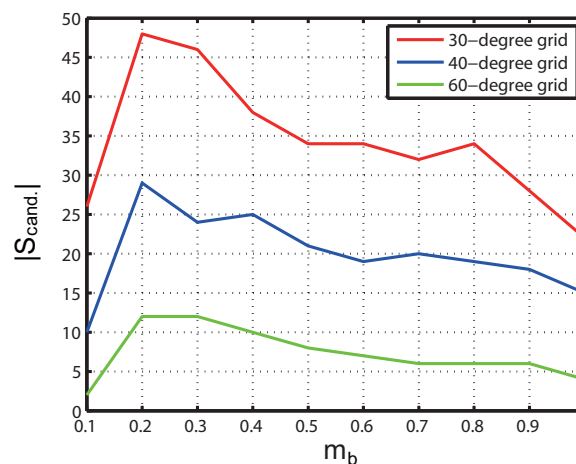


Fig. 3

THE SIZE OF THE NON-COLLISION CANDIDATE SET WITH DIFFERENT m_b VALUES UNDER 3 DIFFERENT DEFINED GRIDS (IN 30-, 40- AND 60-DEGREE INCREMENTS, RESPECTIVELY).

An example of packing results with the target structure randomly sampled in space is illustrated in Fig. 6. After generating the candidate set by minimizing the GCF, and sorting these candidates by the FCF from high to low, the best candidate in the set (Candidate 1 in Fig. 7) shows 1.50, 17.81 and 10.97 degrees of the error in θ_1 , θ_2 and θ_3 , respectively. After running the RRT around this candidate, these errors are further reduced to only 0.79, 2.14 and 0.19 degrees respectively, less than 1.2 % of the total rotation range. Table 2 shows 10 different numerical trials and the mean absolute errors (MAE), $\text{mean}\{\Delta\theta_1, \Delta\theta_2, \Delta\theta_3\}$, are all below 3 degrees.

6. CONCLUSIONS

Macromolecular crystallography has been the traditional workhorse for determining structural models in the field of biophysics. Within macromolecular crystallography, the molecular replacement method has been a highly successful

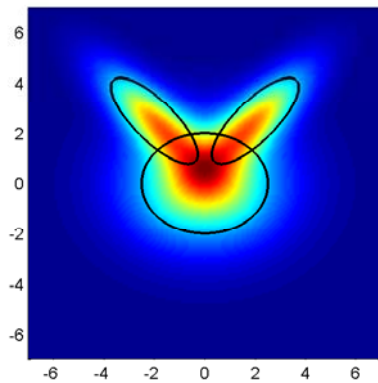


Fig. 4

THE COMPARISON OF THE RABBIT SHAPE WITH THE CONTOURS OF THE GAUSSIAN MIXTURE FUNCTION ($m_b = 0.2$).

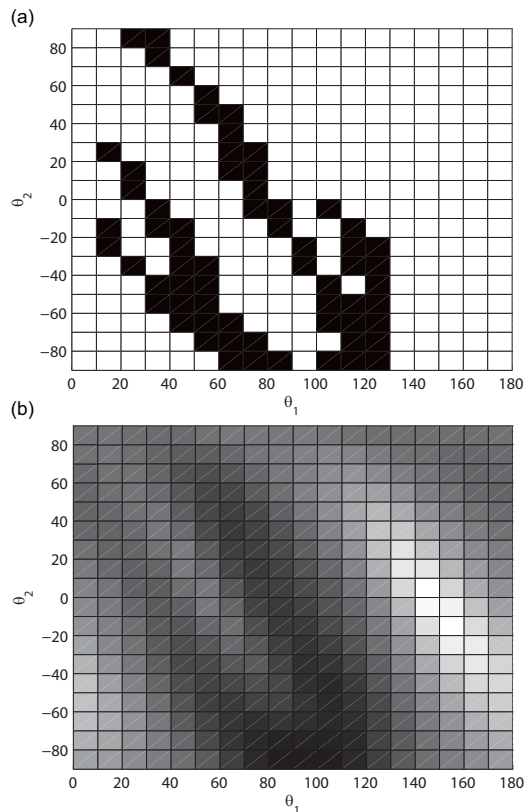


Fig. 5

THE COMPARISON OF (A) COLLISION CHECKING RESULTS WITH (B) GCF VALUES (WITH $m_b = 0.2$) IN THE θ_1 - θ_2 PLANE (WITH $\theta_3 = -90$ DEGREES). IN (A), BLACK PIXELS REPRESENT THE NON-COLLISION CONFIGURATIONS AND WHITE ONES ARE COLLISION FREE. IN (B), THE PIXELS WITH DARKER COLORS REPRESENT THE CONFIGURATIONS WITH LOWER GCF VALUES, AND VISE VERSA.

Table 1

THE DIMENSIONS AND RANGES OF MOTION OF THE RABBIT PACKING MODEL

Dimensions	size of the unit cell	9×6.75
	semi-axis lengths of the face	2; 2.5
	semi-axis lengths of the ears	2.3; 0.92
Range of rotation	face: θ_1 (deg)	$0 \sim 180$
	ears: θ_2, θ_3 (deg)	$-90 \sim 90$

method for providing phasing models to combine with experimental information to obtain protein structures. In this paper we demonstrate that an alternative to molecular replacement, called “phasing by packing” is promising for multi-rigid-domain structures. Numerical results illustrate the potential of this method.

7. ACKNOWLEDGMENTS

We acknowledge NIH Grant R01GM075310 for the support of this work and Dr. E. Lattman for useful discussions.

References

- [1] Hendrickson, W. A. and Ward, K. B., “A Packing Function for Delimiting the Allowable Locations of Crystallized Macromolecules,” *Acta Cryst. A* 32:778-780 (1976).
- [2] Williams, D.E., “Crystal Packing of Molecules,” *Science* 147(3658):605-606 (1965).
- [3] Damiani, A., Giglio, E., Liquori, A.M., Mazzarelli, L., “Calculation of Crystal Packing: A Novel Approach to the Phase Problem,” *Nature* 215:1161-1162 (1967).
- [4] Rossmann, M.G., Blow, D.M., “The Detection of Sub-Units within the Crystallographic Asymmetric Unit,” *Acta Cryst.* 15:24-31 (1962).
- [5] Rossmann, M.G., “Molecular replacement - historical background,” *Acta Cryst. D57*:1360-1366 (2001).
- [6] Navaza, J., “AMoRe: an Automated Package for Molecular Replacement,” *Acta Cryst. A50*:157-163 (1994).
- [7] Collaborative Computational Project Number 4, “The CCP4 suite: programs for protein crystallography,” *Acta Cryst. D50*:760-766 (1994) <http://www.ccp4.ac.uk/>
- [8] Jamrog, D.C., Zhang, Y., Phillips Jr., G.N., “SOMoRe: a multi-dimensional search and optimization approach to molecular replacement,” *Acta Cryst. D59*:304-314 (2003).
- [9] Jeong, J., Lattman, E., Chirikjian, G.S., “A Method for Finding Candidate Conformations for Molecular Replacement Using Relative Rotation Between Domains of a Known Structure,” *Acta Cryst. D D62*, pp. 398-409, 2006.
- [10] Crowther, R.A., and Blow, D.M. (1967). A method of positioning a known molecule in an unknown crystal structure. *Acta Cryst* **23**, 544.
- [11] Crowther, R.A. (1972). The fast rotation function. In *The Molecular Replacement Method*, M.G. Rossmann, ed. New York: Gordon and Breach Science Publishers, 173-178.
- [12] Lattman, E., Love, W.E., “A Rotational Search Procedure for Detecting a Known Molecule in a Crystal,” *Acta Cryst.* B26, 1854-1857, 1970.
- [13] Lattman, E.E., “Use of the Rotation and Translation Functions,” *Methods Enzymol.* 115, 55-77, 1985
- [14] Vagin, A., Teplyakov, A., “MOLREP: an Automated Program for Molecular Replacement,” *J. Applied Cryst.*, 30, 1022-1025 1997
- [15] LaValle, S. M. *Planning Algorithms*, Cambridge University Press, Cambridge, U.K., 2006.
- [16] Chirikjian, G.S., Zhou, S., “Metrics on Motion and Deformation of Solid Models,” *ASME J. Mechanical Design*, Vol. 120, No. 2, June, 1998, pp. 252-261.

Table 2
10 NUMERICAL TRIALS.

Trial	Target			Best θ_1	Cand. θ_2	θ_3	After			RRT			Final errors		
	θ_1	θ_2	θ_3				θ_1	θ_2	θ_3	e_1	e_2	e_3			
1	100.82	-72.21	-3.03	100.32	-90.00	-14.00	101.61	-74.35	-3.22	0.79	2.14	0.19			
2	42.29	64.37	-69.25	43.07	60.00	60.00	42.96	64.29	-67.17	0.67	0.08	2.08			
3	136.67	-68.70	-67.33	120.00	-39.37	-79.98	135.58	-67.54	-68.39	1.09	1.16	1.06			
4	114.21	-63.46	-51.42	120.00	-81.03	-60.00	116.43	-64.71	-49.70	2.22	1.25	1.72			
5	54.83	-49.51	-70.41	61.85	-60.00	-60.00	55.83	-50.37	-69.37	1.00	0.86	1.04			
6	159.67	47.67	-2.65	173.97	26.77	14.89	160.75	44.52	0.43	1.08	3.15	3.08			
7	101.63	-67.65	12.06	114.08	-88.32	31.05	103.72	-70.20	13.08	2.09	2.55	1.02			
8	113.89	-73.69	30.76	120.00	-90.00	38.95	112.72	-74.80	29.70	1.17	1.11	1.06			
9	66.41	27.29	-76.94	60.00	38.95	-90.00	63.20	30.78	-78.49	3.21	3.49	1.55			
10	97.19	-1.59	-86.46	120.00	-39.37	-79.98	100.02	-2.89	-83.53	2.83	1.30	2.93			

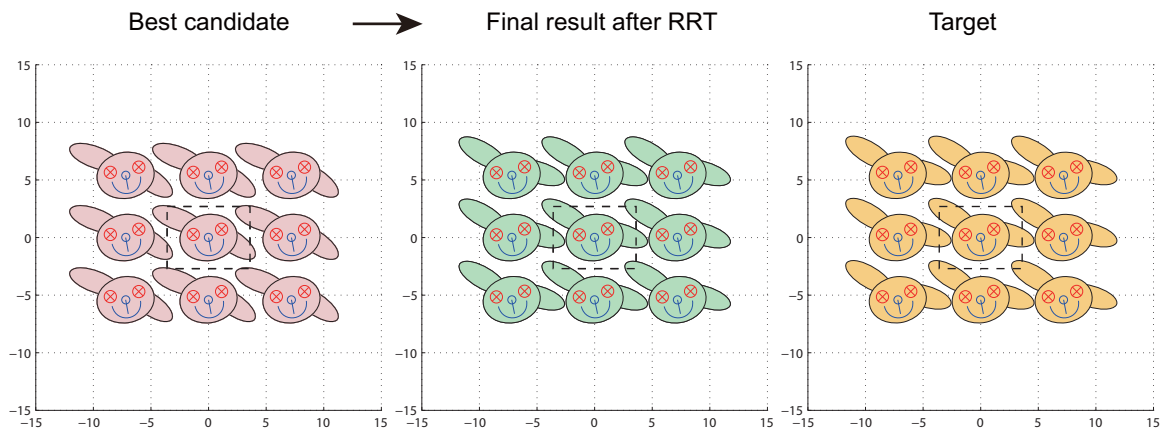


Fig. 6

AN EXAMPLE OF PACKING RESULTS WITH THE TARGET STRUCTURE RANDOMLY SAMPLED IN THE SPACE.

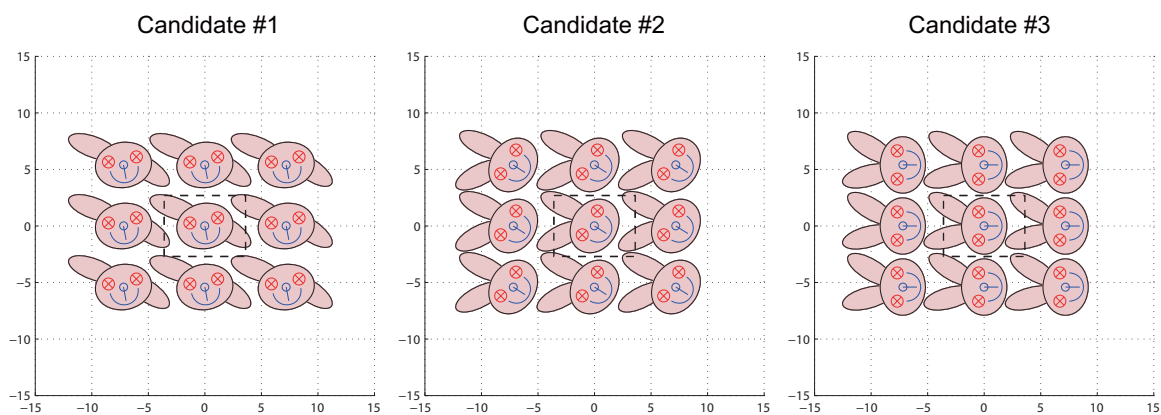


Fig. 7

3 CANDIDATE PACKING ARRANGEMENTS FOR THE EXAMPLE IN FIG. 6.

Biological Data Handling Methods

Pradeep Achan¹, Ajit G Warriar¹, and Bhadrachalam Chitturi²

¹School of Biotechnology, Amrita Vishwa Vidyapeetham University, Amritapuri, Kollam, India

²Department of Computer Science, Amrita Vishwa Vidyapeetham University, Amritapuri, Kollam, India

Abstract -Biological data has more variation in type and format compared to other types of data. Thus, it poses new challenges. However, it encapsulates critical information; thus, handling it is of primary interest. Data handling includes storage and retrieval of data with associated formats and methods of data transfer, data format conversion, algorithms that run on the data and the output methods including visualization of the results. High throughput methods have been yielding biological data at a fast pace. This data includes protein-protein interactions, gene sequences, gene co-expressions, and protein sequences. This data is supplemented with huge amounts of clinical data conveniently captured in electronic medical records and the wet lab data. We describe the current approaches, each with a model system and identify its key contributions. We propose some ideas for biological data handling in the future.

Keywords: biological data handling, cloud computing, data integration, data modeling, semantic web, systems biology

1 Introduction

The term biological data is used in a broad sense. It includes genomics/proteomics data, the data generated from experimental biology, diseases data and patient clinical data. High throughput screening has been yielding large quantity of new data in biology. Micro array analysis provides gene co-expression data, the next generation sequencing, *i.e.* NGS, yields DNA sequences and so on. Even though various types and formats of the data pose challenges the information in the data is vital. Biological data is distributed in various sources; it has redundancy, different formats and naming conventions. A researcher potentially needs the information from various sources. The features that contribute to the difficulties in handling of such data are: 1) the quantity of the data, 2) various sources, formats and naming conventions, 3) the dynamic nature of the data and 4) the complex relationships between several data objects which can be of various types.

The enormity of biological data renders *warehousing* (*i.e.* data warehousing), computing, transmission of the data over the network difficult owing to higher requirements in storage space, computation and bandwidth. Also, integration of large quantities of data is resource intensive. Examples of the data formats are a flat file such as “tab delimited format” or a database dump such as MySQL database dump or an excel spreadsheet. A protein can be addressed with various names in various databases owing to diverse naming conventions. A researcher typically collects genome data,

literature abstracts, protein information, pathways, and 3D structure from Genome database, PubMed, Uniprot, KEGG and PDB respectively [42]. New entries of a given object type and new relationships are continually discovered. For example, a new protein (of object type “protein”) can be discovered. Likewise, a previously unknown interaction can be detected between a pair of proteins present in the database. Thus, the data is dynamic in nature. This causes problems in systems with a warehouse or without it *i.e.* federated system. A central repository will be outdated if new data is added to external sources after the last update. A federated system might become dysfunctional due to schema modification at one or more data sources.

Keeping these databases up to date and in phase with each other is quite challenging, more so in the wake of NGS technologies. Consider a system with a warehouse C which uses data sources $S = \{s_1, s_2, \dots, s_q\}$. As stated earlier, C can have older data compared to S . Also, the data in S can be inconsistent. Consider a scenario where a new gene g and a protein p coded by it are discovered. Let p interact with a known protein q . Say s_i , s_j and s_k have protein (with foreign key to gene), gene and protein-protein interaction (PPI) data respectively. Some of scenarios where the data in S is incomplete are: (a) s_j is updated with g , s_i is updated with p whereas s_k is not updated accordingly (it does not have PPI for p and q), (b) s_j is updated with g whereas s_i and s_k are not correspondingly updated, (c) s_j and s_k are correctly updated whereas s_i is not correspondingly updated. In (a) just the interaction information is missing but it does not have any serious inconsistency. In (b) both the protein and the PPI are missing which is a minor inconsistency because we do not find the protein for a given gene. In (c) the critical link between g and the interacting pair p and q is missing. Thus, C can have two types of problems; *i.e.* it can have outdated data compared to S or it can be in phase with S and yet inherit inconsistency that is inherently present in S . These problems point to the need for frequent access to the information across different databases which are spread across different Internet data sources, consistency check of the data and the practical limitation of having large databases (multi-terra bytes) warehoused centrally due to the limitation of storage space.

Biological data has unique complexity and levels of abstraction as detailed in Section 2. The processing of biological data involves various tasks that depend on the application and the input data. One can broadly subdivide the process into the following chronological sequence of four tasks: a) data acquisition and preprocessing, b) analysis of relationships between data objects c) creating a data model for a given application, and d) creating output.

Data acquisition refers to acquiring the data from the data source(s). Data is often stored in various formats, e.g. flat files, spreadsheets etc. which are not directly conducive to computation. Such data is often converted into a database table; this step is *preprocessing*. *Analysis of relationships* between data objects primarily refers to the domain knowledge; e.g. the relationship between: a protein and a domain, a gene and protein etc. Analyses of the relationships between data objects are represented as structured information in database systems. These are read into application-specific in-memory organization of data. This application specific data organization in database system as well as in-memory data structure can be called as the *data model* of the application. Application can process the data model to create secondary information by selection (retrieving specific pieces of information), aggregation (aggregating information from different sources), or mining (for patterns within the data). The application presents the results of a query as *output*.

Methods for data acquisition and preprocessing are well established and the analysis of relationships is achieved with the expertise provided by biologists. We discuss Output in some detail. Data modeling and the associated task of data integration are more thoroughly covered. Data model which comes from the analyses of relationships can be viewed as a template; when it is executed, it results in data integration.

In Section 2, different approaches for building systems are described with a special focus on the emerging semantic web methodologies. Section 3 details handling of the output. Section 4 gives the features provided by cloud computing. Section 5 details a few recent innovative projects. Section 6 states key findings from different approaches and lists open problems and the work that mitigates some of these problems. It also states some desirable features for the future biological data handling systems.

2 Approaches for biological systems

A system has certain functionality and it is built with a specific approach. In this section, we discuss approaches for building such systems. Subsections 2.1 and 2.2 explore the approaches of the vital aspects of such systems, *i.e.* data integration and data modeling respectively.

2.1 Approaches for data integration

Data integration needs for applications vary considerably with the user who can be a biologist, bioinformatician or a systems biologist. A review of various integration approaches is given in [7], where they are labeled as light to heavy in terms of integration efforts.

Integration techniques which include the use of scripts written in Perl and Python [42] exist. Service based methods like WSDL an XML format provides a model for describing Web services [42]. [20,46] classify data integration approaches into warehousing, mediator or view integration and also as link or navigational. [46] describes the use of Web Services, Distributed Annotation System (DAS) and Globally Unique Identifiers in data integration and also proposes an

approach, termed as “knuckles-and-nodes approach”, where in the source databases remain independent but a few important relationships are stored in special-purpose linking databases. In addition the use of scripting, peer-to-peer systems, semantic web technologies and workflow-based were introduced in [42]. The approaches mentioned in [42, 20, 46] overlap with each other in various aspects; *i.e.* technological choice, methodology etc. Also they are not mutually exclusive but use or depend on some others for effective data integration. Link integrations are used in building systems based on either relational model or semantic web technology.

Archival databases like NCBI, EMBL, DNA Data Bank of Japan, maintained by International Nucleotide Sequence Database Collaboration accept data directly from sequencing labs and are referred as primary sequence database [47]; they aggregate data centrally. Similarly, primary protein sequence databases include PIR and UniprotKB (Swiss-Prot/TrEMBL) which handle the protein sequences. Other systems act as value added integrators of this data such as Ensembl, UCSC Genome Browser, Uniprot and Model organism databases [47]. These provide data in convenient formats for further aggregation and analysis. Secondary data sources like PROSITE, PRINT, Pfam aggregate data centrally and also link to primary data sources by unique identifiers.

Most of the primary and secondary databases link to other information sources through link integration. Some systems are built by power (advanced) users from these primary and secondary sources for custom application systems [47]; they may be general purpose or special purpose systems [37]. We refer to them as *tertiary* systems, *e.g.* BioWarehouse [35], ATLAS [44] and ONDEX [31]. All of these aggregate data. In contrast, TAMBIS [48], BIO-BROKER [1], and SEMEDA [32] use a mediator approach *i.e.* they use a wrapper to access original data sources.

Some other systems [46] store a part of data in a warehouse in addition to the use of mediation for effective integration. In [30] another approach was introduced to integrate gene expression data and proteins stored in data warehouse with annotation data retrieved from public sources using sequence retrieval system. The above mentioned integration methods [30,46] are also termed as hybrid systems. SADI does not store data locally and links with other systems using REST-based [15] web services [54].

The advantages of warehousing approach are: it relies less on network [20], allows faster query performance, allows the system to filter, validate, modify, and annotate the data obtained from the sources [20], *e.g.* BioWarehouse [35]. It also facilitates the integration of locally derived experimental data into the repository. However, it needs large storage (the biological data is semi-structured and is not easily stored in relational databases (or simply *RDBs*) [42] and it must be synchronized with underlying sources for updates [49]. Biological data needs significant computation to be stored in the typical format *i.e.* *RDBs*.

Semantic web is an emerging technology by WWW consortium describing it as “web of data” [22]. An informal definition for the Semantic web technologies could be “comprising of four essential component technologies namely

RDF, RDFS, OWL and SPARQL" [2]. Semantic web uses uniform resource identifier, URI, to represent a data object, mostly in a triple containing *subject*, *predicate* and *object*. This triple, which uses three URIs, is called Resource Description Framework (RDF) [23]. A *triple store* stores this triple (RDF data). RDF represents the information or data as a graph. RDFS and OWL [24] are ontology languages. Querying the RDF graph is done with a querying language similar to SQL called SPARQL [26]. A SPARQL query is denoted by a graph pattern containing the patterns of triples that are similar to RDF triples but are replaced with variables.

Current usage of Semantic web technologies for biological knowledge management has been described in [2]. Knowledge management refers to the process of systematically capturing, structuring, retaining and reusing information to develop an understanding of how a particular system works, and subsequently to convey this information meaningfully to other information systems. [2] lists selected resources and projects which use Semantic web technologies and suggests more prevalent use of it in future systems.

Majority of the data is stored in RDBs and it is difficult for Semantic web technologies to access them. Thus, an application tries to create its own relational to semantic mapping and thereby accessing the relational data using SQL. Semantic web layer can play a great role in integrating relational data into Semantic web technologies, it defines the standard vocabularies, formal models and semantic relations between RDBs [9]. Datagrid [9] framework along with a set of practical semantic tools was used to facilitate the integration of heterogeneous RDBs using Semantic web technologies. OWL [41] is a technique to extract the semantics of a RDB and transform it into RDF/OWL. It extracts the schema information of the data source and converts it automatically into ontology. With this technique every RDB can automatically be an integral part of Semantic web. Thus, web applications can access and query data stored in RDBs using their own built-in functionality [41]. Jiang et al. describe an architecture to expose RDB to Semantic web application using Hibernate [18]. OWL ontology is translated to java classes and then a runtime SPARQL to hibernate query language (HQL) translation algorithm was introduced for efficient run time translations [18]. This method suits queries without cycles and a subset of SPARQL language [18].

2.2 Approaches for data modeling

Data modeling is considered to be the critical task of Biological Data Handling. Some of the open problems in it are covered in [10,14]. Elmasri et al. [10,14] state that ordering (e.g. DNA sequences), 3D structures of proteins and functional processes (e.g. metabolic pathways) as the main characteristics of biological data. Conventional data representation does not explicitly include these characteristics. However, they are biologically relevant and ideally data representation should include a mechanism to represent these characteristics. [10,14] propose a new enhanced ER (EER) schema, notation to represent the same and give methodology to implement the same in a RDB. Ordered relationships are

modeled by extending the relationship concept in two directions 1) allowing related entities to be ordered and 2) allowing the repetitions of a relationship instances. Molecular spatial relationship deals with the representation of 3D structures in conceptual EER modeling. Atoms and amino acids are modeled with molecular spatial relationships and these spatial structures generate the measurement data like bond angles and bond distance. Atom is treated as points and its position is represented with coordinates in space. Process relationships have three basic entities *i.e.* input, output and catalyst. Inputs are used by the process, the outputs are produced by the process and catalysts are needed for the process to work. Biological pathways are examples of process relationship where an output of one reaction becomes the input of another. For example, the output of transcription process, mRNA, serves as an input for the subsequent translation process.

In [10,13] a multilevel EER model for biological processes which incorporates the multilevel concepts and relationships is proposed. [13] highlights biological examples along with their conceptual EER modeling notations to show that multilevel modeling can be effectively used in biomedical domains and introduces the important concept that at different levels of abstraction, data needs to be modeled differently. The method in [13] also introduces various approaches for data source integration namely horizontal and vertical approaches. The advantage of vertical approach over the horizontal approach is that it integrates data sources from different abstraction levels while the horizontal approach facilitates the integration of data source from same level of abstraction.

In RDB systems, data elements are stored in RDB tables and each table contains an entity with primary key and attributes. Two different entities are related through foreign-key relationships between their keys. Such relationships are not formally defined with specific names. So, such relationships cannot be queried upon. In contrast, Semantic web technology uses RDF and the relationship is treated as a first class entity (predicate), referenced by a URI and stored along with subject and object. In RDF, relationships can also be queried (e.g. SPARQL query). This means, the graph of persistent RDF nodes contains the full semantic information about the entities and the relationship between them. In RDBs custom programs are needed for each database schema and the programmer must know the relationship between the tables. Likewise, these relationships are specified in the queries. However when data is stored as RDF graphs, general purpose programs can be written without the knowledge of the underlying RDF graphs, and this could provide a general purpose querying interfaces to the underlying RDF graphs.

2.2.1 Systems biology and data modeling

Systems biology studies introduce another dimension, by requiring different search and modeling needs depending on the user. [8] Introduces different Systems Biology standards that are either accepted or in development. E.g. minimum requirements like MIRIAM and MIASE, the description formats like SBML, SBRML used to represent

data and the associated ontologies like SBO, KiSAO and TEDDY are used to integrate different models to have a better understanding of the complete system. [19] highlights the complexity of biological data as one of the major problems along with the scale of data generated NGS and the scope of the experimental investigations with systems biology. It introduces new data integration architecture Addama. An approach to integrate information management supporting the bottom-up systems biology was introduced in [50]. It proposes to build an automated integration system that can automatically capture the experimental data and integrate it with models.

3 Output methods in data handling

Depending on the nature of the application, output methods can widely differ. Many systems provide knowledge extraction for a human or a computer. Such systems provide search/-results interfaces typically based on a query where the results are displayed as output [17]. Many systems provide structured search capabilities. This is achieved by allowing the input keywords to be associated with specific data elements; providing matching conditions like $>$, $<$, contains etc. and search the underlying data for specific matching criteria [34].

Search results have various presentation styles that include computer readable formats. For knowledge extraction systems, faceted browsing [39] is a suitable style. It is effective in showing biologically relevant data where the result set can be easily filtered and categorized. BioFacets [36] allows a faceted classification *i.e.* dynamic categorization of biological result set. Faceted interfaces go naturally with semantic query search and retrieval systems and can help modeling the biological data. Often output has inter-related information; *i.e.* gene-gene interactions and pathways; which demands visualization to effectively display the search results.

Visualization gives insight into the biological process and hidden relationships between data elements. A survey of visualization tools for biological network analysis highlights the pros and cons of each tool [40]. For visualizing the output data Cytoscape, Ondex, PATIKA [40] etc. provide excellent support. Cytoscape can be enhanced by plugin interfaces [45], it supports Semantic web by importing data from triple store through simple text table or XML-RDF, loading and visualizing RDF data as networks and querying the RDF data with SPARQL. It also helps in developing custom Semantic web applications with Jena and Sesame. It can also be used with other tools like statistical programming language R with sna/ igraph package. For GenomeGraphs [12] an add-on package for R was developed for visualization of genomic datasets. Addama [19] also uses R for its dynamic visualization capabilities.

Often visualization systems provide interactive visualization capabilities. Querying the Semantic web with SPARQL may not be easy for a novice who does not know the structure of the ontology. [29] describes a rewriting of SPARQL to allow users to write queries from their perspective (without knowing the structure of the ontology) but it has limitations. A similar approach was described in [6],

which introduces a semantic approach to process knowledge in two phases *i.e.* constructing a semantic query from the user input and displaying the semantic result using scalable vector graphics. Here, the results are output as an RDF graph, often with interactivity to navigate the RDF graph. For systems that output data to be fed into other computer systems, communication standards, ontology, data integration and minimal specification languages play an important role.

4 Approaches enabled by cloud

Various computational solutions to large scale biological data handling are explained in [43]; specifically cloud computing and heterogeneous computing. Currently, the quantity and the storage of genomic data is a vital issue. Cloud computing plays a vital role in the management of genome informatics [47]. Large datasets that act as a virtual disk are stored in a cloud. It inspired projects like Galaxy [51] to build tools to easily setup clusters on cloud platforms. Problems of large datasets requiring huge storage space, processing power and network bandwidth are largely mitigated by commercial scale cloud enabled approaches [47]. Data source providers can expose the data for many consumers, who can access only the requested data through service oriented approaches from the cloud. Extension systems can co-exist in local systems with the cloud. It may be noted that analytical toolbox for biological data like Bioconductor [16] and Galaxy [51] provides prebuilt images for the popular commercial cloud platform Amazon Elastic Computing Cloud (EC2), thus, eliminating large scale datasets and complex software setups on a local network.

5 Examples of data handling systems

The study of biological data handling systems yields the following aspects.

- Data is either aggregated or linked to.
- For non-warehoused systems mediator is needed.
- Ad hoc data retrieval methods extract data and information in unintended ways.
- Extendibility in functionality (ability to add new functions to the system by scripts/ programs).
- Expandable data models (Open world system).
- Use of semantic relationships between data elements.
- Technology choices (Web services, REST)
- Use of infrastructure (Cloud)
- Systems Biology requirements
- Use of output methods

Here, we explore a few innovative systems to identify the underlying concepts. Sample systems are meant to demonstrate such concepts; they are not comprehensive.

5.1 BIO2RDF

Bio2rdf project [4] gives the standards for a system to use Semantic web technology to cross-link information sources and expose services to each other. Since many of the existing systems are not enabled with these technologies, current implementation of Bio2rdf also transforms the data

into semantically linked formats, and exposes a semantic query front end. That is, it has a warehouse for demonstration purposes. It asserts that if the Bio2rdf standards are implemented by the systems then warehousing of the data and Bio2rdf project itself are not needed. Bio2rdf tries to create a network of coherent linked data across the life science databases and provides various SPARQL endpoints to query the RDF graphs without locally storing the graph [4]. A user can define a SPARQL query in a query form and it can be sent to the triple store, and the results can be sent back to the user. With this approach it is possible to link different databases containing the RDF data using the federated and distributed SPARQL queries. Bio2rdf successfully integrated 163 million documents from a large number of data sources [4].

5.2 SADI

Semantic Automated Discovery and Integration, SADI is a Semantic Web Service (SWS) framework which integrates the data from various sources [54]. It is seen that the web services create an implicit biological relation between the supplied input and the retrieved output, but SADI links the input with the output with a common base identifier and the services are annotated thereby explicitly describing the semantic relation between them [54]. SADI framework attempts to build a virtual database by extracting RDF triples through web services, the data can be queried by SPARQL. SADI has improved upon BioMoby and SSWAP by having a SWS framework that integrates itself more naturally into the Semantic web [54]. SHARE is a mediator system which enables federated querying where resources are exposed as services using the SADI SWS framework [54]. SADI services are also REST-like; there is only a standard basic set of HTTP methods, *i.e.* GET and POST [54]. A GET operation on a given service returns its semantic description, while a POST initiates service execution and returns the same RDF graph with the annotations created by the service [11].

CardioSHARE [52] is a unique framework for querying distributed data and performing data analysis using Semantic web standards. The SPARQL query engine of CardioSHARE retrieves the required data dynamically from web services [52]. CardioSHARE project is built on the strengths of BioMoby [55] and addresses its weakness by replacing its syntax with Semantic web ontologies [52]. It is a prototype application that accesses SADI services in response to SPARQL queries. It was initially designed for the analysis of clinical data on heart disease but can be extended to integrate any type of biological data [52].

SADI addresses the problem that most of biological data is in “deep web” and enables discovery of new information from it [54]. SADI proposes a scaled-down version of web service usage, especially suited to bioinformatics; and thus improves upon the earlier Web Services implementations like BioMoby and SSWAP [54]. SADI tries to expose analytical services as REST-enabled URLs [15] that can be combined to form analytical workflow pipelines. Thus, SADI supports and enables ad hoc extension of its data models and functionality.

5.3 ADDAMA

A recent article [19] highlights the complexity of biological data as a major problem along with the scale of data generated and scope of the experimental investigations with systems biology. It introduces new data integration architecture Addama which has been developed for systems biology investigations. Addama tries to integrate and extend existing enterprise technologies to enable the rapid development of ad-hoc tools, and to provide a robust and scalable software infrastructure [19]. The ongoing research requires an adaptable system which provides an integration framework for the existing software technologies while addressing the user requirements which include universal access, support of discovery process and adaptation to new technologies and usage [19]. Addama meets all the user requirements and it does it by allowing a combination of both enterprise technologies and organic software development models. It supports scientists in the use of heterogeneous data types and through the development of related visualization and analysis tools. It defines service interfaces to integrate selected technologies with the underlying infrastructure [19].

6 Key findings and recommendations

The objectives of all systems are similar; so, the best aspects of all systems can be combined to yield a better approach. Data warehousing still has better performance and reliability, and acceptance from academia and industry. Relationships between entities are lost when E-R diagrams are converted into database schema [33]. These can be restored by adding tables to store relationships and multiplicity to model RDB tuples as RDF. Each RDB entity can have a reference id, as defined in some specific domain-standard ontology. Thus, RDB can be “semantically enriched”.

Warehouse data can achieve data provenance (authentication) by storing information about source and version; this along with conflict resolution methodologies can be used to build automated/semi-automated update cycles. Warehousing systems build custom parsers to convert source data, *e.g.* for Uniprot data, BioWarehouse [35] has a parser with XMLBeans [3] technology and object to relational mapping (ORM) conforming to the DRY principle [27]. A class/object model generator that can take OWL based data models as input and an ORM toolset that generates semantically-enriched RDB schema is desired.

Database systems tend to be a closed-world system but they present a consistent snapshot of the knowledge. Open world data from heterogeneous sources can be inconsistent. Warehousing can be enhanced to have knowledge discovery (KD) capabilities by providing connectors to open-world systems; *e.g.* SADI allows other SPARQL end points from the open web [53]. Results from such queries can be checked for consistency with the standard snapshot version of information. SADI effectively addresses the problem of a researcher having to go to multiple websites [10]. SPARQL gives a system the capability to extend its knowledge store [38], this is highly desirable.

The major disadvantage of distributed querying is performance [38,46,49] which can be mitigated by caching. Extensive research has been performed in the area of caching the SPARQL queries [38, 49]. Here the query result is cached with an idea of reusing the computed results of previously generated queries avoiding the network usage and increasing the robustness of the system by providing a local copy of cached data when the original source data is unavailable [49].

The adoption of Semantic web technologies for data integration needs productivity enhancing tools for programmers. Possibilities for ORM tools to be architecturally enhanced to work with RDF and OWL is referred to in [18].

SPARQL has the potential to be the choice of end user for knowledge management system that uses Semantic web technologies and maintain semantic relationships. More so, if it procures visual query construction methods [6].

Bio2rdf converts the data from other formats into RDF format using RDFizer [21] whereas SADI leaves the data at its original location.

ADDAMA stresses the need for an ad hoc extension of data stores and functionality [19]. Ad-hoc extensions are especially sought if they are easily mastered and are programming language independent.

We argue that in addition to general purpose query capabilities exposed by SPARQL one may build ADDAMA style REST-based data access services into underlying semantic data stores. Addama also provides the process management services layer with REST-like access mechanism and also provides for a coordinating central registry service. Not all ad hoc data inputs from research communities are curated. They are neither sufficiently structured nor formatted to organize them into RDB models. They contain very less details to organize them into RDF-graph, and much less to be mapped to standardized nomenclature systems and ontologies. Such data also can be input into analytical algorithms in addition to well-structured data from well curated public data, Addama supports this use case. ADDAMA uses content repositories in addition to SQL databases for storing ad hoc data inputs. We note that, for any large scale data handling systems to be effective to serve the research community, ADDAMA approach is very important.

Visualization of experimental results and its analysis capability can be provided with programming extensions to large scale systems as illustrated by Addama project. Use of statistical programming language like R [28] is best suited for this purpose.

In SADI where the output is mapped as annotation to the input data structure, it is possible to build pipelines of processes. Also, input and output data structures can be in a common model (RDF graph). Analytical process pipelines are important for biological research to reduce the time taken for knowledge discovery and processing. Further, the addition of Cloud enabled approaches, wherein data source providers can host the datasets in the cloud and the consumers can access only the needed subset of the data through service oriented methods, can solve many problems related to the scale of biological data and also make the systems reusable thereby

reducing the duplication of work. This is our key learning from Galaxy [51].

Our general recommendations are stated here. For future systems, faceted UI is the best choice for visualizing the output. Use of semantic web technologies (controlled vocabularies, ontologies and RDF) is highly desirable. Cloud computing overcomes the issues of huge local repository and outdated data. With proper design, federated approaches can be adapted with minimal deterioration in the data availability and system performance. Service oriented approach, with use of REST is important for large-scale data integration.

7 Acknowledgment

We acknowledge Schools of Biotechnology and Engineering (CS dept.) of Amrita University for the support to conduct the research. We also thank Dr. T.C. Gilliam, Dr. N. Maltsev's team (D.S, S.B., E.B., R.K., B.Q.) and U. Dave of University of Chicago.

8 References

- [1] JF Aldana, M Roldán-Castro, I Navas-Delgado, MM Roldán-García, M Hidalgo-Conde, O Trelles. Bio-Broker: a tool for integration of biological data sources and data analysis tools. *Software: Practice and Experience*, 36(14):1585-1604, 2006.
- [2] E Antezana, M Kuiper, V Mironov. Bio. knowledge management: the emerging role of the Semantic Web tech. *Brief Bioinform*, 10(4):392-407, 2009.
- [3] Apache XML Project. Java XMLBeans, 2003. Available from <http://xmlbeans.apache.org>.
- [4] F Belleau, MA Nolin, N Tourigny, P Rigault, J Morissette. Bio2RDF: towards a mashup to build bioinformatics knowledge systems, *J Biomed Inform*. 41(5):706-16, 2008.
- [5] VY Bichutskiy, R Colman, RK Brachmann, RH Lathrop. Heterogeneous Biomedical Database Integration Using a Hybrid Strategy: A p53 Cancer Research Database, *Cancer Inform*, 2: 277-87, 2007.
- [6] TD Cao. Integrating a Graphical Semantic Query Interface and a SVG-based knowledge presentation method in an Enterprise Knowledge Management System. *Proc. of AUN/SEED-Net Regional Workshop in Information and Communication Technology*, 2009.
- [7] G Carole and S Robert. State of the nation in data integration for bioinformatics, *Journal of Biomedical Informatics*, 41(5), Pages 687-693, October 2008.
- [8] VL Chelliah, N Endler, J C Laibe, C Li, N Rodriguez, N Le Novere, Data Integration and Semantic Enrichment of Systems Biology Models and Simulations, *LNCS*, V.5647:5-15, Jul 2009.
- [9] H Chen, Y Wang, H Wang, Y Mao, J Tang, C Zhou, A Yin, Z Wu. Towards a semantic web of relational databases: a practical semantic toolkit and an in-use case from traditional chinese medicine. *LNCS*, V.4273:750-763, Nov 2006.
- [10] J Chen, S Amandeep: Biological database modeling, *Artech House*, ISBN 13: 978-1-59693-258-6, 2008.
- [11] LL Chepelev and M Dumontier. Semantic Web integration of Cheminformatics resources with the SADI framework, *J Cheminform*. 3:16, 2011.

- [12] S Durinck, J Bullard, PT Spellman, S Dudoit. GenomeGraphs: integrated genomic data visualization with R. *BMC Bioinformatics*, 10:Article 2, 2009.
- [13] R Elmasri, F Ji, J Fu, Y Zhang, & Z Raja: Modeling concepts and database implementation techniques for complex biological data, *Int. J. Bioinformatics Research and Application*, 3(2):366-388, 2007.
- [14] R Elmasri, J Fu, and J Feng. Multi-level conceptual modeling for biomedical data and ontologies integration, *CBMS*, pp.589–594, 2007.
- [15] RT Fielding. Representational state transfer (REST), *Ph.D. Thesis*, University of California, Irvine, CA, 2000.
- [16] RC Gentleman, VJ Carey, DM Bates, B Bolstad, M Dettling, S Dudoit, B Ellis, L Gautier, Y Ge, J Gentry, K Hornik, T Hothorn, W Huber, S Iacus, R Irizarry, F Leisch, C Li, M Maechler, AJ Rossini, G Sawitzki, C Smith, G Smyth, L Tierney, JY Yang, and J Zhang. Bioconductor: open software development for comp. biology and bioinformatics, *Genome Biol.* 5(10): R80, 2004.
- [17] L Guo, J Shanmugasundaram, G Yona. Topology Search over Biological Databases, *ICDE*, 2007.
- [18] J Hao, J Liwei, X Zhuoming. Upgrading the Relational Database to the Semantic Web with Hibernate, *Intl. Conf. on Web Information Systems and Mining*, 227-230, 2009.
- [19] R Hector, K Sarah, S Ilya and B John. An Integration Architecture Designed to Deal with the Issues of Biological Scope, Scale and Complexity, *LNCS*, V.6254:179-191, 2010.
- [20] T Hernandez, S Kambhampati. Integration of biological sources: current systems and challenges ahead, *SIGMODRec.*, 33:51–60, 2004.
- [21] <http://simile.mit.edu/wiki/RDFizers>
- [22] <http://www.w3.org/2001/sw/>
- [23] <http://www.w3.org/RDF/>
- [24] <http://www.w3.org/TR/owl-features/>
- [25] <http://www.w3.org/TR/rdf-schema/>
- [26] <http://www.w3.org/TR/rdf-sparql-query/>
- [27] A Hunt, D Thomas. “Don’t Repeat Yourself.” The Pragmatic Programmer, Addison Wesley, Boston, 2000.
- [28] CDR Ihaka, R Gentleman. R: a language for data analysis and graphics, *J. Comput. Graph. Stat.*, 5(3), 299–314, 1996.
- [29] P Jain, P Yeh, K Verma, C Henson, A Sheth. SPARQL query re-writing for spatial datasets using Partonomy based transformation rules, *Proc. of the Third Intl. Conf. on GeoSpatial Semantics*, 140–158, Springer, 2009.
- [30] T Kirsten, HH Do, C Körner, E Rahm. Hybrid integration of molecular biological annotation data, *Proc. Intl. Workshop on Data Integration in the Life Sciences*, 2005.
- [31] J Kohler, J Baumbach, J Taubert, M Specht, A Skusa, A Rueegg, C Rawlings, P Verier, S Philippi. Graph-based analysis and visualization of experimental results with Ondex, *Bioinformatics* 22:1383–1390, 2006.
- [32] J Kohler, S Philippi, M Lange. SEMEDA: ontology based semantic integration of biological databases, *Bioinf.* 19(18):2420–2427, Dec 2003.
- [33] M Krishna. Retaining Semantics in Relational Databases by Mapping them to RDF, IAT Workshops 2006.
- [34] M Latendresse, PD Karp. An advanced web query interface for biological databases, *Database*, 2010 doi:10.1093/database/baq006.
- [35] TJ Lee, Y Pouliot, V Wagner, P Gupta, DW Stringer-Calvert, JD Tenenbaum, PD Karp. BioWarehouse: a bioinformatics database warehouse toolkit, *BMC Bioinformatics*, 7:170, 2006.
- [36] M Mahoui, ZB Miled, A Godse, H Kulkarni, N Li. BioFacets: Faceted Classification for Biological Information, *18th Intl. Conf. SSDBM*, 225–234, 2006.
- [37] K Marrakchi, A Briache, A Kerzazi et al.. A Data Warehouse Approach to Semantic Integration of Pseudomonas Data, *LNCS*, V.6254:90–105, 2010.
- [38] M Martin, J Unbehauen, S Auer. Improving the Performance of Semantic Web Applications with SPARQL Query Caching, *LNCS*, V.6089:304–318, 2010.
- [39] M Norman, H David, H Lynette et al.. Data shopping in an open marketplace: Introducing the Ontogator web application for marking up data using ontologies and browsing using facets, *Stand Genomic Sci.*; 4(2): 286–292, Apr 2011.
- [40] G Pavlopoulos, AL Wegener, R Schneider. A survey of visualization tools for biological network analysis, *BioData Min*, 1:12, 2008.
- [41] C Perez de Laborda, S Conrad. Bringing relational data into the semantic web using SPARQL and relational OWL, *Proc. ICDE* 55–60, 2006.
- [42] L Raschid. Data Modeling and Data Management for the Biological Enterprise, *OMICS*: 7(1):51-55, Jan 2003.
- [43] E E Schadt, MD Linderman, J Sorenson, L Lee, GP Nolan. Computational solutions to large-scale data management and analysis, *Nat. Rev. Genet.*, 11:647-657, 2010.
- [44] SP Shah, Y Huang, T Xu, MMS Yuen, J Lin, BFF Ouellette. Atlas—a data warehouse for integrative bioinformatics, *BMC Bioinformatics*, 6:34, 2005
- [45] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Research* 13(11):2498-504, Nov 2003.
- [46] L D Stein. Integrating biological databases, *Nature Rev. Genet.* 4:337–345, 2003.
- [47] L D Stein. The case for cloud computing in genome informatics, *Genome Biology*, 11:207, 2010.
- [48] R Stevens, P Baker, S Bechhofer, G Ng, A Jacoby, NWPaton, CA Goble, A Brass. TAMBI: transparent access to multiple bioinformatics information sources, *Bioinformatics* 16(2):184-186, 2000.
- [49] H Stuckenschmidt. Similarity-Based Query Caching, *LNCS*, V.3055:295-306, 2004.
- [50] N Swainston, DJameson, P Li, I Spasic, P Mendes, N Paton. Integrative Information Management for Systems Biology, *LNCS*, V.6254:164-178, 2010.
- [51] J Taylor, I Schenck, D Blankenberg, A Nekrutenko. Using Galaxy to perform large-scale interactive data analyses, *Curr. Protoc. Bioinformatics*, 10.5, 2007.
- [52] BP Vandervalk, L McCarthy, M Wilkinson. CardioSHARE: Web Services for the Semantic Web, *Semantic Web Challenge*, 2008.
- [53] BP Vandervalk, L McCarthy, M Wilkinson. SHARE: A Semantic Web Query Engine for Bioinformatics, *ISWC* pp.367-369, 2009.
- [54] MD Wilkinson, BP Vandervalk, L McCarthy. SADI Semantic Web Services — ‘cause you can’t always GET what you want!, *IEEE Asia-Pacific Services Computing Conference*, pp.13-18, 2009.
- [55] MD Wilkinson, M Links. BioMOBY: an open source biological web services proposal, *Brief. In Bioinform.*, 3(4):331–341, 2002.

A Bioinformatics Approach for Identification of Type III Signal Anchored Proteins in Rice

Amit Katiyar^{1,2}, Shuchi Smita^{1,2}, Dev Mani Pandey², Viswanathan Chinnusamy³ and Kailash Chander Bansal^{1*}

¹National Research Center on Plant Biotechnology, Indian Agricultural Research Institute Campus, New Delhi-110012, India

²Department of Biotechnology, Birla Institute of Technology, Mesra, Ranchi 835 215, Jharkhand, India

³Division of Plant Physiology, Indian Agricultural Research Institute, New Delhi-110012, India

* Corresponding author

Email: kailashbansal@hotmail.com

Phone: 91-11-25843554; Tele Fax: 91-11-25843554

Abstract - Single-pass transmembrane protein (type II, III, and IV) possessing a membrane-spanning domain which targets the protein to the endoplasmic reticulum (ER) membrane. In both type II and III membrane proteins, a single membrane-spanning domain serves both as a signal to initiate insertion and as a membrane anchor. These signal anchor sequences may direct membrane insertion with either an $N_{\text{cyt}}/C_{\text{exo}}$ or $N_{\text{exo}}/C_{\text{cyt}}$ orientation. This study focused on type III proteins, which possess single-anchor sequence but not having N-terminal signal peptide by definition. Type III proteins have the cluster of positively charged residues on the C-terminal side of the signal anchor. The distribution of charged residues flanking the hydrophobic core of the signal sequences play important role in the orientation of signal anchor proteins in membrane. However, the mechanism by which a signal-anchor sequence adopts a particular orientation is still unknown. Here, we performed genome wide screening to identify number of signal anchor proteins in rice genome, which will help to understand the general mechanism of protein orientation in type III membrane proteins.

Keywords: Single-Pas Membrane Protein, Type III Protein, Signal Anchor, Rice

1. Introduction

Subcellular protein sorting, in which proteins travel to their functional organelle within a cell, is an essential feature of cellular life. Typically, protein sorting depends on 'signal' content encoded in their primary structure of the transmembrane proteins. It contains number of hydrophobic and hydrophilic region or domain, which are exposed on one or both sides of the membrane. Single and multiple

membrane-spanning domains containing protein are known as single-pas and multi-pas protein, respectively. Two orientations of signal sequences (NH₂-terminal cleaved or uncleaved signal sequences), have been recognized, which can direct single-spanning membrane proteins to the endoplasmic reticulum (ER) [1]. The NH₂-terminal signal sequences are found on both secreted and membrane proteins [2] and cleaved from the protein by signal peptidase during its translocation across the ER membrane. Second classes of proteins which possess an uncleaved signal sequence target the protein to the ER and stably anchor the protein into the membrane [3, 4]. These proteins are known as signal-anchor (SA) protein and are differ from proteins with a cleaved signal sequence [5]. Two different kinds of orientations ($N_{\text{cyt}}/C_{\text{exo}}$ and $N_{\text{exo}}/C_{\text{cyt}}$), can be possible for single-spanning membrane proteins (Type II, III and IV) during protein targeting towards the ER membrane. In type II proteins the orientation is a luminal C terminus and a cytosolic N terminus, whereas in type III proteins, the orientation is a luminal N terminus and a cytosolic C terminus, which is just the opposite [5, 6]. The type III proteins contain single-anchor sequence but they have lack of N-terminal signal peptide, just like type II proteins. The cluster of positively charged amino acids in type II and III proteins, generally found adjacent to the N-terminal side and on the C-terminal side of the signal anchor sequence, respectively. These positively charged residues change the orientation of proteins in membrane by an uncertain mechanism. Therefore, the identification of new signal anchor proteins helps to determine the biological function and the mechanism of protein orientation. In this study, we did genome wide screening to produce a catalogue of putative signal anchor proteins encoded by the rice genome.

2. METHOD

2.1 Identification of Signal Anchored Proteins

Rice protein sequences for all 12 chromosomes were obtained from “The Institute of Genomic Research (TIGR)” database-version 6.1 (<http://rice.plantbiology.msu.edu/>) [7]. The server TMHMM, Version 2.0 (<http://www.cbs.dtu.dk/services/TMHMM/>) was used to predict transmembrane domain in rice protein sequences. The proteins containing single transmembrane domain within 50 amino acid C-terminal were collected from manual eye inspection. Further, proteins not having N-terminal signal peptides were collected by using SignalP version 3.0 (<http://www.cbs.dtu.dk/services/SignalP/>). The parameters for SignalP v3.0 [8] tool were set as follows: eukaryotes, neural networks and Hidden Markov Model; truncated to first 70 residues. The overall strategy to predict signal anchor proteins were based on *In-silico* approach describe in Figure 1.

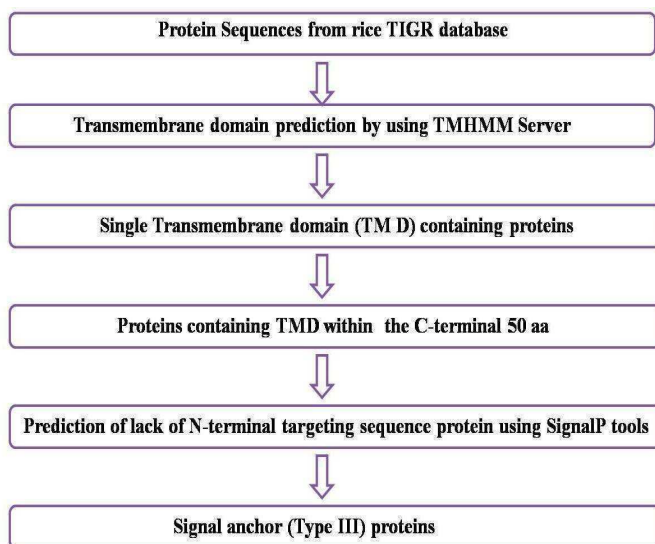


Figure 1: Shows the flow chart of obtaining signal anchor type III proteins.

2.2 Protein Localization and Functional Analysis

The molecular function of predicted signal anchor proteins were analyzed by using TIGR GOSlim Assignments (http://rice.plantbiology.msu.edu/downloads_gad.shtml) and AmiGO blast search (<http://amigo.geneontology.org/cgi-bin/>). The functional annotation was based on TIGR annotation release 6.1 and blast search against *Arabidopsis*, while subcellular localization was based on validated experimental data available on TIGR.

3. Results and Discussion

3.1 Genome-wide identification of Signal Anchored Proteins in rice

Single-spanning transmembrane proteins which contain signal anchor sequence but lack N-terminal peptide sequence are known as type III signal anchor protein (also known as reverse signal anchor) [9]. The function of a signal sequence was reported as targeting to the membrane, membrane insertion and translocation (secreted proteins) or retention (SA proteins). A bioinformatics approach has been previously applied to identify various transmembrane, tail anchored proteins in human, yeast and *Arabidopsis*. Here, we performed genome wide identification of single-pass signal anchored proteins (Type III) in rice using various computation tools. Signal anchor proteins in plants helps to understand general mechanisms about the changes occurred in the protein orientation in membrane. Thus, catalogue preparation of signal anchor proteins encoded by the rice genome is an important step to unravel the biological function. As a first step, we obtained 56,797 protein sequences from TIGR release 6.1 databases. Further, we identified proteins contained single transmembrane domain (TMD) by using transmembrane helix prediction server and we found 5,317 protein members. The next step descends the protein sequence up to 9, 36 as we extracted only those protein members having transmembrane domain within the C-terminal 50 residues. This step identified number of proteins with single TMD near to C-terminus which further required to process for knowing N-terminal signal peptide and if found discarded from list. Remaining 54 protein members contain single transmembrane helix (TMhelix) and lack N-terminal signal peptides are known as signal anchor proteins (Type III) and collected as signal anchor protein catalogue for rice. The gene ontology provide the molecular function of signal anchor proteins and we observed that majority of the protein members were involved in membrane protein transporter activities (Table 1). This exercise will help to understand biological functions of SA proteins in more detail.

Table 1: Shows the molecular function and localization for signal anchored proteins in rice

Rice TIGR Locus	Protein Description	GO-Molecular Function	GO- ID	GO-Cellular Component
LOC_Os01g03850	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os01g06750	verticillium wilt disease resistance protein precursor,	Nucleotide binding	GO:0000166	Plasma membrane
LOC_Os01g09270	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os01g15029	Expressed protein	Peptidyl-prolyl cis-trans isomerase activity	GO:0003755	Unknown
LOC_Os01g38070	Retrotransposon protein	Unknown	Unknown	Unknown
LOC_Os01g38510	protein transport protein SEC61 subunit beta	Transporter activity	GO:0005215	Unknown
LOC_Os01g41860	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os01g51240	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os01g54424	Expressed protein	Unknown	Unknown	Unknown
LOC_Os01g55860	Expressed protein	Unknown	Unknown	Mitochondrial
LOC_Os01g67960	GPI transamidase component	Unknown	Unknown	Endoplasmic reticulum
LOC_Os02g03880	Hypothetical protein	Protein transmembrane transporter activity	GO:0015450	Mitochondrial
LOC_Os02g08180	protein transport protein SEC61 subunit gamma	Protein transmembrane transport activity	GO:0006810	Unknown
LOC_Os02g10370	hrpN-interacting protein from Malus	Unknown	Unknown	Unknown
LOC_Os02g11705	Expressed protein	Unknown	Unknown	Unknown
LOC_Os02g17590	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os02g29400	Expressed protein	Protein transmembrane transporter activity	GO:0015450	Mitochondrial
LOC_Os02g32009	Expressed protein	Unknown	Unknown	Unknown
LOC_Os02g35610	Expressed protein	NADH dehydrogenase (ubiquinone) activity	GO:0005739	Mitochondrial
LOC_Os02g46830	Expressed protein	Unknown	Unknown	Unknown
LOC_Os03g02620	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os03g10160	Pentatricopeptide repeat-containing protein	Structural constituent of ribosome	GO:0003735	Chloroplast
LOC_Os03g12970	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os03g14334	Expressed protein	Unknown	Unknown	Unknown
LOC_Os03g38359	Expressed protein	Unknown	Unknown	Unknown
LOC_Os03g56784	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os04g44760	Hypothetical protein	Carbohydrate transmembrane transporter activit	GO:0015144	Membrane
LOC_Os04g50780	Expressed protein	Unknown	Unknown	Unknown
LOC_Os05g28020	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os05g30830	Retrotransposon protein	Unknown	Unknown	Chloroplast
LOC_Os05g42010	ubiquinol-cytochrome c reductase complex	Unknown	GO:0005739	Mitochondrial
LOC_Os05g50654	Mitochondrial import receptor subunit TOM7-1	Transporter activity	GO:0005215	Unknown
LOC_Os06g14340	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os06g15450	Retrotransposon protein	Iron ion binding	GO:0005506	Membrane
LOC_Os06g16620	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os06g21420	Retrotransposon protein	Unknown	Unknown	Unknown
LOC_Os06g23330	Conserved hypothetical protein	Unknown	Unknown	Unknown

LOC_Os06g44374	protein transport protein SEC61 subunit gamma	Protein transmembrane transport activity	GO:0006810	Unknown
LOC_Os07g04270	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os07g05084	Expressed protein	Transferase activity	GO:0016740	Unknown
LOC_Os07g31194	Expressed protein	Unknown	Unknown	Unknown
LOC_Os07g48244	Ubiquinol-cytochrome c reductase protein	Unknown	Unknown	Mitochondrial
LOC_Os08g03980	Retrotransposon protein	Unknown	Unknown	Unknown
LOC_Os08g06260	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os08g06270	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os08g13120	Retrotransposon protein	Unknown	Unknown	Unknown
LOC_Os08g41680	Expressed protein	Unknown	Unknown	Unknown
LOC_Os09g06930	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os09g25130	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os09g29640	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os10g40010	Expressed protein	Binding activity	GO:0005488	Chloroplast
LOC_Os11g41875	Membrane protein	Unknown	Unknown	Unknown
LOC_Os12g15390	Hypothetical protein	Unknown	Unknown	Unknown
LOC_Os12g32950	Membrane protein	Unknown	Unknown	Unknown

4. Conclusions

Our work provides the list of type III signal anchor proteins from rice genome that not having N-terminal targeting peptide but contain signal sequence and involved in protein targeting to the endoplasmic reticulum (ER). Gene ontology of SA proteins facilitates to understand molecular function of SA proteins. This work forms the foundation for molecular genetic and biochemical analysis that will help understand of the biological function of Type III signal anchor proteins in rice.

5. Acknowledgements

AK thanks to Indian Council of Agricultural Research (ICAR) for supporting this work by the ICAR-sponsored Network Project on Transgenic in Crop (NPTC).

6. References

- [1] Wickner, W. T., and H. F. Lodish (1985) Multiple mechanisms of protein insertion into and across membranes. *Science (Wash. DC)* **230**: 400-407.
- [2] Walter, P., and V. R. Lingappa (1986) Mechanism of protein translocation across the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* **2**: 499-516.
- [3] Spiess, M. and Lodish, H. F. (1986) An internal signal sequence: the asialoglycoprotein receptor membrane anchor. *Cell*, **44**: 177-185.
- [4] Zerial, M., P. Melancon, C. Schneider, and H. Garoff (1986) The transmembrane segment of the human transferrin receptor functions as a signal peptide. *EMBO (Eur. Mol. Biol. Organ.) J.* **5**:1543-1550.
- [5] Lipp, J., and B. Dobberstein (1988) Signal and membrane anchor function overlap in the type II membrane protein 13, CAT. *J. Cell Biol.* **106**:1813-1820.
- [6] Holland, E. C., J. O. Leung, and K. Drickamer: Rat liver asialoglycoprotein receptor lacks a cleavable N-terminal signal sequence (1984) *Proc. Natl. Acad. Sci. USA*, **81**:7338-7342.
- [7] Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek R L, Lee Y, Zheng L, Orvis J, Haas B, Wortman J and Buell C R (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Research* **35**: D883-D887
- [8] Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**: 783-795
- [9] Von Heijne, G. and Manoil, C (1990) Membrane proteins—from sequence to structure. *Protein Eng.*, **4**:109-112.

Application of W-curves and TSP to Clustering HIV1 Sequences

Douglas Cork^{1,2,3}, Steven Lembark⁴, Nelson Michael^{1,5}, Jerome Kim^{1,5}

US Military HIV Research Program¹; Henry M. Jackson Foundation for the Advancement of Military Medicine², Rockville, MD. 20850, BCPS Dept., Illinois Institute of Technology³, Chicago, IL. 60616, Workhorse Computing⁴, Woodhaven, NY., Walter Reed Army Institute of Research⁵, Rockville, MD.

Abstract - *The high mutation rate in HIV-1 makes it difficult to treat and analyze. Monitoring the evolution of drug resistance requires frequent re-sequencing, but comparing and visualizing the progress is difficult. One difficulty is simply locating the areas of interest: gaps and crossover mutations make it difficult to isolate clinically significant sequences for comparison. Effectively displaying the results of comparisons grouped according to multiple regions is also a problem. Our comparison algorithm based on the W-curve helps automate the comparison process, producing results suitable for clustering via a modified solution to the Traveling Salesman Problem (“TSP”). Appropriate color-coding of the TSP results allows us to display the results of multiple comparisons effectively for single samples or time-series. The results can be useful for providing guidance in treatment, analyzing the membership in anonymous study populations, tracking the evolution of drug resistance in populations, or rates of co-infection within study groups.*

Key words: HIV-1 Genomic Clustering, W-curve, Traveling Salesman Problem, Drug Resistance

1 Disclaimer

The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting true views of the Dept. of the Army or the DoD.

2 Introduction

One step in managing HIV infection response is monitoring individuals and study populations for evolution of drug resistance. Given *in vivo*

replication kinetics with more than 10^9 new cells infected every day, each and every possible single-point mutation occurs between 10^4 and 10^5 times per day in an HIV infected individuals [1]. This high level of variation leaves identifying the markers for drug resistance difficult, requiring multiple manual steps in many cases. Especially for population studies, an automated process that can identify and compare drug resistance sites would be an enormous help. The W-curve's scoring process produces a set of localized comparison values that can be used as landmarks that can be used to automate the process of locating relevant areas for comparison. Clustering new samples with known ones using the Traveling Salesman Problem (“TSP”) can automate the grouping of new samples into drug-resistance categories. Combining them leads to an automated process for tracking medication in individuals or analysis of groups.

3 Background

Analyzing the data in many HIV studies is made difficult by a combination of HIV's genetics and the sources of data. Patients in many HIV clinics are anonymous, often because they are illegally engaged in prostitution, illegal drug use, or homosexual sex. Studies of groups look for new strains and how they spread and have to check for changing members of the sample population; individual patients have to be sequenced frequently to evaluate appropriate drugs.

HIV's high mutation and crossover rates combined by discontinuous sequences for the drug responses make the analysis difficult. The high mutation in areas between the regions of interest confound any analysis based on whole genes or regions:

there is simply too much white noise on the genome to use large portions of it for comparison. High rates of mutation and co-infection leave many individuals with multiple strains of HIV, further confounding the analysis, and HIV strains have frequent crossover mutations, making things even worse. As an example, studies of cross-clade neutralization produce effectively random results [2].

Treatment of HIV is still largely a manual process: patients have to be sampled and sequenced frequently and doctors have to make informed decisions on how to deal with evolution of the strains infecting them. Presenting sequence comparison results is particularly important: physicians have to evaluate the similarity of a single individual to multiple known drug-resistant strains.

In the US 98% of the infections are type-B, and the single FDA-approved program for treatment using genetics treats type-B only. The U.S. Army has to treat soldiers who acquire HIV all over the world and only 85% of their cases are type-B, leaving them without any good options for 15% of their cases. Being able to at least classify the non-B cases and evaluate their treatment outcomes effectively would be a huge help.

4 Methodology

4.1 W-curves

The W-curve was originally designed as the basis for a graphical tool for visualizing very large regions of

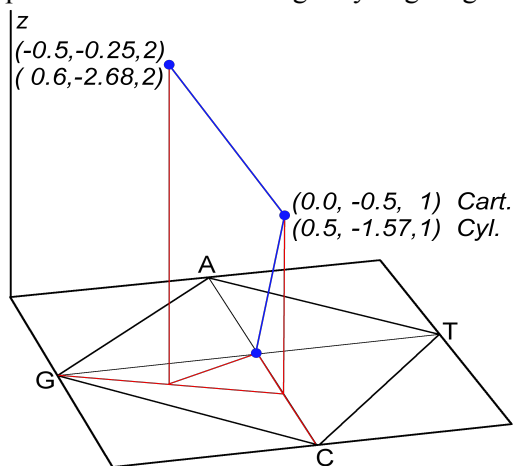


Figure 1: W-curve generation showing initial points for curve "CG" [4].

DNA [3]. Over time it has been adapted for numerical comparison as well [4,5,6]. By abstracting the genetic sequences into a three-dimensional space, W-curves offer a wider range of comparisons rather than comparisons based on searching, aligning and tree-building (assuming a mutation model) with uni-dimensional strings of characters. W-curves make it easier to find patterns in sequences or locate common features between genes. Their 3D nature also makes it possible to align smaller features than are possible with string based techniques. The design of our comparison utilities builds on these capabilities to provide a technique for matching the small regions of HIV-1's CD4 epitopes along its *gp120* gene [7].

The order of bases along the corners is significant: number of hydrogen bonds (2 or 3) and chemical structure (purine or pyrimidine) share quadrants around the square. This means that most synonymous SNP's in the gene sequences will leave the curves in the same quadrant. This keeps our same-quadrant measure small for SNP's.

We use a two-pass process for comparing W-curves. The first pass produces an array of alignment regions with starts, stops, and a difference measure that we call "chunks". SNP's increase the difference measure, gaps show up as differences in the starting values on successive chunks of the comparison, indels as successive chunks with no change in the relative offset. The second pass summarizes the chunks into whatever measures are useful, for example by averaging the differences over the length of a sequence.

The advantage of chunking the results first is that similar chunks can be used to locate landmarks for aligning sequences with one another. Areas with small differences provide an automated means of locating the offsets between start and stop values in the sequences. Given a library of sequences with known landmarks such as points of known drug resistance, we can score a single incoming sequence against all of them.

4.2 TSP and Clustering

The Traveling Salesman Problem ("TSP") is quite easy to describe but difficult to solve. The problem is to take a list of distances between cities and make a tour of them which visits each city once

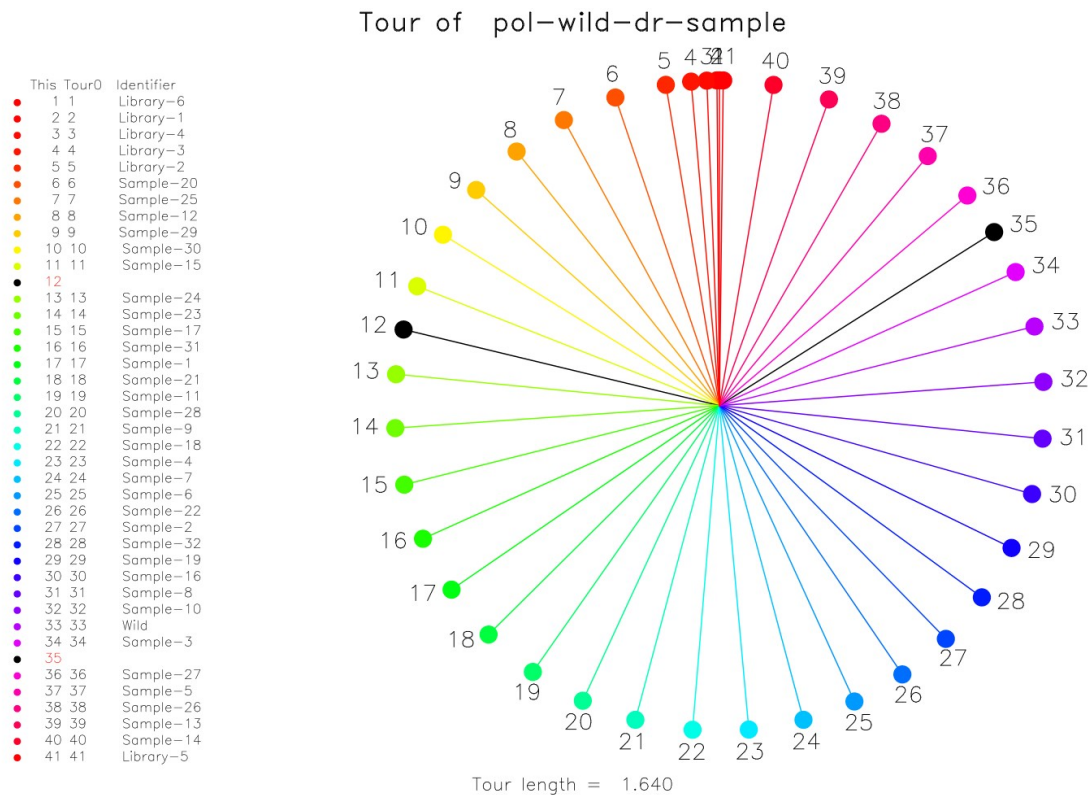


Figure 2: Example of tour generated by R's TSP package comparing POL sequences of the wild type with a library of simulated drug-resistant (“dr”) and random sample strains of HIV-1 [6]. The library strains cluster tightly together at the top (red) with the wild strain (#33) outside the drug-resistant cluster to the right. The approach shown here works well if individuals are sampled over time: pre-assigning the colors from a first tour makes it relatively easy to watch if samples from individuals drift into different groups.

with the least total distance. Substitute cost and the algorithm will find the “cheapest” route through all of the cities. As it turns out, this problem is NP-complete, requiring analysis of all routes to guarantee the least distance. Much work has gone into developing heuristics for solving this problem and there are fast algorithms for approximating the solution.

The utility of TSP problems is that an optimal tour will cluster the closest cities together. If the difference measures are for genes, they can similarly be clustered on any region of interest. A number of techniques for determining inter- and intra-clade distances have been developed. One technique developed by Climer and Zhang at Washington University is to add a fixed number of “dummy

cities” to the list [8]. Each dummy has a small distance to all other cities (we use 2^{-20}). The non-zero distance leaves these cities in the intra-cluster gaps. We display the resulting tours as color-coded pinwheel diagrams. Appropriate color-coding makes these relatively easy to analyze individually or compare to one another.

4.3 Generating and Analyzing TSP Clusters with R

The R statistical package includes a TSP library, available from CRAN. We have used it here to generate an approximate solution for clustering the genes. In our case an optimal tour is not required: any good approximation will cluster the genes properly. Our approach starts with a square

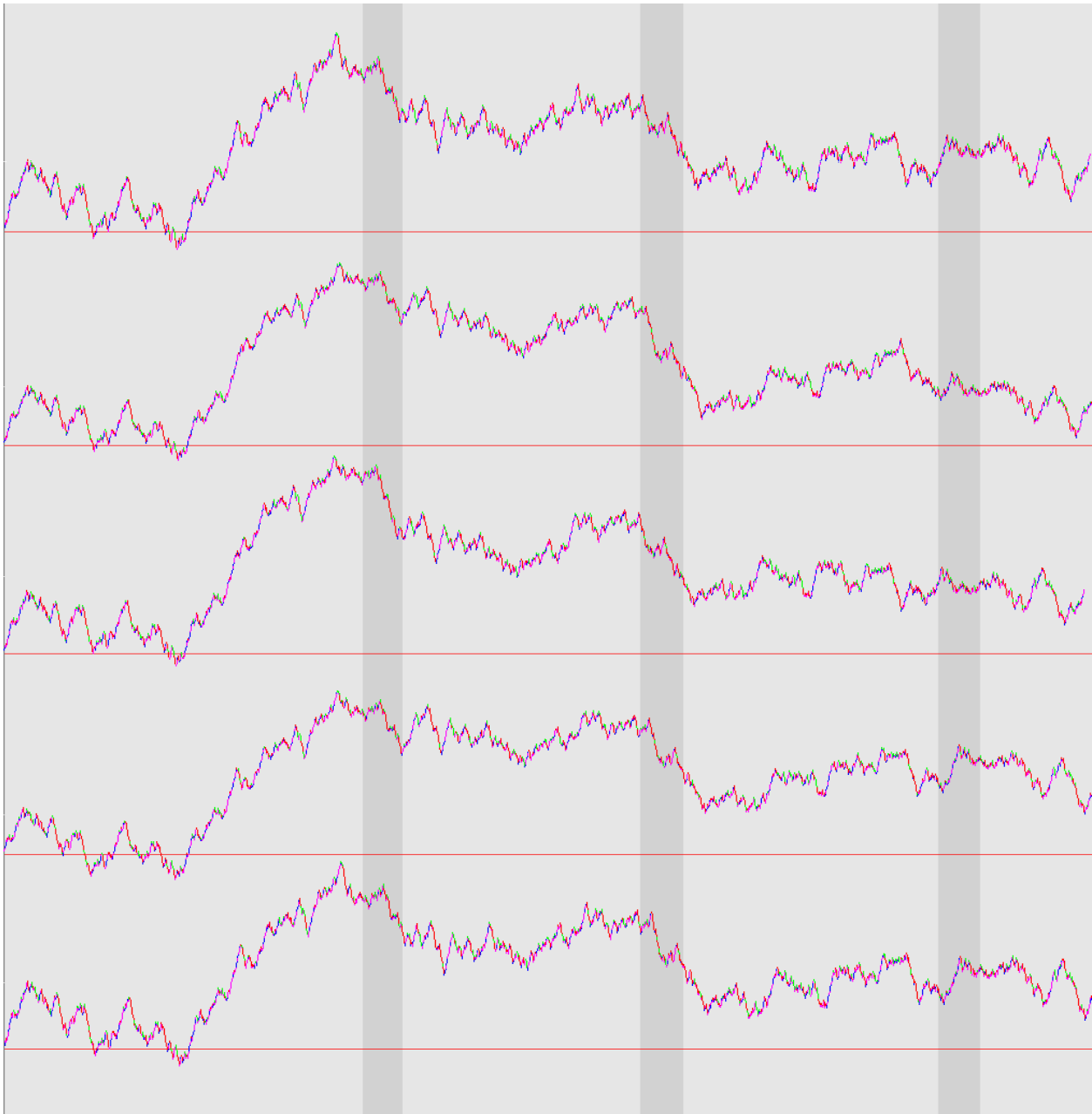


Figure 3: Example W-curves POL gene from wild (top) and example drug resistant HIV-1 strains [10]. Local features of the W-curve geometry can provide landmarks for isolating smaller regions of the sequences for comparison [6]. Highlighted regions show some areas with more easily visible variations between the curves.

distance matrix having zeros on the diagonals. The TSP package does not require a symmetric matrix but in our case we use one.

The colors shown in Figure 2 are assigned by generating a list of 1024 colors and rotating the tour

so that it starts with the first library sequence. After that the colors are assigned by taking the fractional tour length times 1024. For example, if the total tour length were 20 and one of the sequences fell at a cumulative distance of 9 then it would be assigned a color of $9 / 20 * 1024$, or 460.

We have found that this works better than simply assigning the colors sequentially from one to the sample size. Assigning the colors by position along the tour gives some additional visual feedback from the similarity of colors. This can be seen in Figure 2 where the closely related “library” sequences are all red. We have found this color scheme also works well for comparing tours generated from different sections of the same genes: even if the nodes appear in different orders on different graphs the similar colors grouping together help identify the clusters.

A slightly different color scheme can be used to observe the evolution of drug resistant strains in a population. In that case coloring the time scale allows us to watch which direction the group is evolving. With the library sequences colored red and the population samples green through blue over time, the migration of blue dots towards library samples is easy to pick out. The library samples can be drug resistant, or susceptible to different drugs. Either way, the progression across clusters is relatively easy to view.

This approach also works well for integrating samples of an individual over time. We can display results of comparing various regions of an individual to a library of sequences with known clinical results. With the library samples in one color and the individual's sequence of samples colored over time we can see how the various samples migrate between clusters. This provides a nice way to integrate treatment information about an individual that may rely on samples of unrelated genes.

4.4 Combining the TSP and W-Curve

The TSP approach shown here will work for any comparison mechanism: ClustalW, Fasta, or Blast provide a suitable square comparison matrix and generate the graphic results from R. Our use of the W-curve has an advantage due to chunks: we can automate scoring discontinuous regions that may have differing alignments. This matters with HIV-1 where the high rate of SNP's leaves too much white noise in larger areas and the high rate of gaps makes locating the often small, discontinuous areas causing drug resistance difficult. The curve's geometry also provides us a more feature-rich environment in

which to compare the curves. Figure 3 shows the wild (HXB2 standard) and five drug resistant sequences [9,10]. Differences in geometry are visible, even when viewed at different scales [6]. The geometric representation also offers us more options for approximate matching using discrete spatial mathematics than string comparison techniques allow.

5 Conclusions

The W-curve provides us with a way of using landmarks to identify regions of interest and score only the relevant portions of sample sequences. The TSP with Climer & Zhang's boundary techniques offers a fast, effective way to cluster genes. The R statistical package provides us with the tools to analyze and color-code the results for analysis. Taken together this provides us with a useful tool for comparing the status and evolution drug response for HIV-1.

6 Acknowledgments

Military HIV-1 Research Project (MHRP)/Henry Jackson Foundation (HJF) and Workhorse Computing. This work was funded by MHRP (Military HIV-1 Research Project) cooperative agreement (W81XWH-07-2-0067-P00001) between the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., and the U.S. Department of Defense (DOD). The cohort work was supported through NICHD by R01 HD34343-03 and partly by a cooperative agreement between the Henry M. Jackson Foundation for the Advancement of Military Medicine and the U.S. Dept. of Defense. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors wish to thank Pranamee Sarma and Scott Zintek for their assistance preparing the W-curve graphics used in this paper.

7 References

- [1] Coffin JM, (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 267, 483-489.
- [2] Brown BK, Sanders-Buell E, Rosa Borges A, Robb ML, Birx DL, et al. (2008) Crossclade neutralization patterns among HIV1 strains from the six major clades of the pandemic evaluated and compared in two different models. *Virology* 375: 529–538.
- [3] Wu D, Roberge J, Cork DJ, Nguyen BG, Grace T (1993) Computer visualization of long genomic sequences, in *Visualization 1993*, IEEE Press, New York City, New York, CP 33:308–315.
- [4] Cork DJ, Lembark S, Tovanabutra S, Robb ML, Kim JH (2010) W-curve Alignments for HIV1 Genomic Comparisons. *PLoS ONE* 5(6): e10829.
doi:10.1371/journal.pone.0010829
- [5] Cork D, Lembark S, Tovanabutra S, Sanders-Buell E, Brown B, Robb M, Wiczorek L, Polonis V, Michael N, Kim J. Application of W-curves and TSP to Clustering HIV-1 Sequences by Epitope. 847-853
<http://www.cs.uga.edu/~hra/2010-proceedings-final/biocomp/volume-i.pdf>
- [6] Supplemental data and code used to generate the W-curves and TSP results shown here are available at
<http://www.bioinformatics.org/wcurve/>.
- [7] Zhou T, Xu L, Dey B, Hessel AJ, Van Ryk D, Xiang SH, Yang X, Zhang MY, Zwick MB, Arthos J, Burton DR, Dimitrov DS, Sodroski J, Wyatt R, Nabel GJ, Kwong PD (2007) Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Feb* 15;445(7129):732-7.
- [8] Climer S, Zhang W (2004) Take a walk and cluster genes: a TSPbased approach to optimal rearrangement clustering. *ACM International Conference Proceeding Series*; Vol. 69, p. 22.
<http://portal.acm.org/citation.cfm?id=1015419>
- [9] Leitner T, Korber B, Daniels M, Calef C, Foley B. (2005) HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences., <http://hiv.lanl.gov>
- [10] Stanford HIV RT and Protease Sequence Database. See supplemental data for specific resistant drug studies [6].
<http://hivdb.stanford.edu/pages/genotype-rx.html>

Electronic Health Records and Mobile Technology Use in Northwest Florida Medical Practices

E. Rodgers, Ph.D.

Department of Computer Science, University of West Florida, Pensacola, FL, USA

Abstract - *This paper describes the ongoing analysis of surveys and on-site visits to review the characteristics of EHR software and mobile technology used in selected rural hospitals, medical clinics, and practitioners in Northwest Florida. The paper follows up on previously published findings and compares the results to state-wide and nation-wide statistics for urban and rural practices.*

Keywords: Mobile Technology, EHR

1 Introduction

Even with its reputation as a technology innovator, the United States lags the rest of the world in the medical use of mobile technology and the use of Electronic Health Records (EHR). Researchers have found that 99% of the physicians in the Netherlands use EHRs, and the EHR adoption rate for Australia, Italy, Norway, Sweden, and the United Kingdom is 94% or higher. The U.S. adoption rate of 17% to 30% was only higher than that for Canada among the world's leading economies [1].

1.1 Rural Healthcare Characteristics

The quality of healthcare in rural areas of the United States has steadily declined over the past 25 years. The rural population of the U.S. represents about 25% of the total U.S. population, but only 10% of the total number of U.S. physicians practice in rural areas. In addition, rural areas of the U.S. have lost over 500 hospitals over the past 25 years through consolidation and economic hardships. Not only does this void result in reduced coverage and longer commutes for treatment, but it also reduces access to specialized treatment for rural citizens. Ironically, at a time when the use of mobile technology and the use of EHRs could help make up for these losses in rural coverage, the U.S. medical community still lags the world in such meaningful use [2].

2 How Mobile Technology Can Improve Rural Healthcare

It has been shown that mobile technology can improve the quality of rural healthcare. Mobile technology includes the use of portable computers, Personal Digital Assistants (PDAs), mobile phones and smart phones, global positioning system (GPS) devices, and other wireless hardware.

2.1 Examples of Mobile Technology Usage in Healthcare

Some medical applications of mobile technology include remote patient monitoring, remote diagnosis and treatment, on-site diagnosis and prescribing, physician, nurse, or hospital to patient communication, and online patient medical record storage and retrieval. For example, in a study in rural Washington State, nurse practitioners used PDAs loaded with a pharmacology program and medical decision making software. The PDAs provided access to medical information not normally available in such remote areas. The use of the PDAs was compared to traditional means with respect to participants' responses to ease and speed of access, speed of response, decreased need to use the Internet to seek clarification or to ask questions of others, increased understanding of their roles and responsibilities, and recognition of real world practices. Evaluation of participant responses from the group using PDAs yielded significantly better results than the group using traditional methods. In addition, during the test, information gained and the use of the PDA's decision making software enabled two patients with critical problems to be diagnosed whose conditions or drug interactions would have otherwise been missed [3].

2.2 Mobile Technology Usage and EHRs can also Reduce Medical Errors

The use of mobile technologies and EHRs can play an even more important role than improving the quality of rural healthcare. It has been reported that medical errors are the fifth leading cause of death in the U.S. [4]. The uses of mobile technology and EHRs have the potential to significantly reduce errors due to incorrect incomplete, misunderstood, or missing information. These technologies can provide more consistent data entry and effective interfaces to ensure the integrity of medical information. Sharing of patient information among practitioners would also be timelier.

3 Status of Mobile Technology and EHR use in Rural Northwest Florida

Table 1 indicates that physicians in rural areas of Florida are less likely to use EHRs, mobile technology to communicate with patients, and PDAs for medical applications than their urban counterparts. Rural and urban Florida physicians, however, used mobile computers at about the same rate. Rural NW Florida physicians showed about the same rate of use as other rural Florida physicians in all categories [1], [5].

Table 2, which compares rural NW Florida to the U.S., indicates similar results as those shown in Table 1, which compares rural NW Florida to other areas of the state. [1], [6], [7].

Table 1. Percentage of Florida Practices Using EHR and Mobile Technologies compared to Rural NW Florida.

Technology Usage	Florida Urban	Florida Rural	NWFL Rural
EHR	24%	17%	15%
Patient communication	17%	8%	6%
Mobile computer	81%	77%	78%
PDA	38%	32%	31%

Comparing the Tables, urban Florida physicians and urban U.S. physicians use EHRs and mobile technology for patient communication at about the same rate, but Florida urban physicians use mobile computers and PDAs less. Rural NW Florida physicians and rural U.S. physicians showed about the same rate of use in all categories.

Table 2. Percentage of U.S. Practices Using EHR and Mobile Technologies compared to Rural NW Florida.

Technology Usage	U.S. Urban	U.S. Rural	NWFL Rural
EHR	23%	17%	15%
Patient communication	19%	7%	6%
Mobile computer	85%	80%	78%
PDA	43%	30%	31%

4 References

- [1] E. Rodgers & K. Rodgers. "Use of Mobile Technology in Rural Medical Practices"; *Proceedings of the Western Decision Sciences Institute*. (Apr 2011).
- [2] NHRA. "What's Different about Rural Healthcare?"; *National Rural Health Association*. <http://www.ruralhealthweb.org/go/left/about-rural-health> (2009).
- [3] M. Rice. "Enhancing Rural Health Care Using PDAs"; *PDA Cortex*. http://www.pdacortex.com/Rural_Health_Education_Using_PDAs.html. (n.d.).
- [4] P. Rajendran. "Ethical Issues Involved in Disclosing Medical Errors"; *JAMA* 286(9) 1078. (2001).
- [5] N. Menachemi, A. Langley, & R. Brooks. "The Use of Information Technologies Among Rural and Urban Physicians in Florida"; *Journal of Medical Systems* 31(6) 483-488. (Aug 2007).
- [6] C. Hsiao et al. EMR/EHR Systems of Office-Based Physicians: United States, 2009 and Preliminary 2010 State Estimates. *CDC/NCHS*. http://www.cdc.gov/nchs/data/hestat/emr_ehr_09/emr_ehr_09.pdf. (Dec 2010).
- [7] Poon, E. et al. Assessing the Level of Healthcare Information Technology Adoption in the United States. *BMC Medical Informatics and Decision Making*. <http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1343543&blobtype=pdf>. (Jan 2006).

Anode Effects On Microbial Fuel Cell Efficiency

M. Brant³, G. Chu⁶, M.W. Claire², J. Curnutt¹, E. Gomez¹, A. Gonzalez⁴, C. Gott⁴, M. Grigsby⁴, R. Hovanesian⁴, G. Kaladjian⁴, J. Losch³, A. Nguyen, A. Olano⁴, G.W. Payton⁴, A. Razzak⁴, K. Rotunno⁴, S. Saleemi⁴, A. Scheppelmann³, K.E. Schubert¹, G. Solis⁴, E. Statmore⁵, K. Symer³,

¹School of Computer Science and Engineering, California State University, San Bernardino,

²Virtual Planetary Laboratory, University of Washington ³California Polytechnic State University, San Luis Obispo

⁴California State Polytechnic University, Pomona ⁵San Francisco State University ⁶NASA Ames

Abstract—*The use of microbial fuel cells to detect anaerobes and power experiments in remote extreme environments is examined by a team of scientists and student teachers in a joint venture of the California State University system and NASA Ames. The economic viability of carbon fiber electrodes is tested by comparing their performance to graphite rods, which are commonly used. A connection between concentration of life and voltage output is indicated, which provides a potentially easy and powerful test for the presence of life in extreme environments such as found on Mars. Pollution free but low power density microbial fuel cells are shown to demonstrate viability as a power source for sensing applications.*

Keywords: Life in extreme environments, microbial fuel cell, life on Mars

1. Introduction

Two great questions confront scientists working in astrobiology and extreme environments. The first is how to detect life outside our planet, and the second is how to power equipment in extreme environments. In this paper we suggest a potential solution to both problems.

1.1 Detecting Life in Extreme Off-world Environments

Recent studies, see [4], have cast doubts on the interpretation of the 1976 experiments conducted by Viking, which indicate no life in the soil samples analyzed. The basic problem with Viking was that its experiments were destructive to organics, and required chlorine to be in a chloride salt form, when it turns out from Mars Phoenix Lander that chlorine is in perchlorate form. The new Mars Science Lab to be launched soon will hopefully clarify some of the results, but still the question remains as to what is a good way to detect life on another planet that has an extreme environment.

Ideally, we would like to submit the samples to a wide range of tests, but this is economically not viable. Detection of life is dependent on the form the life takes, so to detect life

with a minimum of equipment requires some assumptions to be made. We postulate a few simple ones.

- First, if Mars had life, it was carbon based. Other elements like silicon has been suggested but we have no idea exactly what such a life would look like and thus it would be very hard to detect, and harder to prove.
- Second, if Mars had carbon-based life, it likely had microbial life somewhat similar to earth. This is reasonable for a variety of reasons, including ease of formation, durability, necessity in an ecosystem, possibility of cross-fertilization from impacts, and if not true it is unlikely we could guess the exact nature of life on Mars.
- Third, if Mars had microbial life, it likely had a range of anaerobes, as the atmosphere likely had higher carbon dioxide and less free oxygen than Earth.
- Finally, if Mars had anaerobes, metal reducing ones similar to *Geobacter sulfurreducens* were likely to exist, as they are widely dispersed on earth, and would fit the Mars well.

Metal reducing anaerobes could be present in the ground of Mars, and could even still be alive in Martian lava tubes, which would provide a viable system for them to still survive. Testing for the presence of metal reducing anaerobes, is thus a logical step to verifying if Mars has or had life.

1.2 Powering Equipment in Extreme Environments

One of the greatest challenges of exploring extreme environments is the difficulty of finding long-term, self-sustaining, and indefinitely sustainable power sources for the kinds of equipment that may need to remain in place for extended periods under incredibly inhospitable conditions. In the case of a Mars rover, for example, scientists and engineers can leverage the power of the sun using solar panels to charge and recharge on-board batteries. In other situations anchored to the ground, wind can be harnessed to provide a renewable energy source. Underwater, near the deep ocean thermal vents, engineers and scientists can

access the heat rising from the Earth's core to power sensors or equipment for exploration within a certain range. As scientists and engineers begin to explore more and more extreme and remote regions that are lacking in these more obvious energy sources, a new question arises: specifically, what types of practical, cost-effective, self-sustaining power sources can we tap into in regions like the bottom of Earth's oceans - the kinds of places where sunlight does not reach, pressure is unbearable, oxygen is scarcely available, and our ability to fix or replace parts is extremely limited?

One of the most promising energy sources for this kind of exploration is so simple and omnipresent it might easily be overlooked: namely, the energy generated by the microbial life forms that can be found anywhere from the deepest reaches of the ocean to the shallowest mud puddle in a suburban back garden. If we could design and engineer a fuel cell that harnesses the energy of microbes in the deep ocean - or anywhere else in the Universe where they might be found - we could conceivably develop simple power cells fueled by microbes going about their daily business of metabolizing glucose, acetate, other organic materials, or even metals. And we could potentially do so for extended periods - for as long as a power source is needed.

1.3 Science Education

This paper presents the preliminary results of the work carried out by a group of science and math oriented student teachers under the supervision of scientists from several universities and two NASA labs, as part of the NASA/CSU Spaceward Bound Project funded by the California State University's Mathematics and Science Teacher Initiative Program. A major subgoal of this paper is to demonstrate how actual science can be performed in classrooms, and to instigate further classroom projects in this vein.

2. Experimental Setup

The purpose of this experiment was to test the viability of detecting low density anaerobes as well as to compare the potential energy yields and cost-benefit models that could be achieved using different forms of carbon anodes - in this case, comparing the efficacy of four one-half inch diameter graphite rods and six yards of six-inch woven carbon fiber tape. By deliberately starting from materials that are both abundant and inexpensive, our goal was to refine the design parameters for this new type of microbial fuel cell so that it could be more quickly, easily, and cheaply manufactured and deployed to remote locations anywhere they might be useful, from the deep mud of Earth's ocean floor to the potential use on Mars, Europa, or anywhere else.

The basic design of a battery is elegantly simple. It has two key parts: the anode and the cathode. In a bacterial battery, microbes attach to the anode. As they metabolize food, they create energy which wants to be released in oxygen. Since there is no oxygen at the anode, the energy

transfers to the cathode through a wire. The movement of energy along the wire from the anode to the cathode creates a current which can be harnessed as usable energy. The use of microbes in this situation utilizes the chemical energy produced by the microbes and converts it directly into electricity as well as converting substrate into a source of electrons to complete the battery circuit. The bacteria are kept on the anode along with their food source. They convert their food source (often glucose) into CO₂, protons and electrons. The protons and electrons are then used for energy for the battery. Several different food sources have been tested including waste water, organic waste stream, and recycling waste water. All of these are rich in nutrients and provide amiable conditions for bacterial growth.

Although the concept of a microbe powered battery is wonderful in theory, in practice we needed to focus on inexpensive, widely available materials that could be used to amplify the amount of power being created. Typically, researchers have used graphite to make these batteries. For this experiment, four graphite rods were used (1/18 m²) on one battery and 6' of carbon fiber (2 m²) on the second battery. It is expected that the graphite rods generate more current energy per unit area but the expected increased area of carbon fiber will make a more cost efficient product. The carbon fiber is expected to have a lower cost per V-A than the graphite rods.

One very promising design has been researched and prototyped by a research group at The Pennsylvania State University. They have done work using both one and two chambered microbial fuel cell batteries as well as flat plate and salt bridge designs. All designs are tested and compared by the researchers in an effort to discover the most reliable design for production of energy with microbes. The microbes they use come from waste water, air, or soil. All bacteria are grown on graphite rods, or plates that provide a food source for the microbes.

Another design with great potential has been developed by the Geobacter Project, based at the University of Massachusetts, Amherst. Their design emphasizes their discovery of a strain of Geobacter with increased capacity for power production in microbial fuel cells which can be utilized in fuel cells as well as give insight for the mechanisms by which long range electron transfer operate in biofilm. Dr. Lovely and his fellow researchers hold that Geobacter sulfurreducens is a good choice to study electron transfer and power production under high pressure. They are working toward creating selective pressures to drive the new strain of bacteria to evolve that prefers to grow on the anodes of the batteries so that they can be more efficiently utilized in fuel cell design.

If it can be proven cost-effective and energy-efficient, this new energy source could be even more promising than its predecessors. Our research is meant to help advance the conversation about these new energy sources and to

demonstrate the ability [viability] of such energy sources and materials to support deep sea mud batteries as an energy source for exploration in extreme, remote, and muddy places.

2.1 Graphite Rods

In this experiment, we used a battery constructed of graphite rods instead of fabric as a control, since batteries of this type have been constructed before. We used a series of four graphite rods, available at most home improvement stores, as electron sources. When placed in anoxic mud, anaerobic bacteria will grow on these rods, producing free electrons as byproducts of their metabolism.

Each rod has a cylindrical shape and is about 30.5 cm long and 1.27 cm in diameter. The surface area of each rod, excluding the tops, is equal to $2\pi rh$ where r is the radius and h is the height. Since we know the diameter, d , the surface area is given by πdh . Therefore, surface area = $30.5 * 1.27\pi = 38.74\pi$ cm squared. Graphite is the same material the lead of your pencil is made of. We chose graphite, because it is a good conductor of electricity and is relatively soft and therefore easy to drill into.

Holes were carefully and gently drilled into the tops of the rods. The holes were approximately 0.238 cm in diameter and 7 cm in depth. The hole in each rod was just wide enough to allow a small amount of 'breathing room' for the wire (i.e. it wasn't a tight fit). In order to not damage the rods when drilling, each rod was wrapped at the tip with electrical tape. Copper house wire about 15 cm long was inserted into each hole and sealed in with a combination of solder material and hot glue. The hot glue will act as a sealant to prevent the bacteria from eating away at the rods. The rods were divided into two pairs and each pair of copper wires was connected to twist-on wire connectors (wire nuts). A fifth wire was then added to each wire connector that connected the two pairs and a sixth wire led from one of the wire connectors to the rest of the battery circuit. It was expected that the graphite rods would prove to be more durable than the graphite fabric, because the rods are much thicker and more sturdy.

Although graphite rods provide less overall voltage and current as compared to the carbon fiber sheets (due to less surface area), it is a great conductor of heat and electricity. It also has excellent corrosion resistance as well as a high resistance to fracture. This will provide us with more consistency for an extended period of time when returning to the site to retrieve data.

2.2 Carbon Fiber

Two types of Microbial fuel cells are considered, a traditional graphite-rod based battery and a new carbon fiber based battery. Neither of these use a proton-exchange membrane and both only utilize inexpensive materials, with the intent to optimize the cost to electricity produced ratio.

The carbon fiber based battery is set up with two sheets of fiber: a long piece placed approximately one meter underground in anaerobic conditions, and a short piece placed so that it is exposed to oxygen (either above ground or in water). The long piece will act as the anode, through which electrons generated from cyanobacteria will enter the system. The short piece will act as the cathode, towards which the electricity will flow. The testing box, with the resistors, will be placed between the two sheets.

The purpose of the long carbon fiber sheet is to act as a host for the cyanobacteria. As the microbacteria undergo bioelectrogenesis, the released electrons travel through the carbon fibers, which offer the path of least resistance, towards the cathode. Since the carbon fibers are unidirectional along the length of the sheet, three additional bare copper wires were interwoven and soldered, with tin, perpendicular to the fibers. This will allow the electricity in the fibers to be directed towards a main wire which is attached to the ends of the three copper wires. This main wire will lead directly to the test box. All exposed wires that will be in the anaerobic conditions are covered with hot glue in order to prevent bacteria from using the copper as a food source.

The long sheet was placed in a 'lasagna' shape, and local mud was placed between each layer. The carbon fiber sheet acting as the cathode does not necessarily have to be of this material, but any excellent conducting material may be used. The only purpose for this sheet is to create a potential difference for the electrons to flow through the circuit, thus creating a current. Anaerobic bacteria use a final electron acceptor other than oxygen to complete their electron transport chain. For this test location, the final electron acceptor may be sulfur, as it is abundantly found around the area.

Compared to the graphite rod battery, the surface area of the fiber buried in the ground is much larger for the same cost and can thus accept more electrons from the microbacteria.

2.3 Test Box

To verify electrical production from the mud battery, there needs to be means of measuring the current and voltage. We built a switched testing circuit to simplify these measurements and enclosed it in a waterproof "test box." The test box is controlled by a simple switch to open the circuit through a pair of testing leads, or close the circuit through the battery in default use. The test box has a dial which can be turned from the off position to one of five other positions which passes the current through resistors of various strength as described in Section 2.4.

Following a circuit diagram, we constructed the inner workings of the test box by soldering to the connection endpoints of a 6-way rotary switch and soldering the contact points to secure them. The output wires from the dial lead were attached to test leads which can be probed by a multi-meter to get a reading of the electrical yield of the mud

battery. The circuit continues through slide switches and to additional wires that lead outside the box to the main battery.

2.4 Current and Voltage Measurement Setup

We wanted to be able to test the battery at different levels of resistance. Using a small circuit board from Radio Shack, some resistors ranging from 10Ω to $100,000\Omega$, a soldering iron, and some soldering material we soldered the resistors to the circuit board. Altogether we soldered 4 rows of 5 resistors each to the board. Each row had a 10, 100, 1,000, 10,000 and a 100,000 ohm resistor. You can determine the rating of a resistor by using the established color code. The body of each resistor contains a band of three or four colors that indicate the rating. In some cases there is an extra band on the body that indicates the resistor's tolerance or accuracy rated as a percentage. The tolerance is typically 1%, 2%, 5% or 10%.

According to Ohm's law the current flowing through a resistor is directly proportional to the voltage. In other words $V = IR$, where V is voltage, R is resistance and I is current. The bacteria in the mud and the oxygen in the air create a flow of electrons through the circuit. Each of the five resistors is connected to an independent circuit that can be activated with the turn of a circular switch. Since the switch has five different settings and each setting is connected to only one circuit the circular switch is called a six pole two throw switch. We can control how much current flows through the battery just by turning the switch to a different setting. For a fixed voltage V the larger the resistance the smaller the current and vice versa.

The only difficulty we encountered with constructing the resistor pack was in the soldering. We had to be very precise to make sure only the exposed wire of the resistor was soldered to the board and that the board itself did not come into contact with soldering iron. This required a lot of patience!

3. Data Analysis

In our first experiment, rich, dark anaerobic mud with a strong associated odor was used to verify the operation of the fuel cell. Open circuit voltages of 0.6 volts were measured immediately on graphite rods, and served as a baseline for comparing the main experiment.

For our main experiment, we used a grey mud, which did not have a strong odor. There were two benefits of this selection. First it allowed us to verify if measurements of life could be taken of a sample that was not as rich to begin with. Several samples were also collected for lab testing. Second, this allowed us to examine the viability of a microbial fuel cell in sub-optimal soil conditions.

For the first several days, only open circuit voltages could be measured, so only open circuit measurements will be discussed in this paper. On the morning of the first day, the voltage on the graphite rods were measured to be 55mV and

the voltage on the carbon fiber was 110 mV. As expected the carbon fiber, performed significantly (two times) better in the same soil conditions. Eight hours later, the voltage on both anodes had risen: the graphite measured 64mV and the carbon fiber measured 130mV. The voltages were checked repeatedly over the next 20 minutes and the voltage never varied more than a few mV.

The circuit was then closed to allow easy flow of electrons, and presumably the maximum growth of the microbes. The voltage on the graphite rods was measured one day later to be 117mV, no measurement of the carbon fiber was taken at this time. The carbon fiber was measured the next morning and found to be 236mV. A wire on the graphite rods had come loose in the mean time and was detected and fixed at this point, but the two systems were no longer both in identical states so exact comparison was no longer possible. While the measurements were not taken at the same time, the general trend of the carbon fiber having twice the voltage was continued.

The circuit was closed again and left alone for a month. At this point the carbon fiber had risen to 0.36V. A new soil sample was taken and the mud was both noticeably darker and had a much stronger smell.

4. Conclusions

The results are very preliminary, but several things seem likely. First, carbon fiber provided a superior performance to graphite rods, for a comparable price, suggesting it is a viable candidate for future power generation techniques. Second, even in a soil sample with likely a low concentration of microbes, an easily measurable voltage was obtained, verifying that it can be used to measure life. Third, the voltage grew simultaneous with the increase in other indicator of life, verifying that even in extremely low concentrations, a small probe could be left behind in a "grow" state (closed circuit), and returned to later to be measured for any increase. The preliminary results thus indicate that carbon fiber microbial fuel cells are a potentially viable source of power in remote, extreme environments, and that the fuel cell life detector is a reasonable candidate for detecting life on other planets.

References

- [1] Cunningham, A.: Microbial moxie. *Science News* **169**(5) (2006)
- [2] Cushing, G., Titus, T., Wynne, J., Christensen, P.: Themis observes possible cave skylights on mars. *Geophysical Research Letters* **34**(L17201) (2007)
- [3] Lovley, D.R.: Powering microbes with electricity: direct electron transfer from electrodes to microbes. *Environmental Microbiology Reports* (2010)
- [4] Navarro-González, R., Vargas, E., de la Rosa, J., Raga, A.C., McKay, C.P.: Reanalysis of the viking results suggests perchlorate and organics at midlatitudes on mars. *J. Geophys. Res.* **115**(E12010) (2010)

Emerging Viral Agents at Risk in Global Health Approaches to Early Diagnosis and Prompt Therapy

Giulio Tarro^{1,2,3,*} and Ciro Esposito⁴

¹Department of Biology, Center for Biotechnology, Sbarro Institute for Cancer Research and Molecular Medicine, Temple University, Philadelphia, PA, USA.

²Committee on Biotechnologies and VirusSphere, World Academy of Biomedical Technologies, UNESCO, Paris, France.

³Foundation T. & L. de Beaumont Bonelli for Cancer Research, Naples, Italy

⁴Unit of Virology, D. Cotugno Hospital, Naples, Italy.

*Correspondence to: Giulio Tarro, Via Posillipo 286, 80123 Naples, Italy

e-mail: gitarro@tin.it gitarro@teletu.it

Abstract - Emergence of a novel swine-origin influenza A (H1N1) virus in humans was detected in April 2009 in Mexico, Canada and USA. The swine-origin influenza virus (S-OIV) showed a unique genome composition not detected before. The S-OIV caused outbreaks of febrile respiratory infection from mild to severe disease throughout the world. This paper describes the tracking of H1N1 in Campania, the most affected Italian region and the one in which the highest number of flu-related deaths occurred. Here, we discuss the possible causes of these high incidence and mortality rates and their implications on the public opinion and the prevention campaign.

Key words: influenza A; H1N1, swine-origin influenza virus (S-OIV); flu virus prevention

1 Introduction.

The history of flu viruses teaches that influence originates from birds, usually aquatic, then it is transferred to man through the leap into pigs. The promiscuity of the herds, facilitates this transition and then the spread. Three pandemics caused by influenza A viruses, which occurred in the 20th century, have all had this origin: the 'Spanish flu' (1918, H1N1), the 'Asian flu' (1957, H2N2) and the 'Hong Kong flu' (1968, H3N2).

The 2009 H1N1 influenza virus acted during the winter in Australia and New Zealand yielding a pattern effect for the treatment of patients during the winter in the Northern Hemisphere [1].

The performance of rapid diagnostic test for the detection of novel influenza A (H1N1) virus was evaluated by the Centers for Disease Control and Prevention [2].

The findings of severe respiratory disease concurrent with the circulation of H1N1 influenza was proved by the aforementioned test [3].

Even the potential impact of pandemic influenza during the Hajj pilgrimage was taken in account to reduce the substantial effect on the crowd to spread the infection [4], [5].

Previous observation of hospitalized patients with the 2009 H1N1 influenza in the USA during springtime indicated how to cope with patients showing severe medical conditions [6]. Pregnant women were at increased risk for complication from pandemic H1N1 virus infection [7]. Same critical illness was reported in children and suggested planning responses in intensive care units with swine derived H1N1 virus [8].

The story of the 2009 H1N1 influenza has been tracked since the beginning [9] and the experience of the previous pandemics was the key to afford on time screening procedures and to promote specific vaccine programs all over the world [10]. The current concepts on the emergence of influenza A viruses are reported in many review articles [11], [12].

Persons who were born before 1957 had a reduced risk of infection [13]. Furthermore cross-reactive antibody responses were measured in people vaccinated with 1976 swine influenza vaccines [14]. Therefore, a good portion of older adults had pre-existing cross reaction antibodies to the 2009 pandemic H1N1 influenza virus [15]. This is similar to what happened with the recent strains of 1957 Asian flu (A2) for which it was demonstrated the presence of antibodies in older segment of that population. In Asian influence there were obviously strains with dominant characters, other than those that had characterized the previous years, but similar to those of the strains widespread much before (1889-90 pandemic). This is consistent with Burnet theory on the origin of new epidemic strains. Most old people have antibodies directed towards the antigens from the strains with which they were in contact. As age progresses the immunity spectrum broaden reflecting the ample repertoire of polyvalent antibodies generated following the contact with many primary and secondary antigens present in viral strains encountered during the years. But each contact with a flu virus of type A involves not only specific antibodies, but also an increase in those directed towards the strain responsible for the first flu infection of the subject (phenomenon of Davenport or doctrine of original sin). In this way, the immunization to a particular strain increases in a certain period, limiting further distribution of the virus and creating a selective advantage, for some viral variants, to multiply and spread. The new strains will be in

conditions of growing in hosts, regardless of whether they have or not an immunologic experience with the previous strains. As a result, shortly after the appearance of a new type, the old forms will disappear and the new family will become dominant for a period that usually covers 10-20 years, in which there is, for the emergence of minor antigenic variation, the subdivision in various subtypes.

The outcome of a new epidemic strain may, therefore, be regarded as a developmental process involving both the characteristics of the strain and the susceptibility of the population. For a viral strain to reach a wide distribution, its antigenic characteristics must ensure that it escapes the neutralization of the host antibodies and of the surrounding population. So the outbreaks will happen with those strains that have dominant antigens that fit the deficiency, or better, the absence of the antibody in the population. It seems, in conclusion, that the flu virus shows an ability and an aptitude for survival owing to the emergence of new models that allow the virus to affect populations still partly immune to previous antigenic forms. According to this view, the changes in the influenza A can be designed in a single meaning, in the context of a principle and of an evolutionary progress, from Burnet said immunological drift.

Here, we report the data on the H1N1 influenza in the Italian region Campania, which resulted the most affected by the S-OIV and the one with more lethal cases. We discuss the possible causes of these high incidence and mortality rates as well as their implications on the public opinion and the prevention campaign.

2 Results.

Among the Italian regions that were most affected by the S-OIV, Campania was leading for incidence of the infection, with a rising number of flu-related fatalities in its main town, Naples. This obviously generated some panic among the population. The Virology Laboratory of Cotugno Hospital in Naples is the sole center for the surveillance on the virus approved by the Italian Ministry of Health for Campania region. This allows us to make a wide comparison of cases, helping to correlate all the different diagnosis. 5706 diagnostic tests were performed at the Virology Laboratory of Cotugno Hospital in Naples starting on April 28, 2009 (3 days after the WHO alert) until December 31, 2009. The method used for the detection of S-OIV was a real-time reverse-transcriptase-polymerase chain reaction assay [2]. Of these 5706 tests, 40,80% (2329) resulted to be positive. In May, 2 out of 25 tests were found to be positive for the H1N1 virus and corresponded to the first two positive patients in Campania. Only few tests were performed in April, May and June (3, 25 and 11 respectively). Whereas during and after the summer the number of tests performed increased, peaking in November: 222 in July (48,64% positive), 127 in August (52,76% positive), 396 in September (30,30% positive), 999 in October (53,65% positive), 3103 in November (45,47% positive), 820 in December (10,36% positive). In Campania the peak of

influenza occurred during the 44th week of 2009 and preceded of about two weeks the incidence peak at national level.

Of the 2329 patients who were positive for H1N1 infection, 1284 (55,10%) were males and 1045 (44,90%) females; similar percentages were found for negative patients (56,40% males versus 43,60% females) suggesting that gender does not seem to affect the incidence rates.

Most patients who reached the Virology Laboratory of Cotugno Hospital were from the main town Naples (4290 patients, 77,0% of which were positive). Whereas 1416 were from the other Campania provinces Salerno (824 patients, 12,62% of which were positive), Caserta (382 patients, 6,35% of which were positive), Avellino (161 patients, 3,13% of which were positive), Benevento (49 patients, 0,90% of which were positive). The number of tests performed reflects the number of inhabitants belonging to each Campania province. In fact, according to data from the National Institute of Statistics (<http://demo.istat.it/>) in 2009 Naples was the most populated province followed by Salerno, Caserta, Avellino and Benevento. However, the percentage of patients who resulted positive for H1N1 infection was much higher in Naples compared to the other Campania provinces. This is probably due to the higher population density in the main town Naples, which favors the infection spreading.

In Campania the age group from 7 months to 10 years, including 634 patients, showed the highest percentage of incidence for H1N1 infection (28,85%), (table 1). This is consistent to what observed at the National level, in fact, in Italy the age group from 0 to 14 years resulted the most affected, as reported by INFLUNET the surveillance network for influenza coordinated by the Italian Ministry of Health.

	A (0-6 months)	B1 (7 months 10 years)	B2 (11-17 years)	C1 (18-27 years)	C2 (28-35 years)	C3 (36-49 years)	D (over 50 years)
2197 TOTAL	51	634	321	406	200	292	293
% on total positive	2,36%	28,85%	14,61%	18,47%	9,10%	13,29%	13,33%
F-%F on total for group of age	21- 41,2%	267- 42,10%	154- 48,00%	180- 44,30%	92- 46,00%	131- 44,85%	131- 44,70%
M-%M on total for group of age	30- 58,8%	367- 57,90%	167- 52,00%	226- 55,70%	108- 54,00%	161- 55,15%	162- 55,30%

Table 1. Positive Patients Classified According to the Age Group.

3 Conclusions.

Although the number of victims caused by H1N1 influenza is decidedly inferior to other pandemics [6], [15] a potential risk of a panic syndrome existed because of a bad information or a scarce knowledge of the phenomenon. . The virus, that was

first detected in Mexico, reached other parts of the world as happens for all the types of influenza virus [14]. While for the SARS a direct contact was necessary, through the so-called droplets of Pflugge, this swine-derived influenza spreads to more distance through the air and is very contagious. .

Among the Italian regions that were most affected by the S-OIV, Campania was leading for incidence of the infection and flu-related fatalities. This can be in part due to the fact that Campania is the most densely populated Italian region, which obviously favors the spreading of the infection. In fact, also at the regional level, the main town Naples, which is the most densely populated Campania province, had the higher percentage of flu incidence compared with the other provinces.

Consistent with the incidence data relative to the whole Italian population the most affected group resulted to be the younger population. This is probably owed to both the higher population density due to the scholarization and to the more promiscuous behaviors.

As for the high mortality rates, in our opinion, the data about Campania may appear to be higher than other areas in Italy because Cotugno hospital in Campania has medical specialists qualified to make detailed diagnoses to determine if H1N1 is the main or a co-factor in mortality cases. Similar centers, which are able to compare cases to see if H1N1 is the main or a co-factor of mortality are not so readily available throughout other regions. This means that data from other regions (especially from the South) might not be as accurate as Campania's data. It's also reasonable to suppose that given a lack of capability in many cases to determine a precise diagnosis for the H1N1 virus, it may not always be possible to know when the virus is a main factor or a co-factor in the mortality of a patient [16].

Also, for what it may concern the increased deaths observed in Naples and/or in Campania during the peak of the novel influenza A (H1N1), we can remember the Will Rogers phenomenon because the element being moved (S-OIV infected samples) to the Virology laboratories of the Cotugno Hospital was above the current average of the set it was entering. By definition, adding it to the new set will raise the incidence of H1N1 virus infection and then the mortality average.

The analysis of the factors that contribute to higher flu incidence is important not only to address the panic issues among the population but has also implications on the prevention campaign.

The massive campaign for vaccination across Italy helped to stop the spread of the virus, which while not very aggressive, is very contagious.

From the first symptoms through convalescence, an episode of H1N1 flu lasts about 10 days. The epidemic itself, however, could possibly last for months, since several human variants of the flu may merge with H1N1 to create a new and possibly more dangerous and harmful viral variant [11]; [12].

The vaccination against the influenza is the most effective method to prevent the illness. From the moment of the isolation of a new flu virus, one must wait for the preparation

of a new specific vaccine to be ready for the next influenza season in Autumn [17].

The vaccine against the virus prevents the flu in 70-80% of cases. It takes about two to three weeks after the injection to develop antibodies for the virus [18].

Vaccines are free and can be administered by family doctors or pediatricians for children. It is recommended, but not obligatory, for children between 6 months and 2 years of age. The Ministry of Health also provides vaccinations to all hospital-based doctors and medics, blood donors and chronically ill patients up to age 65. The last group of patients who will be vaccinated include healthy people between 6 months and 27 years. The prototype vaccine did not cause any particular collateral damages [19] and only a single dose is necessary for protection [20].

The authors declare no conflict of interests.

4 References.

1. SA Webb , V Pettila, I Seppelt, et al. Critical care services and 2009 H1N1 influenza in Australia and New Zealand. *N Engl J Med* 361:1925-34, 2009.
2. Evaluation of rapid influenza diagnostic tests for detection of novel influenza A (H1N1) Virus - United States, 2009. *MMWR Morb Mortal Wkly Rep* 58:826-9, 2009.
3. G Chowell, SM Bertozzi, MA Colchero, et al. Severe respiratory disease concurrent with the circulation of H1N1 influenza. *N Engl J Med* 361:674-9, 2009.
4. SH Ebrahim, ZA Memish, TM Uyeki, TA Khoja, N Marano and McNabb SJ. Public health. Pandemic H1N1 and the 2009 Hajj. *Science* 326:938-40, 2009.
5. ZA Memish, SJ McNabb, F Mahoney, et al. Establishment of public health security in Saudi Arabia for the 2009 Hajj in response to pandemic influenza A H1N1. *Lancet* 374:1786-91, 2009.
6. S Jain, L Kamimoto, AM Bramley, et al. Hospitalized patients with 2009 H1N1 influenza in the United States, April-June 2009. *N Engl J Med* 361:1935-44, 2009.
7. DJ Jamieson, MA Honein, SA Rasmussen, et al. H1N1 2009 influenza virus infection during pregnancy in the USA. *Lancet* 374:451-8, 2009.
8. P Lister, F Reynolds, R Parslow, et al. Swine-origin influenza virus H1N1, seasonal influenza virus, and critical illness in children. *Lancet* 374:605-7, 2009.
9. FS Dawood, S Jain, L Finelli, et al. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med* 360:2605-15, 2009.
10. RP Wenzel, MB Edmond. Preparing for 2009 H1N1 Influenza. *N Engl J Med* 361:1991-3, 2009.
11. SM Zimmer, DS Burke. Historical perspective--Emergence of influenza A (H1N1) viruses. *N Engl J Med* 361:279-85, 2009.
12. DM Morens, JK Taubenberger and AS Fauci. The persistent legacy of the 1918 influenza virus. *N Engl J Med* 361:225-9, 2009.

13. DN Fisman, R Savage, J Gubbay, et al. Older age and a reduced likelihood of 2009 H1N1 virus infection. *N Engl J Med* 361:2000-1, 2009.
14. C Soares. Pandemic payoff. *Sci Am* 301:19-20, 2009.
15. K Hancock, V Veguilla, X Lu, et al. Cross-reactive antibody responses to the 2009 pandemic H1N1 influenza virus. *N Engl J Med* 361:1945-52, 2009.
16. Virological surveillance of human cases of influenza A(H1N1)v virus in Italy: preliminary results. *Euro Surveill* 14, 2009.
17. M Enserink, J Cohen. Virus of the year. The novel H1N1 influenza. *Science* 326:1607, 2009.
18. XF Liang, HQ Wang, JZ Wang, et al. Safety and immunogenicity of 2009 pandemic influenza A H1N1 vaccines in China: a multicentre, double-blind, randomised, placebo-controlled trial. *Lancet*; 375:56-66, 2010.
19. Z Vajo, F Tamas, L Sinka and I Jankovics. Safety and immunogenicity of a 2009 pandemic influenza A H1N1 vaccine when administered alone or simultaneously with the seasonal influenza vaccine for the 2009-10 influenza season: a multicentre, randomised controlled trial. *Lancet*; 375:49-55, 2010.
20. E Plennevaux, E Sheldon, M Blatter, MK Reeves-Hoche and M Denis. Immune response after a single vaccination against 2009 influenza A H1N1 in USA: a preliminary report of two randomised controlled phase 2 trials. *Lancet*; 375:41-8, 2010.

Applications of Artificial Immune Systems

Suhair H. Amer, Ph.D.

Department of Computer Science, Southeast Missouri State University, Cape Girardeau, MO, USA

Abstract—*In recent decades, Artificial Immune Systems (AISs) have appeared as an approach dealing with security systems and classification problems. This paper briefly surveys AIS basic concepts, features and principles, approaches and applications.*

Keywords: Artificial Immune Systems, applications of AIS, biologically inspired systems

1 Introduction

An Artificial Immune System (AIS) is a biologically inspired computing which is currently investigated to solve many problems. Such a method was inspired by the Human Immune System (HIS) that can detect and defend against harmful and previously unseen invaders. AISs have been built for a wide range of application domains including document classification, fraud detection, and network- and host-based intrusion detection. Section 2 will discuss the basic concepts used to build an AIS. Section 3 will discuss AISs features and principles that distinguish them from other methods. Section 4 discusses the philosophies or approaches of applying immune system concepts. Section 5 discusses examples of applications of AISs. Finally, section 6 concludes the paper.

2 Artificial Immune systems Basic Concepts

To implement a basic AIS, four decisions have to be made: encoding, similarity measure, selection and mutation.

2.1 Initialization and Encoding

It is very important to choose a suitable encoding [1] to insure the algorithm's success. The antigen and antibody should be defined in the context of an application domain. For example, antigens can represent intrusion data instances, and antibodies bind to antigens identifying an intrusion.

2.2 Similarity or Affinity Measure

A good matching algorithm guarantees that the AIS works properly. The primary response in the immune system [2] uses learning mechanism to identify antigens that

have not been detected by a detector before. When a B cell is activated after binding to a pathogen, it starts cloning itself and the cloned cells then undergoes a somatic hyper mutation to create child B cells with mutated receptors. Then all B cells compete with each other.

2.3 Negative Selection

In the negative selection algorithm [1], a set of trusted behavior describing self is defined. During the initialization of the algorithm, a large number of detectors are created. Then these detectors are subjected to a matching algorithm that compares them to self behavior. Any matching detector will be eliminated and those that do not match are selected which explains the term negative selection.

2.4 Somatic Hyper mutation

Somatic hyper mutation [1] is an optional process associated with negative selection. Rather than ignoring matching detectors in the first phase of the algorithm, they can be mutated to save time and effort. Also, depending on the degree of matching, the mutation could be more or less strong.

2.5 Cross-Reactivity and Associate Memories

When a B-cell encounters subsequent antigens it responds quicker (secondary response) in which the memory cells for an earlier antigen quickly start producing large quantities of a specific antibody. In general, B-cell receptors do not require an exact match to an antigen to be activated. Therefore, some memory cells can react to new antigens producing a secondary response which is termed, the cross-reactive memory [3].

3 AIS Features and Principles

In general AISs have the following desirable features and principles [4]:

- **Distributed:** the presence of an infection is determined locally with no central coordination taking place.
- **Scalability:** communication and interaction between components are localized and there is little

overhead associated when the number of components is increased.

- **Multi-layered:** security is achieved by combining multiple layers of different mechanisms to provide high overall security.
- **Diversity:** it is less likely that the security vulnerabilities in one system be widespread.
- **Robustness:** No single component or cell of the human immune system is essential and can be replaced.
- **Autonomy:** no outside management or maintenance is required as classification and elimination of pathogens and repairing self is done locally.
- **Adaptability:** The immune system learns to detect new pathogens, and retains the ability to recognize previously seen pathogens through immune memory.
- **No secure layer:** Any cell in the human body can be attacked by a pathogen or even another immune system cell.
- **Dynamically changing coverage:** since the immune system cannot maintain a set of detectors large enough to cover the space of all pathogens, it maintains a random sample of its detector repertoire circulating throughout the body.
- **Identity via behavior:** identity is verified through the presentation of peptides, or protein fragments.
- **Anomaly detection:** ability to detect pathogens that has never been encountered before.
- **Flexibility or Imperfect detection:** By accepting imperfect detection, the immune system increases the flexibility with which it can allocate resources.
- **Detector replication:** The human immune system replicates detectors to deal with replicating pathogens.
- **Memory :** the immune system reacts more rapidly the second time against pathogens that are similar to the ones that were encountered previously.
- **Implicit policy specification:** definition of self in the immune system is empirically defined by monitoring proteins that are currently in the body.

4 Immune System Approaches

Application of immune system concepts can be based on the following distinct philosophies [5]:

4.1 Negative Selection (NS)

Negative selection concepts are concerned with eliminating immature cells that bind to self antigens. This allows the HIS to detect non-self antigens without mistakenly detecting self-antigens.

4.2 Danger Theory

The Danger Theory describes which data should be represented. It focuses on the presence of dangerous signals and goes beyond and overcomes many of the limitations of self–non-self selection [1].

4.3 Immune Network Theory

The hypothesis of the immune network theory states that the immune system maintains an idiotypic network of interconnected B-cells for antigen recognition. These cells both stimulate and suppress each other in certain ways that lead to the stabilization of the network. For example, two B-cells connect if their shared affinities exceed a certain threshold, and the strength of the connection is directly proportional to the affinity they share [1].

4.4 Clonal Selection Principle

Clonal Selection Principle [1] describes the basic features of an immune response to an antigenic stimulus. Only the cells that recognize the antigen proliferate and are selected against those that do not.

4.5 Idiotypic Networks

The Idiotypic network hypothesis [1] builds on the recognition that antibodies can match other antibodies as well as antigens. This could be used to explain how the memory of past infections is maintained and could result in the suppression of similar antibodies. In general, the nature of an Idiotypic interaction can be either positive or negative.

4.6 Other methods

Although negative selection and the danger theory are the most popular approaches in AIS for intrusion detection, some researchers choose to create AIS based on alternative ideas. For example, Forrest et. al [6] build an intrusion detection system (IDS) based on an explicit notion of self within a computer system. The system was host-based, examining specifically privileged processes. The system collected self-information to construct a database of normal commands.

5 Artificial Immune systems Applications

This section briefly introduces some application areas where AIS have been applied.

5.1 Recommender Systems

Collaborative filtering (CF) [7][8] is one of the common applications of AIS. CF is the term for a broad range of algorithms that use similarity measures to obtain recommendations. In general, any problem domain where users are required to rate items is amenable to CF techniques. For example, commercial applications are called recommender such as movie recommendation. Traditionally, recommended items are treated as black boxes and recommendations are based purely on the votes of neighbors, and not on the content of the item. A user profile which consists of the preferences of a user that is usually a set of the user's votes on an item. These profiles are then compared to build a neighborhood.

Morrison and Aickelin [9] applied idiotypic network theory to build their web site recommender AIS based system. The idiotypic network theory states that interaction in the immune system do not only occur between antibodies and antigens but also between antibodies and each other. Therefore, the antibody may be matched by other antibodies. This activation can spread throughout the population. In general, the interaction may have a positive or a negative effect on a particular antibody-producing cell. Morrison and Aickelin idea is that antibodies that are very similar to each other had their concentrations reduced. This allowed the creation of a set of users that are similar to a user but quite still different to each other which enhances the recommendation accuracy of the system.

Hsieh et. al [10] employed AIS to deal with classification problem. In their paper, an AIS algorithm is developed and applied to a two-group classification problem. They discuss a Taiwanese banking industry example and the financial ratios of each bank from 1998 to 2002 were collected. Their system had a 10% better performance than the three soft computing early warning systems (GNN, CBR and BPN). Their AIS outperforms the statistical early warning systems (LR and QDA) at least 24%.

Singh and Nair [11] outline a robot controller based on a combination of the innate and adaptive immune systems. The learner robot learns to accurately follow a track. It can sense when it is on the track and when it loses it. If it loses the track, it first tries to find it on its own and then requests the assistance of a helper robot, who will guide it back to the track. The general idea is to have the learner robot learn to navigate weak portions of the track autonomously, without

losing the track and having to be guided back by the helper. The proposed immune system has two type of response governed by separate innate and adaptive subsystems.

Burgess has developed Cfengine [12], an autonomous agent and a middle-to-high level policy language for building expert systems to administrate and configure large computer networks. The system adapts the danger model using autonomous and distributed feedback and healing mechanism triggered when a small amount of damage is detected. Cfengine automatically configures large numbers of systems on a heterogeneous network with an arbitrary degree of variety in the configuration.

5.2 Security based systems

Security systems may include virus detection and intrusion detection systems. Virus detection is viewed as a self-non-self discrimination problem. Targets such as legal user activities, legal application usage activities, and uncorrupted data are monitored as self and the AIS are expected to discriminate them from illegal user activities, illegal application usage activities, and virus infected data.

The Computer Virus Immune System (CVIS) approach [13] is able to perform virus analysis, repair infected files and propagate the analysis results to other local systems. In addition, CVIS was designed to operate under a distributed environment using autonomous agents. The system was tested against the TIMID virus, which infects .com files within a local directory. The test reports showed the sensitivity of detection and error results on different matching thresholds. It showed a detection rate of up to 89% but had a very high scalability problem since it required approximately 1.05 years for generated antibodies to scan an 8GB hard disk drive. However, novel concepts such as life span, activation threshold and co-stimulation were investigated.

Sarafijanovic and Le Boudec [14] built an immune-based system to detect misbehaving nodes in a mobile ad-hoc network. The authors considered a node to be functioning correctly if it adhered to the rules laid down by the Dynamic Source Routing (DSR) protocol. Each node in the network monitored its neighboring nodes and collected one trace per monitored neighbor. Four events were sampled over fixed and discrete time intervals to create a series of data sets. This created a binary antigenic representation.

Stillerman et. al [15] introduced an immunity-based intrusion detection approach that was particularly applicable to Common Object Request Broker Architecture (CORBA) applications. CORBA is a popular common messaging middle-ware that enables the communication of distributed objects for distributed applications. The authors employed the same approach reported in [16] to detect a misuse

attacks performed by a legal user of the system. The results showed that the system was able to detect anomalies without high false positive error rates.

Dasgupta [17] provided the conceptual view and a general framework of a multi-agent anomaly based intrusion detection system and response in networked computers. The immunity based agents in the system roamed around nodes and monitored network situation. Each agent can recognize other activities and can take appropriate actions according to its predefined security policies. The agent can adapt to its environment dynamically and can detect novel and known attacks. Network activities were monitored on the user, system, process and packet levels.

Pagnoni and Visconti's [18] NAIS IDS is inspired by innate immune mechanisms. Their immune system is a multilayer defense system. The innate immune system is the first line of defense which is able to recognize self quickly. Their system compiles a list of all observed process names during a training period containing only normal usage. A set of "digital macrophages" is then created which monitors the system and generates an alert when any previously unseen process name is observed.

6 Conclusion

The human immune system has been successful in defending different human organs against a wide range of harmful attacks. Negative selection and Danger Theory are two of the commonly used philosophies in building artificially immune based systems. In general, to implement a basic artificial immune system, four decisions have to be made: encoding, similarity measure, selection and mutation. This paper briefly discussed the features and principles that make AIS desirable for building applications dealing with security and recommender systems.

7 References

- [1] U. Aickelin and D. Dasgupta. Artificial Immune Systems Tutorial. To appear in *Introductory Tutorials in Optimization, Decision Support and Search Methodology* (eds. E. Burke and G. Kendall), Kluwer, 2005.
- [2] Stephanie Forrest and Steven A. Hofmeyr. Immunology as information processing, In *Design Principles for the Immune System and Other Distributed Autonomous Systems*, edited by L.A. Segel and I. Cohen. Santa Fe Institute Studies in the Sciences of Complexity. New York: Oxford University Press. (2001).
- [3] S. Forrest, S. Hofmeyr. Engineering an Immune System. *Graft*, Vol. 4, No. 5, 2001, 5-9
- [4] A. Somayaji. Operating System Stability and Security through Process Homeostasis. PhD thesis, University Of New Mexico, 2002.
- [5] J. Kim, P. Bentley, U. Aickelin, J. Greensmith, G. Tedesco and J. Twycross. Immune System Approaches to Intrusion Detection - A Review. *Natural Computing*, Springer, in print, pp XXX. 2007.
- [6] S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff. A sense of self for Unix processes. In *Proceedings of 1996 IEEE Symposium on Computer Security and Privacy*, pp. 120-128 (1996).
- [7] S. Cayzer and U. Aickelin. A Recommender System based on the Immune Network. *Proceedings CEC*, pp 807-813. 2002.
- [8] D. L. Chao and S. Forrest. Information Immune Systems. In *Proceedings of the First International Conference on Artificial Immune Systems (ICARIS)*, pp. 132-140 2002.
- [9] T. Morriso and U. Aickelin. An AIS as a Recommender System for Web Sites. *1st International Conference on AIS*, pp 161-169. 2002.
- [10] Jih-Chang Hsieh, Shih-Hsin Chen and Pei-Chann Chang. Application of Artificial Immune System in Constructing a Financial Early Warning System: An Example of Taiwanese Banking Industry. *Proceeding ICICIC '07. Proceedings of the Second International Conference on Innovative Computing, Information and Control*. IEEE Computer Society Washington, DC 2007.
- [11] C. T. Singh and S. B. Nair. An Artificial Immune System for a Multi Agent Robotics System. In *Proc. of the 4th World Enformatika International Conference on Automation Robotics and Autonomous Systems (ARAS 2005)*, pages 308–311, 2005.
- [12] M. Burgess. Evaluating cfegine's immunity model of site maintenance. In *Proceeding of the 2nd SANE System Administration Conference (USENIX/NLUUG)*, 2000.
- [13] P. K. Harmer, P. D. Williams, G. H. Gunsch, and G. B. Lamont. An artificial immune system architecture for computer security applications. *IEEE Transactions on Evolutionary Computation*, 6(3):252-280, June 2002.
- [14] S. Sarafijanovic and J.-Y. Le Boudec. An Artificial Immune System Approach with Secondary Response for Misbehavior Detection in Mobile Ad-Hoc

Networks. *IEEE Transactions on Neural Networks, Special Issue on Adaptive Learning Systems in Communication Networks*, 16(5):1076–1087, 2005.

- [15] M. Stillerman, C. Marceau, and M. Stillman. Intrusion detection for distributed application. *Communications of the ACM*, 42(7):62-69, July 1999.
- [16] S. A. Hofmeyr, S. Forrest and A. Somayaji. Intrusion detection using sequences of system calls. *Journal of Computer Security*, 6 (1998), 151--180.
- [17] D. Dasgupta. Immunity-based intrusion detection systems: A general framework. In *Proc. of the 22nd National Information Systems Security Conference (NISSC)*, October 1999.
- [18] A. Pagnoni and A. Visconti. An innate immune system for the protection of computer networks. In *Proc. of the 4th International Symposium on Information and Communication Technologies*, pages 63–68. Trinity College Dublin, 2005.

On a Body Sway Model while Maintaining Upright Posture during Exposure to a Stereoscopic Movie on a Liquid Crystal Display

Hiroki Takada[†], Kazuhiro Fujikake, and Masaru Miyao, *Member, IEEE*[†]

Abstract— It is known that a mathematical model of the body sway can be developed by a stochastic process. The authors have succeeded in finding the nonlinearity in the potential function. Regarding to the mathematical model, we applied an index, sparse density (SPD), of stationary stabilograms for detecting instability due to the motion sickness (simulator sickness), which occurs when a human attempts to maintain an upright posture. In this study, subjects in a standing position were stimulated by stereoscopic movies on a liquid crystal display (LCD). We also measured the degree of determinism in the dynamics of the sway of the center of gravity of the subjects. The Double-Wayland algorithm was used as a novel method. As a result, the dynamics of the body sway in the presence of the stimulus as well as in its absence were considered to be stochastic. The structural changes in the potential function during exposure to the conventional three-dimensional images could be detected by using the SPD.

I. INTRODUCTION

The human standing posture is maintained by the body's balance function, which is an involuntary physiological adjustment mechanism called the righting reflex [1]. In order to maintain the standing posture when locomotion is absent, the righting reflex, centered in the nucleus ruber, is essential. Sensory signals such as visual inputs, auditory and vestibular inputs, and proprioceptive inputs from the skin, muscles, and joints are the inputs that are involved in the body's balance function [2]. The evaluation of this function is indispensable for diagnosing equilibrium disturbances such as cerebellar degenerations, basal ganglia disorders, or Parkinson's disease in patients [3].

Stabilometry has been employed to evaluate this equilibrium function both qualitatively and quantitatively. A projection of a subject's center of gravity onto a detection stand is measured as an average of the center of pressure (COP) of both feet. The COP is traced for each time step, and the time series of the projections is traced on an x-y plane. By connecting the temporally vicinal points, a stabilogram is created, as shown in Fig 1. Several parameters such as the

area of sway (A), total locus length (L), and locus length per unit area (L/A) have been proposed to quantize the instability involved in the standing posture, and such parameters are widely used in clinical studies. It has been revealed that the last parameter particularly depends on the fine variations involved in posture control [1]. This index is then regarded as a gauge for evaluating the function of proprioceptive control of standing in human beings. However, it is difficult to clinically diagnose disorders of the balance function and to identify the decline in equilibrium function by utilizing the abovementioned indices and measuring patterns in the stabilogram. Large interindividual differences might make it difficult to understand the results of such a comparison.

Mathematically, the sway in the COP is described by a stochastic process [4]–[6]. We examined the adequacy of using a stochastic differential equation and investigated the most adequate equation for our research. $G(\mathbf{x})$, the distribution of the observed point \mathbf{x} , is related in the following manner to $U(\mathbf{x})$, the (time-averaged) potential function, in the stochastic differential equation (SDE), which has been considered as a mathematical model of the sway:

$$U(\bar{\mathbf{x}}) = -\frac{1}{2} \ln G(\bar{\mathbf{x}}) + \text{const.} \quad (1)$$

Actually, $G(\mathbf{x})$ is estimated by the histogram of the time series data. The nonlinear property of SDEs is important [7]. There were several minimal points of the potential. In the vicinity of these points, local stable movement with a high-frequency component can be generated as a numerical solution to the SDE. We can therefore expect a high density of observed COP in this area on the stabilogram.

The anterior-posterior direction y was considered to be independent of the mediolateral direction x [8]. Stochastic differential equations (SDEs) on the Euclid space $\mathbf{E}^2 \ni (x, y)$

$$\begin{aligned} \frac{\partial x}{\partial t} &= -\frac{\partial}{\partial x} U_x(x) + w_x(t) \\ \frac{\partial y}{\partial t} &= -\frac{\partial}{\partial y} U_y(y) + w_y(t) \end{aligned}$$

have been proposed as mathematical models that generate the stabilograms [4]–[7]. In numerical analysis, pseudorandom numbers were generated as white noise terms $w_x(t)$ and $w_y(t)$. Constructing the nonlinear SDEs from the stabilograms (Fig. 1) in accordance with Eq. (1), their temporally averaged potential functions U_x , U_y have plural minimal points, and fluctuations could be observed in the neighborhood of the

This work was supported in part by Grant-in-Aid for Scientific Research of Japanese Ministry of Education, Science, Sports and Culture (No. 23790658).

Takada Hiroki is with Graduate School of Engineering, University of Fukui, 3-9-1 Bunkyo, Fukui City, Fukui JAPAN (corresponding author to provide phone: +81-776-27-8795; fax: +81-776-27-8955; e-mail: takada@u-fukui.ac.jp).

K. Fujikake was with Institute for Science of Labour, 2-8-14 Sugao, Miyamae-ku, Kawasaki 216-8501, Japan (e-mail: fujikake@nagoya-u.jp).

M. Miyao is with Graduate School of Information Science, Nagoya University, Chikusa-ku, Nagoya 464-8601, Japan (e-mail: miyao@itc.nagoya-u.ac.jp).

minimal points [7]. The variance in the stabilogram depends on the form of the potential function in the SDE; therefore, sparse density (SPD) is regarded as an index for its measurement.

The analysis of stabilograms is useful not only for medical diagnosis but also for achieving the control of upright standing for two-legged robots and for preventing falls in elderly people [9]. Recent studies suggest that maintaining postural stability is one of the major goals of animals, [10] and that they experience sickness symptoms in circumstances where they have not acquired strategies to maintain their balance [11]. Riccio and Stoffregen argued that motion sickness is not caused by sensory conflict, but by postural instability, although the most widely known theory of motion sickness is based on the concept of sensory conflict [11]–[13]. Stoffregen and Smart (1999) report that the onset of motion sickness may be preceded by significant increases in postural sway [14].

The equilibrium function in humans deteriorates when viewing 3-dimensional (3D) movies [15]. It has been considered that this visually induced motion sickness (VIMS) is caused by the disagreement between vergence and visual accommodation while viewing 3D images [16]. Thus, stereoscopic images have been devised to reduce this disagreement [17]–[18].

VIMS can be measured by psychological and physiological methods, and the simulator sickness questionnaire (SSQ) is a well-known psychological method for measuring the extent of motion sickness [19]. The SSQ is used herein for verifying the occurrence of VIMS. The following parameters of autonomic nervous activity are appropriate for the physiological method: heart rate variability, blood pressure, electrogastrography, and galvanic skin reaction [20]–[22]. It has been reported that a wide stance (with midlines of the heels 17 or 30 cm apart) significantly increases the total locus length in the stabilograms of individuals with high SSQ scores, while the length in those of individuals with low scores is less affected by such a stance [23]. We wondered if noise terms vanished from the mathematical model (SDEs) of the body sway. Using our Double-Wayland algorithm [24], we evaluate the degree of visible determinism for the dynamics of the sway.

We propose a methodology to measure the effect of 3D images on the equilibrium function. We assume that the high density of observed COP decreases during exposure to stereoscopic images [15]. The SPD would be a useful index in stabilometry to measure VIMS. In this study, we verify that reduction in body sway can be evaluated using the SPD during exposure to a new 3D movie on an LCD.

II. MATERIAL AND METHOD

A. Participants

Ten healthy subjects (age, 23.6 ± 2.2 years) voluntarily participated in the study. We ensured that the body sway was not affected by environmental conditions. Using an air conditioner, we adjusted the temperature to 25°C in the exercise room, which was kept dark.

B. Material

The subjects stood without moving on a detection stand of a stabilometer (G5500; Anima Co. Ltd.) with their feet together. The subjects were positioned facing an LCD monitor (S1911- SABK, NANA O Co., Ltd.) on which three kinds of images were presented in no particular order: (I) visual target (circle) whose diameter was 3 cm; (II) a new 3D movie that shows a sphere approaching and going away from subjects irregularly; and (III) a conventional 3D movie that shows the same sphere motion as in (II) which was created using the Olympus power 3D method [25]. The new stereoscopic images (II) were constructed by Olympus Power 3D method. The distance between the wall and the subjects was 57 cm.

C. Design

The subjects stood on the detection stand in the Romberg posture for 1 min before the sway was recorded. Each sway of the COP was then recorded at a sampling frequency of 20 Hz during the measurement; subjects were instructed to maintain the Romberg posture for the first 60 s and a wide stance (with the midlines of heels 20 cm apart) for the next 60 s. The subjects viewed one of the images, i.e., (I), (II), or (III), on the LCD from the beginning till the end. The SSQ was filled before and after stabilometry.

D. Calculation Procedure

We calculated several indices that are commonly used in the clinical field [26] for stabilograms, such as “area of sway,” “total locus length,” and “total locus length per unit area.” In addition, new quantification indices that were termed “SPD,” “total locus length of chain” [27] and the translation error [28] were also estimated. The translation error (E_{trans}) is calculated in order to evaluate the degree of determinism for dynamics that generate a time series. E_{trans} represents the smoothness of flow in an attractor, which is assumed to generate the time series data.

III. RESULTS

The results of the SSQ are shown in Table 1 and include the scores on nausea (N), oculomotor discomfort (OD), disorientation (D) subscale and total score (TS) of the SSQ. No statistical differences were seen in these scores among images presented to subjects. However, increases were seen in the scores for N and D after exposure to the conventional 3D movie, (II) *Cross-point 3D*. In addition, the scores after exposure to the new 3D images were not very different from those after exposure to the static one, (I) *Pre*. Although there were large individual differences, sickness symptoms seemed

Table 1 Subscales of the SSQ after exposure to 3D movies

Movies	(II) <i>Cross-Point 3D</i>	(III) <i>Power 3D</i>
N	8.6 ± 2.6	14.3 ± 4.8
OD	17.4 ± 3.4	16.7 ± 4.0
D	16.7 ± 6.2	22.3 ± 9.3
TS	16.4 ± 3.7	19.8 ± 5.8

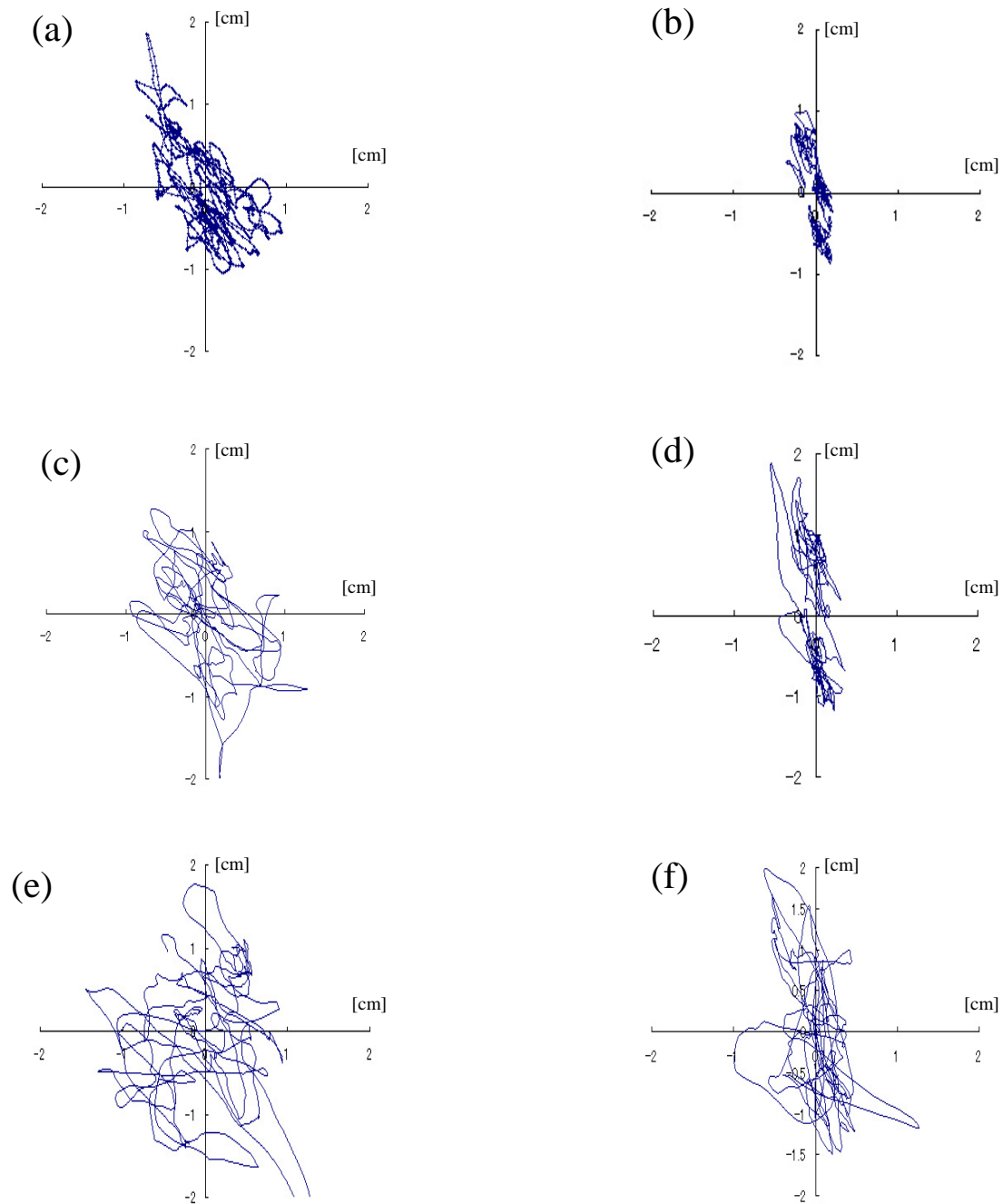


Fig. 1. Typical stabilograms (sway of the COP) observed when subjects viewed a static circle (a-b), the new stereoscopic movie (c-d), and the conventional 3D movie (e-f) [29].

to appear more often with the conventional 3D movie.

Typical stabilograms are shown in Fig. 1. In these figures, the vertical axis shows the anterior and posterior movements of the COP, and the horizontal axis shows the right and left movements of the COP. The amplitudes of the sway that were observed during exposure to the movies (Fig. 1c–1f) tended to be larger than those of the control sway (Fig. 1a–1b). Although a high density of COP was observed in the stabilograms (Fig. 1a–1b, 1e–1f), the density decreased in stabilograms during exposure to the

conventional stereoscopic movie (Fig. 1c–1d). Furthermore, stabilograms measured in an open leg posture with the midlines of heels 20 cm apart (Fig. 1b, 1d, 1f) were compared with stabilograms measured in the Romberg posture (Fig. 1a, 1c, 1e). COP was not isotropically dispersed but characterized by much movement in the anterior-posterior (y) direction (Fig. 1b, 1f). Although this trend is seen in Fig. 1d, the diffusion of COP was large in the lateral (x) direction and had spread to the extent that it was equivalent to the control stabilograms (Fig. 1a).

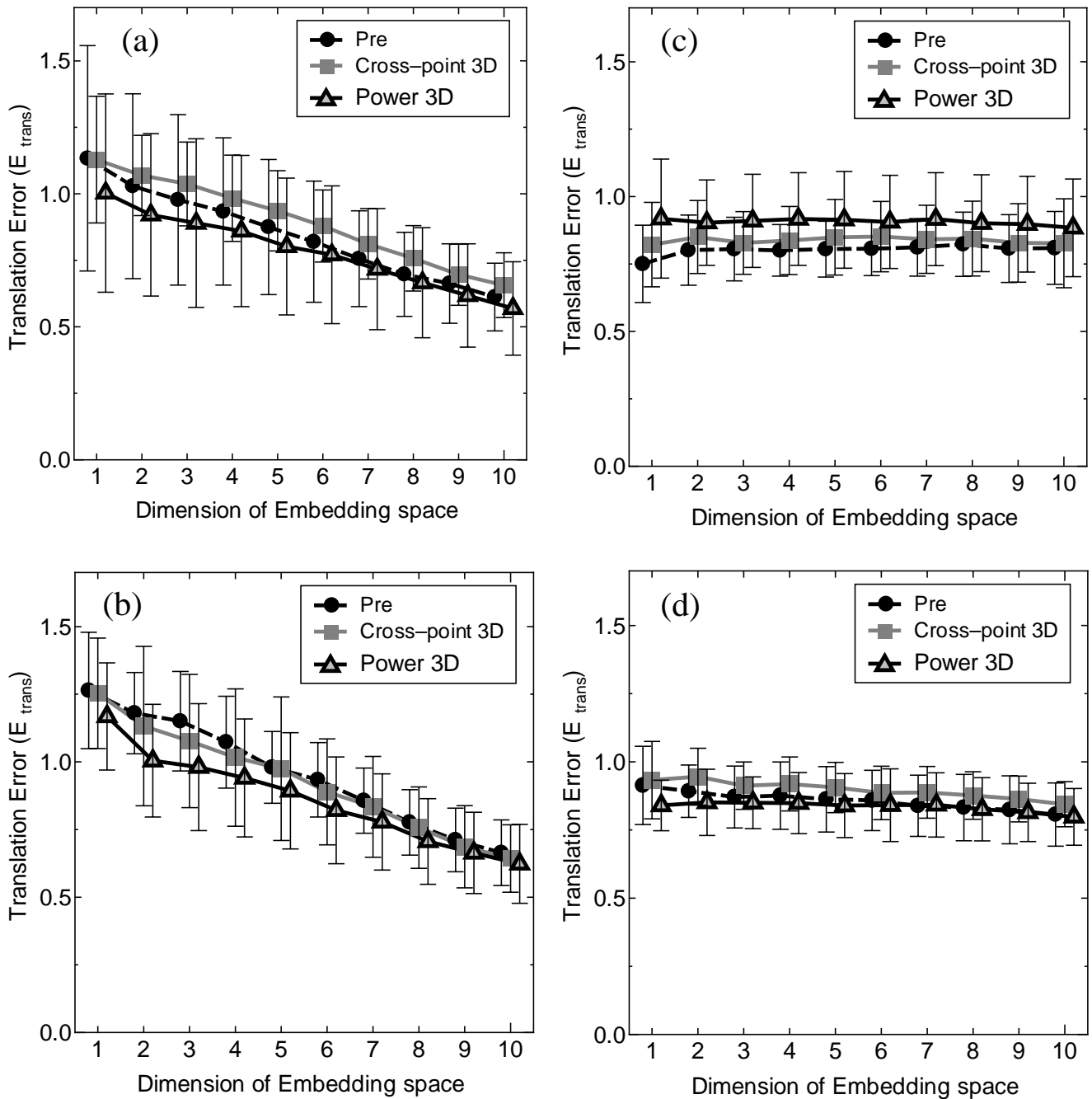


Fig. 2. Mean translation error (E_{trans}) for each embedding space. Translation errors were estimated from a lateral component of stabilograms (a)–(b), and temporal differences of the time series (c)–(d). Subjects maintained the Romberg posture (a), (c), and a wide stance (b), (d).

Results of the Double-Wayland algorithm are shown in Fig. 2 and Fig. 3. Whether subjects were exposed to the 3D movies or not, E_{trans} derived from the temporal differences of those time series x, y was approximately 1. These translation errors in each embedding space were not significantly different from the translation errors derived from the time series x, y although E_{trans} derived from the time series y is less than 1 for any embedding space without exposure to any of stereoscopic movies.

According to the two-way analysis of variance (ANOVA) with repeated measures, there was no interaction between factors of posture (Romberg posture or standing posture with their feet wide apart) and images (I), (II), or (III)). Except to the total locus length per unit area and the total locus length of chain, main effects were seen in the both factors (Fig. 4). On the other hand, any indicators could find a main effect in the postural factor ($p < 0.01$).

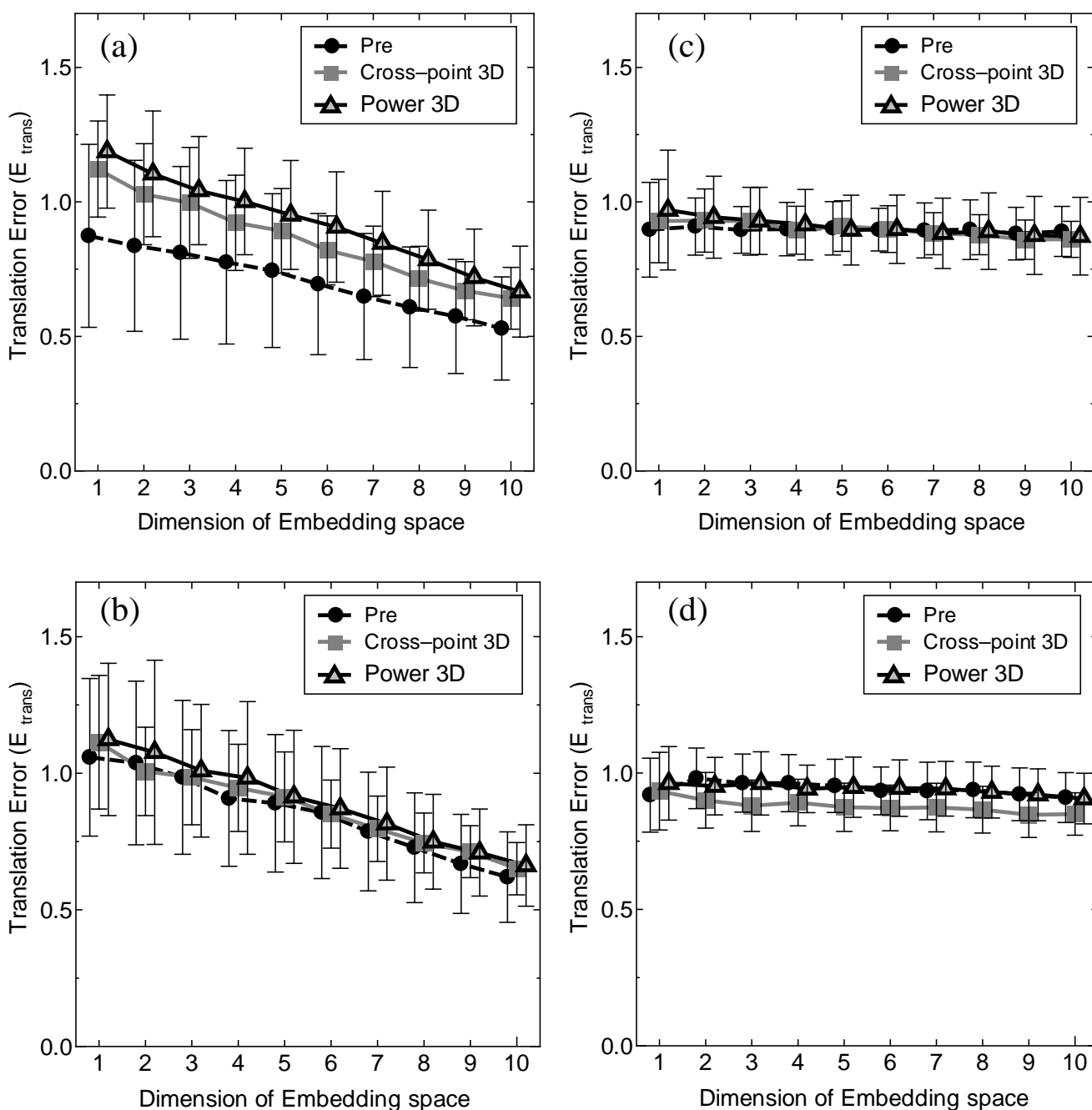


Fig. 3. Mean translation error (E_{trans}) for each embedding space. Translation errors were estimated from a anterior/posterior component of stabilograms (a)–(b), and temporal differences of the time series (c)–(d). Subjects maintained the Romberg posture (a), (c), and a wide stance (b), (d).

IV. DISCUSSION

A theory has been proposed to obtain SDEs as a mathematical model of the body sway on the basis of the stabilogram. According to Eq. (1), there were several minimal points of the time-averaged potential function in the SDEs (Fig. 1). The variance in the stabilogram depends on the form of the potential function in the SDE; therefore, the SPD is regarded as an index for its measurement. The movies,

especially stereoscopic images, decrease the gradient of the potential function. The new 3D movie (II) should reduce the body sway because there is no disagreement between vergence and visual accommodation. The reduction can be evaluated by the SPD during exposure to the movies on an LCD screen. Performing a one-way analysis of variance for a posture with wide stance, we have succeeded in estimating the decrease in the gradient of the potential function by using the SPD as shown in Fig. 4a ($p < 0.05$).

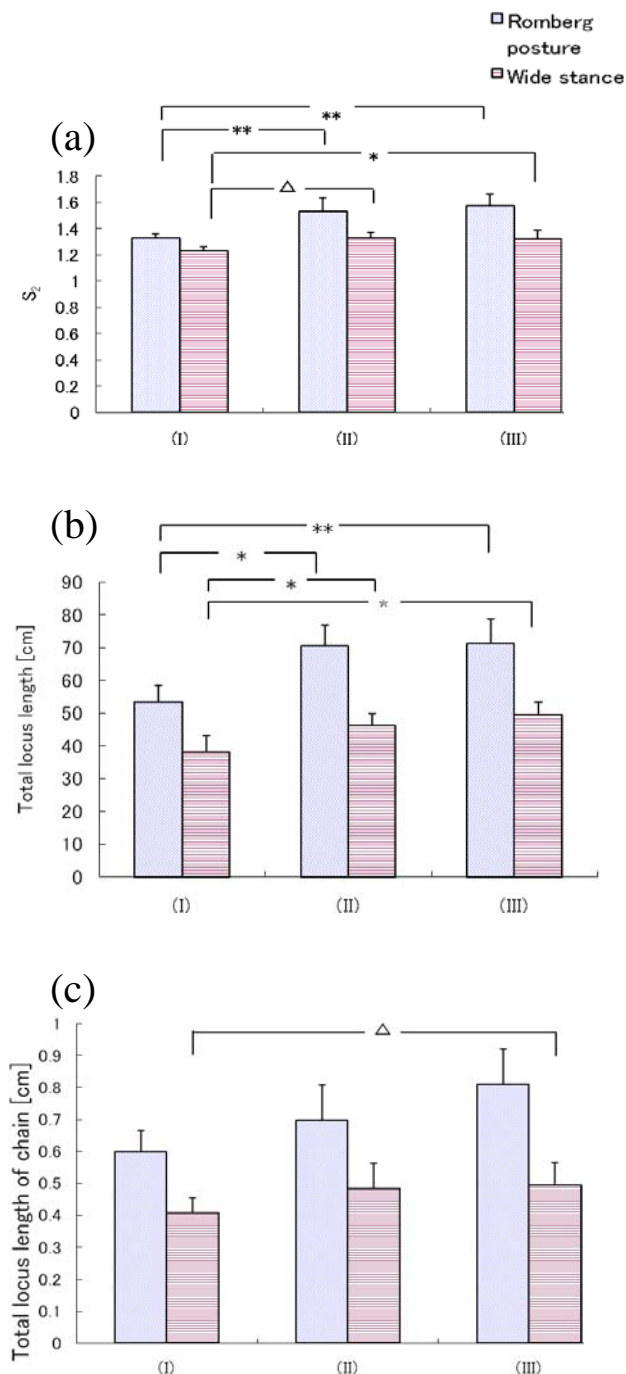


Fig. 4 Typical results of the two-way ANOVA with repeated measures for indicators [29]: the SPD (a), the total locus length (b), and the total locus length of chain (c) (** $p < 0.01$, * $p < 0.05$).

In this study, we mathematically measured the degree of determinism in the dynamics of the sway of COP. The Double-Wayland algorithm was used as a novel method. $E_{\text{trans}} > 0.5$ was obtained by the Wayland algorithm (Fig. 2-3), which implies that the time series could be generated by a stochastic process in accordance with a previous standard [30]. The threshold 0.5 is half of the translation error resulting from a

random walk. The body sway has been described previously by stochastic processes [4]-[7], which was shown with the Double-Wayland algorithm [31]. Moreover, $0.8 < E_{\text{trans}} < 1$ obtained from the temporal differences of these time series exceeded the translation errors estimated by the Wayland algorithm, as shown in Fig. 2b. However, the translation errors estimated by the Wayland algorithm were similar to those obtained from the temporal differences, except for Fig. 2b, which agrees with the abovementioned explanation of the dynamics to control a standing posture. The exposure to 3D movies would not change it into a deterministic one. Mechanical variations were not observed in the locomotion of the COP. We assumed that the COP was controlled by a stationary process, and the sway during exposure to the static control image (I) could be compared with that when the subject viewed 3D movies. Indices for stabilograms might reflect the coefficients in stochastic processes although the translation error did not exhibit a significant difference between the stabilograms measured during exposure to the new 3D movie (II) and the conventional 3D movie (III).

Indices for stabilograms might reflect the coefficients in stochastic processes although the translation error did not exhibit a significant difference among the exposure to images (I), (II), and (III) as shown in Fig.2-3. With respect to the Romberg posture, the total locus length during exposure to 3-D movies was significantly greater than that to the static one (I) which could not induce the VIMS (Fig. 4b). We considered that the 3-D images on the LCD decrease the gradient of the potential function. Moreover, the new 3D movie (II) might reduce the body sway because there is no disagreement between vergence and visual accommodation. The reduction could be evaluated by the SPD during exposure to the movies on an LCD screen while subjects maintained upright posture with the wide stance (Fig. 4a). We have succeeded in estimating the decrease in the gradient of the potential function by using the SPD. We concluded that the metamorphism in the potential function during exposure to the conventional 3-D images could be detected by using the SPD.

Multiple comparisons indicated that the SPD S_2 during exposure to any of the stereoscopic movies was significantly larger than that during exposure to the static control image (I) when subjects stood in the Romberg posture (Fig.4a). The standing posture would become unstable because of the effects of the stereoscopic movies. As mentioned above, structural changes occur in the time-averaged potential function (1) with exposure to stereoscopic images, which are assumed to reflect the sway in center of gravity.

Scibora et al. concluded that the total locus length of subjects with prior experience of motion sickness increases with exposure to a virtual environment when they stood with their feet wide apart [23], whereas, in our study, the degree of sway was found to be reduced when the subjects stood with their feet wide apart (Fig.1b, 1d, 1f) than when they stood with their feet close together (Fig.1a, 1c, 1e). As shown in Fig. 1d and 1f, a clear change in the form of the potential function (1) occurs when the feet are wide apart. The decrease in the gradient of the potential might increase the total locus length.

Regardless of posture, the total locus length during exposure to the 3D movies was significantly greater than that during exposure to the control image (Fig.4b). However, the SPD during exposure to the conventional stereoscopic movie (III) was significantly larger than that during exposure to the control image (I) when they stood with their feet wide apart (Fig.4a). The total locus length of chain simultaneously tended to increase when subjects were exposed to the conventional 3D images (III) compared that when they were exposed to (I) (Fig.4c). Hence, we noted postural instability with the exposure to the conventional stereoscopic images (III) by using these indicators involved in the stabilogram (SPD and total locus length of chain). This instability might be reduced by the Olympus power 3D method.

REFERENCES

- [1] T. Okawa, T. Tokita, Y. Shibata, T. Ogawa, H. Miyata, "Stabilometry - Significance of Locus Length Per Unit Area (L/A) in Patients with Equilibrium Disturbances," *Equilibrium Res.*, vol.55(3), pp.283-293, 1995.
- [2] K. Kaga, "Memaino Kouzo: Structure of vertigo," Tokyo: Kanehara, 1992, pp.23-26 [in Japanese].
- [3] T. Okawa, T. Tokita, Y. Shibata, T. Ogawa, H. Miyata, "Stabilometry-Significance of locus length per unit area (L/A)," *Equilibrium Res.*, vol.54(3), pp.296-306, 1996.
- [4] J. J. Collins, C. J. De Luca, "Open-loop and closed-loop control of posture: A random-walk analysis of center of pressure trajectories," *Exp. Brain Res.*, vol. 95, pp.308-318, 1993.
- [5] R. E. A. Emmerik, R. L. Van Sprague, K. M. Newell, "Assessment of sway dynamics in tardive dyskinesia and developmental disability: sway profile orientation and stereotypy," *Moving Disorders*, vol.8, pp.305-314, 1993.
- [6] K. M. Newell, S. M. Slobounov, E. S. Slobounova, P. C. Molenaar, "Stochastic processes in postural center-of-pressure profiles," *Exp. Brain Res.*, vol.113, pp.158-164, 1997.
- [7] H. Takada, Y. Kitaoka, Y. Shimizu, "Mathematical Index and Model in Stabilometry," *Forma*, vol.16 (1), pp.17-46, 2001.
- [8] P. A. Goldie, T. M. Bach, O. M. Evans, "Force platform measures for evaluating postural control: reliability and validity," *Arch. Phys. Med. Rehabil.*, vol.70, pp.510-517, 1989.
- [9] K. Fujiwara, H. Toyama, "Analysis of dynamic balance and its training effect-Focusing on fall problem of elder persons," *Bulletin of the Physical Fitness Research Institute*, vol.83, pp.123-134, 1993.
- [10] T. A. Stoffregen, L. J. Hettinger, M. W. Haas, M. M. Roe, L. J. Smart, "Postural instability and motion sickness in a fixed-base flight simulator," *Human Factors*, vol.42, pp.458-469, 2000.
- [11] G. E. Riccio, T.A. Stoffregen, "An Ecological theory of motion sickness and postural instability," *Ecological Physiology*, vol.3(3), pp.195-240, 1991.
- [12] C. Oman, "A heuristic mathematical model for the dynamics of sensory conflict and motion sickness," *Acta Otolaryngologica Supplement*, vol.392, pp.1-44, 1982.
- [13] J. Reason, "Motion sickness adaptation: a neural mismatch model," *J. Royal Soc. Med.*, vol.71, pp.819-829, 1978.
- [14] T. A. Stoffregen, L. J. Smart, B. J. Bardy, R. J. Pagulayan, "Postural stabilization of looking. *Journal of Experimental Psychology*," *Human Perception and Performance*, vol.25, pp.1641-1658, 1999.
- [15] H. Takada, K. Fujikake, M. Miyao, Y. Matsuura, "Indices to Detect Visually Induced Motion Sickness using Stabilometry," *Proc. VIMS 2007*, pp.178-183, 2007.
- [16] T. Hatada, *Nikkei electronics*, vol.444, pp.205-223, 1988.
- [17] R. Yasui, I. Matsuda, H. Kakeya, "Combining volumetric edge display and multiview display for expression of natural 3D images," *Proc. SPIE 6055, 0Y1-0Y9*, 2006.
- [18] H. Kakeya, "MOEvision:simple multiview display with clear floating image," *Proc. SPIE 6490, 64900J*, 2007.
- [19] R. S. Kennedy, N. E. Lane, K. S. Berbaum, M. G. Lilienthal, "A simulator sickness questionnaire (SSQ): A new method for quantifying simulator sickness," *International J. Aviation Psychology*, vol.3, pp.203-220, 1993.
- [20] S. R. Holomes, M. J. Griffin, "Correlation between heart rate and the severity of motion sickness caused by optokinetic stimulation," *J. Psychophysiology*, vol.15, pp.35-42, 2001.
- [21] N. Himi, T. Koga, E. Nakamura, M. Kobashi, M. Yamane, K. Tsujioka, "Differences in autonomic responses between subjects with and without nausea while watching an irregularly oscillating video," *Autonomic Neuroscience Basic and Clinical*, vol.116, pp.46-53, 2004.
- [22] Y. Yokota, M. Aoki, K. Mizuta, "Motion sickness susceptibility associated with visually induced postural instability and cardiac autonomic responses in healthy subjects," *Acta Otolaryngologica*, vol.125, pp.280-285, 2005.
- [23] L. M. Scibora, S. Villard, B. Bardy, T.A. Stoffregen, "Wider stance reduces body sway and motion sickness," *Proc. VIMS 2007*, pp.18-23, 2007.
- [24] H. Takada, T. Morimoto, H. Tsunashima, T. Yamazaki, H. Hoshina, M. Miyao, "Applications of Double-Wayland algorithm to detect anomalous signals," *FORMA*, vol.21 (2), pp.159-167, 2006.
- [25] T. Nishihara, H. Tahara, "Apparatus for recovering eyesight utilizing stereoscopic video and method for displaying stereoscopic video," *US Patent 7404639*, 2008.
- [26] J. Suzuki, T. Matsunaga, K. Tokumatsu, K. Taguchi, Y. Watanabe, "Q&A and a manual in Stabilometry," *Equilibrium Res.*, vol.55(1), pp.64-77, 1996.
- [27] H. Takada, Y. Kitaoka, S. Ichikawa, M. Miyao, "Physical Meaning on Geometrical Index for Stabilometry," *Equilibrium Res.*, vol.62(3), pp.168-180, 2003.
- [28] R. Wayland, D. Bromley, D. Pickett, A. Passamante, "Recognizing determinism in a time series," *Phys. Rev. Lett.*, vol.70, pp.530-582, 1993.
- [29] H. Takada, K. Fujikake, T. Watanabe, S. Hasegawa, M. Omori, M. Miyao, "On a method to evaluate motion sickness induced by stereoscopic images on HMD," *Proceedings of the IS&T/SPIE 21st Annual Symposium on Electronic Imaging Science and Technology*, pp.72371P-1 2009.
- [30] T. Matsumoto, R. Tokunaga, T. Miyano, I. Tokuda, "Chaos and Time Series," Tokyo: Baihukan, 2002, pp.49-64 [in Japanese].
- [31] H. Takada, Y. Shimizu, H. Hoshina, Y. Shiozawa, "Wayland tests for differenced time series could evaluate degrees of visible determinism," *Bulletin of Society for Science on Form*, vol.17(3), pp.301-310, 2005.

An Assistive Technology Computer Control System

Eduardo J. Alberti, Tatiany C. Kazmiecak and Alessandro Brawerman
Computer Engineering Department, University of Positivo
Curitiba, Brazil
{eduardoalberti, tatiany, brawerman}@up.edu.br

Abstract— With the advent of technology and digital inclusion, people of all social classes have been presented to the computing world. However, there is a population that, by physical limitations, has no easy and non-expensive means of interacting with a microcomputer, the motor-disabled individuals. This project presents an assistive technology system, which helps motor-disabled individuals to interact with a personal computer, consisting of a control system that captures the individual movements of the tongue and sends the signals to control the computer. A virtual keyboard, with the ability to complete words based on the user's vocabulary, is also proposed, improving even more one's experience. Experiments in real use scenarios are presented to state the feasibility of the system.

Keywords- Assistive technology, computer control system, motor-disabled individual.

I. INTRODUCTION

Brazil has currently approximately 183.9 million inhabitants, of which 24.3 million have some form of physical or mental disability. According to the IBGE (research year 2000), 1.4 million of these people are tetraplegic [1].

Tetraplegia, also known as quadriplegia, is paralysis caused by illness or injury to a human that results in the partial or total loss of use of all their limbs and torso. The loss is usually sensory and motor, which means that both sensation and control are lost [2].

The Brazilian Law 7853 of 1989 supports the physically disable individuals ensuring the exercise of individual and social rights of persons with disabilities, citing Article 1 § 1: "In the application and interpretation of this Law shall be considered the basic values of equal treatment and opportunity, social justice, respect for human dignity, welfare, and others listed in the Constitution or justified by general principles of Law" [3]. One can consider core values of equal treatment and opportunity to be education access, information and personal independence.

With the objective of providing personal independence and digital inclusion to motor-disabled people, mainly tetraplegic, this project presents a non-invasive assistive technology system that captures, interprets and transmits signals resulting from movements of the tongue, allowing the interaction of users with a computer.

Besides the assistive technology system, a virtual keyboard, with the ability to complete words based on the user's vocabulary, is also proposed.

This article not only describes the specification and development of the assistive technology system and the virtual keyboard, but also presents experiment results obtained when using the system. It is divided as follows, some of the most relevant related work is presented in section II, the specification and development of the system is showed in section III, section IV brings the experimentation and validation results, and finally section V concludes the work.

II. RELATED WORK

This section presents similar works whose objective is to help impaired individuals to obtain a certain degree of liberty and improve their way of life.

In [4], the authors present an educational game with modification in controls so that disability individuals can play and enjoy the game. They proposed a new Sudoku game for people whose motion is impaired, called Sudoku Access. Their special interface allowed the control of the game either by voice or by a single switch. As in our system, their solution did not focus every disabled person, but at least it benefits a lot a small portion of that group.

In [5], the authors present a great project called the Camera Mouse System. The objective is the same as in our project, i.e., to provide computer interaction for people with severe disabilities. The system tracks the user's movements with a video camera and translates them into the movements of the mouse pointer on the screen. Body features such as the tip of the user's nose or finger can be tracked. The visual tracking algorithm is based on cropping an online template of the tracked feature from the current image frame and testing where this template correlates in the subsequent frame. The location of the highest correlation is interpreted as the new location of the feature in the subsequent frame.

The project presented in [5] and our project are very similar regarding the objective and managing the mouse in a different way to interact with the computer. However, we include the design of a virtual keyboard to allow users to type words and have a better experience, using any application provided by the underlying operating system. Besides and more important, our assistive technology hardware will be

increased not only to control a computer, but also a wheelchair, a TV, and other electro-electronic utilities.

A very interesting research study was conducted in [6]. The objective was to formulate a better practice model for the application of VR (Virtual Reality) intervention for adults with intellectual and developmental disabilities (IDD). The research group participated in an 8 week VR program using GestureTek's IREX video capture technology operated by the local caregiver staff. The VR programs were found to attract full participation by the participants at moderate levels of IDD but some difficulties were found in fully engaging all individuals at severe levels of IDD. Different commercial VR systems were used and were found to be usable by health-profession students and local caregiver staff. Significant improvements in physical fitness were demonstrated by the research group.

Finally, in [7] a prototype wheelchair with legs for people with motor disabilities was proposed. The objective was to demonstrate the feasibility of a completely new approach to mobility. The authors' prototype system consisted of a chair equipped with wheels and legs, and is capable of traversing uneven terrain and circumventing obstacles. The important design considerations, the system design and analysis, and an experimental prototype of a chair were discussed. The results from the analysis and experimentation tried to show the feasibility of the proposed concept and its advantages

III. SPECIFICATION AND DEVELOPMENT

The development of a system like this, capable of bringing some independence to the handicapped, is of great value to our society. By including the tetraplegic in the digital realm, the system makes them able to perform daily tasks that they were not able before, providing personal independence and digital inclusion.

With this assumption, a research was initiated to develop a comfortable, practical and aesthetically pleasing device, which is inserted in the user's mouth. According to [8], the tongue, being an organ attached to the lower face (jaw) and floor of mouth, fits better to prosthesis installed on the lingual surface of lower incisors, thus making the movements more comfortable and less tiring.

With the aid of dental professionals we have developed a model of partial prosthesis, composed of two micro-switches and a mini joystick that can be nicely accommodated in the user's jaw. The joystick moves to 4 sides (up, down, left, right) and can also be used as a mouse wheel when pressed. The two micro-switches are employed as the left and right mouse buttons. The connection between the prosthesis buttons and the controller circuit is made by wire resin, reducing the thickness of wires, allowing an individual to use the prosthetic device with his/her mouth closed. Figure 1 depicts the prosthesis being constructed.

The user communication with his/her personal computer is done in a wireless fashion, sending all motion captured to the PC through radio signals. The software in the PC captures, interprets and executes the commands sent by the users.

A. Prosthesis

Before dealing with peripherals or the software, the development of the dental prosthesis was initiated. For this, it was taken into account some features of the electronic components involved (joystick and buttons) as protection against rust, against liquids and the possibility of a user to get a shock from the components. The parameters were analyzed and the electronic components were isolated to ensure user's safety.

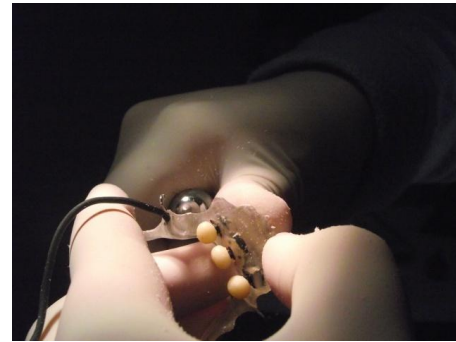


Figure 1 – Initial prosthesis model.

Since we adopted the use of a dental prosthesis, each user has to pass through a clinical phase consisting of an anatomical and physiological oral analysis. This might be considered a disadvantage since there is not a unique prosthetic device that would fit every individual. However, the construction of a dental prosthesis is costless, making the project feasible and non-expensive.

Figure 2 shows the final version of the prosthetic device. Note that the switches have been modified from earlier versions. The ones adopted in the final prosthesis showed to be more comfortable and easy to use.



Figure 2 – Final prosthesis model.

B. Assistive Technology System – Hardware

The project designs the hardware as a set of two elements, the transmitter module, inserted in the user's mouth and specified in Figure 3, and the receiver module, connected to a PC and specified in Figure 4.

The transmitter module is responsible for the data acquisition, interpretation and transmission. The

microcontroller, central core of this module, starts the variables used to capture movement and the RF (Radio Frequency) communication system. After the initialization step, the firmware in the microcontroller remains in a loop periodically checking the communication doors and obtaining new values for the Joystick, left and right switches. The serialized data is sent to the RF module, providing information to the receiver module.

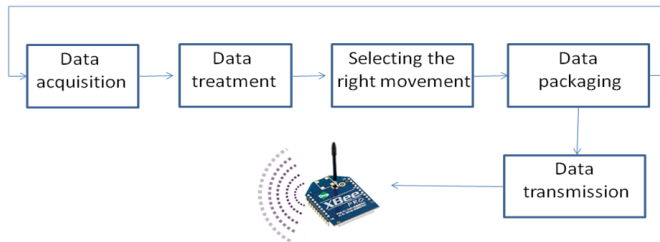


Figure 3 – Transmission module diagram.

When connected to the PC USB port, the receiver module is powered up. This module is responsible for interpreting commands from the transmitter module and encoding data using the HID Windows Class to control the mouse cursor, i.e., data sent by the user controls the mouse movement and mouse button clicks.

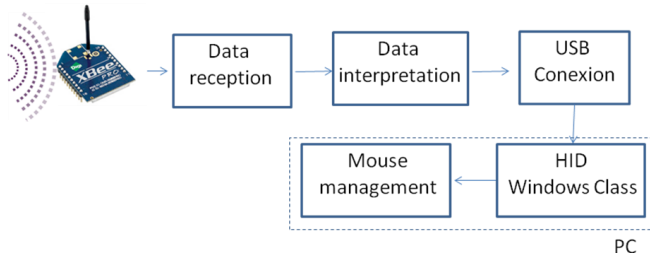


Figure 4 – Reception module diagram.

The HID Windows class is an interface for identifying human interface devices. It consists of device description classes providing information from BIOS setup and microcomputer manufacturer [9]. It allows mouse controlling without requiring the installation of additional drivers.

The commands received via radio frequency are interpreted, packaged and sent to the operating system in a vector format with four positions, each representing an operation to control the mouse, as shown in Table 1.

Table 1- Data sent to the computer in order to manage the mouse.

Vector Position	Action
0	Type of mouse click.
1	Pointer speed in the X axis.
2	Pointer speed in the Y axis.
3	Rotate the mouse wheel.

The PIC microcontroller 18F4550I was chosen because it offers the USB communication interface (necessary for connecting to the PC). Communication via radio transceiver is accomplished by the Fbee® modules in conjunction with the MiWi™ protocol stack, supporting peer to peer

communication. The P2P architecture allows a decentralization of the network, thus each module can have both roles of client and server, eliminating the need for a manager or infrastructure associated with it [10].

Powered with a DC voltage of 5V each module consumes approximately 80mA CC. The transmitter module, remotely located with the user, is powered by a 9V battery. This voltage is regulated internally. The receiver module is powered by the PC USB port.

The maximum distance between the modules is given by the RF modules maximum power. In this case, the option to use the Fbee transmitters has provided a maximum distance of approximately 150 meters between the transmitter and the receiver module, which in our case is much more than the necessary.

C. Software

With the hardware ready to use, the individual can interact with the computer by controlling the mouse. However, that was all that he/she could do. To improve the user's experience a virtual keyboard was also developed. Through this keyboard the user can type words and paste them to any application. In this fashion, one can use the Web through the interaction of our keyboard and a Web Browser, one can type words in an Editor, prepare a datasheet or presentation, ultimately one can use any application as a regular user.

Since our users control the mouse by moving their tongue, the virtual keyboard was designed to minimize the need for moving the mouse when typing. An intelligent algorithm was developed making the virtual keyboard able to complete words according to the probability of writing.

The probability calculation to complete a word is based in its incidence, making the algorithm flexible according to the user's vocabulary. To correctly complete a word, the algorithm queries the most probably word in a database. Initially, the database must be populated with a large number of words in a certain language, Portuguese in our case. Each word has a counter that accounts for its occurrence. Each time a word is used, its counter is incremented. Thus, the algorithm simply searches the word with the greatest counter beginning with the typed characters. The search only begins after the inclusion of at least two characters.

The searching algorithm returns the 50 most used words that start with the characters entered. From the feedback, the user is presented with the most probable word, i.e., the word with the highest occurrence. After selecting the word the algorithm writes it in the virtual keyboard text box. Words that have not yet been used and are not in the database might be easily inserted using the virtual keyboard.

Figure 5 depicts the virtual keyboard when a word is being typed. In this example the user wants to type "engineer", which in Portuguese is "engenharia". Note that after the two initially characters have been typed, the algorithm correctly completes the word automatically for the user. The user selects an application to send the word, in this case the Notepad is

selected, and press a button to paste it. The application is always showed in the background.

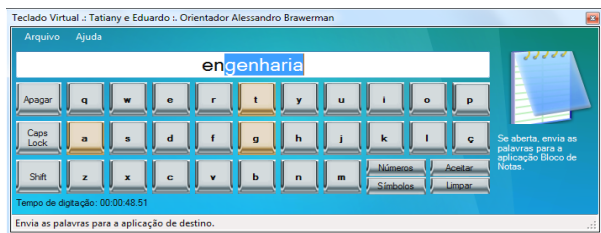


Figure 5 – Virtual keyboard model 2.

Not all possible applications are known by the virtual keyboard a priori, only the most common ones. If the user needs to interact with a new application, he/she only needs to register that application within the keyboard, which will interact with the desired application through the underlying operating system.

Because the system constantly queries the database, indexes were created on the tables in order to maintain a predefined structure to speed up searches.

A brief study on the virtual keyboard design was conducted. The idea is to minimize the need for moving the mouse when typing, hence making the user experience more comfortable. Afterwards, two interface models were projected. The first model, depicted in Figure 6, contains few buttons, and directional arrows to navigate to the next letters. The following most probably letters are emphasized in six smaller buttons located at the top of the keyboard, just below the text box. In the second model, depicted in Figure 5, the keys are arranged as in a standard keyboard with a larger size. The following most probably letters are also emphasized, but they are colored to stand out, instead of having them in special buttons.



Figure 6 – Virtual keyboard model 1.

The development of two interface models allowed a comparative study presented in the next section.

IV. TESTING AND VALIDATION

This section presents results obtained by conducting experiments in real use situations. The experiments performed were testing prosthesis adaptability, virtual keyboard performance and user's safety. The testing scenario employed a laptop computer with Intel® Dual Core 1.86GHz processor, 2GB of RAM and Windows® Seven as the operating system.

A. Prosthesis adaptability and virtual keyboards performance

Every assistive technology requires some training beforehand. Our system is not different. By wearing the prosthetic device developed, the user may feel some difficulty at the beginning. With this assumption, tests were performed to assess the time it takes to learn how to use the system satisfactorily and quickly.

The tests take into account the two virtual keyboard models presented earlier and indicate which one would be better to use. To perform these tests, users were given a summary of this work for typing in both models.

To calculate the performance of our searching algorithm we analyze how many characters the algorithm needed to correctly complete the word when it was first typed. The results showed that 57.89% of the words were completed correctly with less than three typed characters and 42.11% of the words were completed with an error of gender or needed four or more typed characters. It can be concluded that for every 100 words, 58 were correctly suggested in the very first time. After that, the error rate drops to less than 10%.

Another experiment was performed to check which virtual keyboard would present the smaller route to type words, i.e., which keyboard model would require less tongue movements from our user. Using the software OdoPlus [11] to measure the distance traveled by the mouse pointer, words containing opposite characters were typed.

Considering that the mouse pointer was initially located on the center of each keyboard the characters "za" were typed. It was found that in model 1, the distance traveled was 13.14cm, while in model 2 was 8.24 cm, representing a saving of 37.19% in the route. In a second test, typing the characters "ba" the distance was 2.87 cm in model 1 and 9.76 cm in model 2, which represents 240.07% increase on the route. In general, it can be noticed that model 2 is more intuitive because it is similar to regular keyboards, which ultimately leads to a higher performance, despite the greater distance to be traveled by the mouse pointer. The results are presented in Figure 7.

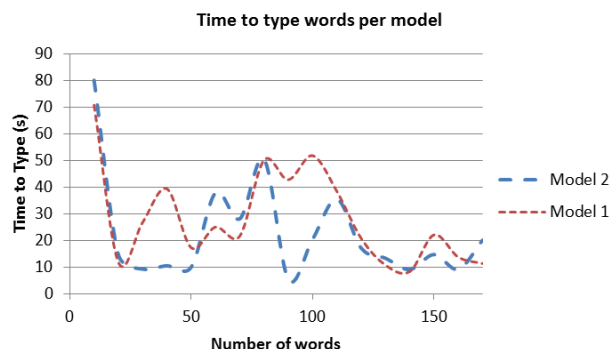


Figure 7: Time to complete words per keyboard model.

B. Algorithm adaptation and efficiency

The algorithm adaptation and efficiency experiment aims to determine the evolutionary behavior of the algorithm as the database gets populated when words are typed.

We used pieces of sections in this scientific article, for a total of 1000 words typed. Results showed that there is an increased efficiency of the suggestions in most sections entered. Table 2 shows the efficiency for every 250 words typed within text of the same subject. Considering the equation 1, which states that efficiency is equal to the saved typed characters divided by the total typed characters, the efficiency was calculated. Thus, efficiency is a rate that measures the saved characters when typing a text.

As it can be noticed, efficiency increases as the text is typed, stating that the algorithm gets used to the user's vocabulary. For instance, when typing the Hardware section, the efficiency rate was 26.60% at the beginning of the typing process, going up to 28.70% at the end of the process.

There is however some difficulty in helping to write in two situations: words of different genders, which is a problem depending on the language, like Portuguese, but not English, and differing words using their form in singular or plural. Suppose the database has stored that the words "player" occurred 10 times and "players" with 9 occurrences so far. The problem arises when one needs to write the second word. Since the singular form has occurred more times, the algorithm will keep suggesting the first word and to get the second word the user will have to type the plural form entirely.

Equation 1: Algorithm efficiency

$$efficiency = \frac{saved_chars}{total_chars}$$

Table 2- Efficiency of the algorithm suggestions

Theme	Beginning of typing	End of typing	Average
Abstract	4.30%	11.90%	8.10%
Introduction	16.40%	34.40%	25.40%
Hardware	26.60%	28.70%	27.65%
Results	25.30%	38.40%	31.85%

C. Reliability and prosthesis safety experiment

The reliability and prosthesis safety experiment aims to investigate the possibility of voltage existence in the prosthetic device and what, if any, is the prosthesis efficiency loss in real use conditions.

Since the prosthesis was the first piece built of this work, it could undergo a more extensive usability testing. During 10 months of use, tests were made in order to obtain an average wastage rate and loss of effectiveness rate.

The results showed that although through the naked eye the prosthesis seems totally isolated, micro-grooves allowed the infiltration of saliva into the buttons and joystick. In about nine months the prosthesis contacts did not respond. A more detailed analysis revealed that the contacts of the buttons were covered with verdigris, which resulted in loss of contact. These results allowed the observation of a definite time for loss of efficiency.

Furthermore, to improve the prosthesis quality, it was immersed in an aqueous solution simulating saliva. With the assistance of a precision multimeter, the voltage and current on the prosthesis were measured. The results were within the expected. The voltage presented in the prosthesis was 6 μ V and the current was 2mA to 5mA, variation occurred with the button pressed and unpressed. The conclusion of this particular experiment is that there is no harm to our user when wearing the dental prosthetic device.

V. CONCLUSION

This work presented an assistive technology system composed by a set of hardware and software. The system aims to allow the user, a disabled person, mainly a tetraplegic individual, to interact with a computer managing the mouse and typing words in order to use Web Browsers, Word Processors, Datasheet Applications and any other application offered by the underlying operating system.

To allow the user to interact with the computer, the system is formed by two pieces of hardware, the transmission module, which captures the tongue movements through a dental prosthetic device, interprets them and packages this input as data to be sent. The data package is received by the receiver module, unpackaged and by using the HID Windows Class, the movements made by the user are translated in mouse movements.

Two virtual keyboards were also designed in order to let the user type words by pressing the switches in the dental prosthesis. The user types words and chose to what application he/she wants to send the data typed. An intelligent algorithm was also presented aiming to accomplish word completion. Experiments showed that after the user types three, or sometimes less, characters the algorithm was able to achieve a 57.89% correctness rate for words typed for the very first time. The rate increases as that word is repeated in our user's vocabulary, achieving more than 90% of correctness rate.

Another important experiment measured the efficiency of our completion algorithm. The efficiency is a rate that measures the saved characters when typing a text. As it was showed by the results, efficiency increases as the text is typed, stating that the algorithm gets used to the user's vocabulary. Finally, a third experiment proved that there is no harm to our user when wearing the dental prosthetic device.

Future work includes increasing our assistive technology system to not only interact with a computer, but also control a wheelchair, a TV and other electro-electronic utilities, improving even further the impaired individuals' way of life.

REFERENCES

- [1] CEDIPOD. Census of disability in Brazil. Information and Documentation Centre of Disabled People. 2010. [Online] Available at: <http://www.cedipod.org.br/lbge1.htm>. Last viewed on 02/10/2011.
- [2] Bromley, Ida. Tetraplegia and Paraplegia, A guide for Physiotherapists. 6 ed. Philadelphia. Churchill Livingstone – Elsevier. 2006.
- [3] The Brazilian Government. Law 7853 of October 24th 1989. [Online] Available at: <http://www.planalto.gov.br/CCIVIL/LEIS/L7853.htm>. Last viewed on 02/10/2011.
- [4] Norte, Stéphane and Lobo Fernando G., Sudoku access: a sudoku game for people with motor disabilities. Proceedings of the 10th international ACM SIGACCESS Conference on Computers and Accessibility, 2008.
- [5] Betke, M.; Gips, J.; Fleming, P., The Camera Mouse: visual tracking of body features to provide computer access for people with severe disabilities. IEEE Transactions on Neural Systems and Rehabilitation Engineering. Volume 10, Issue 1, 2002.
- [6] Lotan, M. Yalon-Chamovitz, S. Weiss, P.L., Lessons learned towards a best practices model of virtual reality intervention for individuals with intellectual and developmental disability. Virtual Rehabilitation International Conference, 2009.
- [7] Wellman, P. Krovi, V. Kumar, V. Harwin, W., Design of a wheelchair with legs for people with motor disabilities. IEEE Transactions on Rehabilitation Engineering. Volume 3, Issue 4, 1995.
- [8] Reinhard, Putz and Reinhard, Pabst. Sobotta: Human Anatomy Atlas. 22. ed. Guanabara Koogan, Volume 1, 2006.
- [9] USB Implementers Forum. HID Information. [Online] Available at: <http://www.usb.org/developers/hidpage>. Last viewed on 02/10/2011.
- [10] Microchip Technologics Inc. Datasheet: MRF24J40, IEEE 802.15.4 2.4GHz RF Transceiver. Electronic publication, 2006.
- [11] UOL. OdoPlus. [Online] Available at: <http://ziggi.uol.com.br/downloads/odoplus>. Last viewed on 02/10/2011.

A Proposal of a Comprehensive Medical Emergency Decision Support System

Asma AlJarullah, Samir El-Masri

Department of Information Systems, College of Computer and Information Sciences,
King Saud University, Riyadh, Saudi Arabia

Abstract - *The aim of pre-hospital emergency medicine is to save lives, minimize sanitary harm and restore the quality of life as good as possible. Emergency pre-hospital care has the challenge of being time critical in nature. It requires rapid decision making despite the limited information around the patient which contributes to the high risk of medical errors. Although many clinical decision systems have been proposed many decades ago for the purpose of improving the quality of healthcare and reducing medical errors in clinics and emergency departments, but none of them had introduced a design of a decision support system for the pre-hospital emergency treatment. This paper introduces a high level design for a comprehensive medical emergency decision support system (CMEDSS). The major contribution of this paper is that it provides a framework for a medical emergency decision support system that addresses the challenges of pre-hospital emergency treatment through the use of the patient's electronic health record (EHR) and artificial intelligence techniques during the decision making process.*

Keywords: Emergency expert system; Intelligent decision support system; Electronic health record; Pre-hospital and emergency treatment.

1 Introduction

The aim of emergency medicine is to save lives, minimize sanitary harm and restore the quality of life as best as possible. Pre-hospital care has the challenge of being time critical in nature.

The literature has posted the potential role of health information technology in reducing emergency medical response times [1, 2] and improving the level and type of care provided to a patient through emergency care [3, 4, 5].

Indeed, next generation medical emergency systems have been identified as an essential component of healthcare systems that should enable decision support for an integrated voice and data emergency communications system [6]. To this aim, the Hatfield Report [7] provided

recommendations toward upgrading medical emergency systems infrastructures so that they can sufficiently address improvements and opportunities made available by existing technologies such as Internet Protocol (IP) networking standards, voice over IP (VOIP) communications, location identification techniques and public safety answering point (PSAP) processes and resources [8].

A systematic literature review by Garg et al. [9] of 100 studies concluded that "Clinical decision support systems improved practitioner performance in 64% of the studies and improved patient outcomes in 13% of the studies". Another literature systematic review of 70 studies by Kawamoto et al. [10] found that "Decision support systems significantly improved clinical practice in 68% of trials."

Clinical decision support (CDS) systems, with the potential to minimize practice variation and improve patient care, have begun to surface throughout the healthcare industry. Clinical Decision Support Systems are "active knowledge systems which use two or more items of patient data to generate case-specific advice" [11]. This implies that a CDSS is simply a DSS that is focused on using knowledge management in such a way to achieve clinical advice for patient care based on some number of items of patient data. CDSSs are aimed at supporting clinical diagnosis and treatment plan processes; and promoting use of best practices [12].

This paper presents a high level design for a Comprehensive Medical Emergency Decision Support System (CMEDSS) based on artificial intelligence that takes into consideration the patient's electronic health record in order to improve the quality of the decision making process in terms of both speed and accuracy.

The use of the patient's electronic health record in the medical emergency decision making process is very important. It has been reported in [15] through a systematic review with healthcare practitioners how having preexisting patient information (e.g., medications, pre-existing medical conditions, allergies, blood type, etc.) could significantly reduce data collection time and help to reduce medical

errors and to increase quality of care provision across the emergency response continuum of care.

Furthermore, the CMEDSS is designed in a way that permits the countries that didn't yet implement the concept of electronic health record (EHR) in their national healthcare centers to use the CMEDSS. This is because the decisions it makes don't totally depend on the patient's EHR, rather they are based on both the patient's EHR and his current status. So the CMEDSS can still function in the absence of the EHR.

The CMEDSS is proposed for emergency care ambulances and emergency departments where more accurate decisions are needed in critical time.

The paper is organized as follows: Section 2 provides a literature review on the related work on emergency medical DSSs. Section 3 provides an overview on the framework of the proposed CMEDSS. Section 3 discusses the process of the CMEDSS. And section 5 discusses the architecture and the components of the CMEDSS. Finally section 6 concludes the paper.

2 Related Work

Although medical decision support systems have been discussed extensively in the research, but researches introducing this technology into the area of medical emergency are too rare. We can summarize the researches that had introduced this technology to the area of emergency medicine below.

The first medical emergency decision support system was MEDAS - Medical Emergency Decision Assistant System, which was designed in 1980. MEDAS is a knowledge-based interactive diagnostic system which assists in diagnosis of multiple disorders in human body [13]. The knowledge base consists of disorder patterns that constitute the background medical information required for diagnosis in the emergency and critical care medicine. This system is designed to provide the clinician with decision aids from the time the patient is first seen in the emergency department until the immediate risk of life has been minimized. The system includes life support protocols, diagnosis, recommendations for data acquisition, guidelines for therapy, storage and retrieval of the patient record, and a consultant library that may be accessed in real time. An automatic knowledge acquisition system for MEDAS has been proposed in [14] which assist physicians to build, test, and verify the knowledge base for MEDAS without the involvement of knowledge engineer.

In [15, 16, and 17] a prototype has been made for a decision support system for medical Triage. Triage has been

defined as the process of categorization of casualties based on their need for medical attention [16, 17]. In medical triage, the treatment category determines the level of urgency of medical attention, and decisions based on nurse's primary observations must be produced in the shortest time possible. In emergency departments in Australia, the triage nurses use the Australian Triage Scale (ATS) to guide them through the triage decision-making process [18]. The Australian College of Emergency Medicine (ACEM) adopts the Australian Triage Scale (ATS) as part of its triage policy [19]. Because the accuracy of triage decisions has a major impact on whether or not a patient may receive medical intervention in an appropriate time frame, it is critical for the health outcomes of the patient. It is envisaged that by providing decision support tools to assist triage nurses in producing correct and timely triage decisions that are consistent with standard triage scales, triage decision support systems can contribute to the improvement of quality of life of triage patients and also reduce costs occurring from misappropriation of resources [20].

The CMEDSS proposed in this paper combines the strengths of the medical emergency decision support systems discussed above in addition to the usage of patient's electronic health record (EHR) as an input to the DSS in emergency ambulances and departments, which is the major contribution of this paper.

3 The CMEDSS Framework

The CMEDSS is based on three aspects; intelligent decision support system, national electronic health records and web-enabled decision support system. The following subsections give an overview on each of them.

3.1 Intelligent Decision Support System (IDSS)

Decision making work is now becoming more 'knowledge oriented' [22] and the need for more 'knowledge-driven' decision making support has laid the foundation to many artificial intelligence approaches and furthered the development of Intelligent Decision Support Systems (IDSS) [15].

Intelligent Decision Support Systems (IDSS) is a term that describes decision support systems that make extensive use of artificial intelligence (AI) techniques. The aim of the AI techniques embedded in an intelligent decision support system is to enable these systems to support decision makers by gathering and analyzing evidence, identifying and diagnosing problems, proposing possible courses of action and evaluating the proposed actions to be performed by a computer, whilst emulating human capabilities as closely as possible [23].

These DSSs are person-computer systems with specialized problem-solving expertise. The "expertise" consists of knowledge about a particular domain, understanding of problems within that domain, and "skill" at solving some of these problems [24]. IDSSs have been called suggestion DSS [25] and knowledge-based DSS [26].

3.2 National Electronic Health Records (EHR)

With the national movement towards data interoperability and standards development, electronic health record concept is at the heart of health informatics. Its purpose can be understood as a complete record of patient encounters that allows the automation and streamlining of the workflow in health care settings and increases safety through evidence-based decision support, quality management, and outcomes reporting [27].

An electronic health record (EHR) is an evolving concept defined as a systematic collection of electronic health information about individual patients or populations [28]. Such records may include a whole range of data in comprehensive or summary form, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal stats like age and weight, and billing information. [27].

The need for a better quality of service, unique identification of electronic health records, and efficient monitoring and administration requires a uniform and nation-wide organization for service and data access [29]. Several healthcare organizations worldwide are moving toward nationwide implementation of interoperable electronic health records. For instance, Canada Health Infoway [30] is an organization that provides specifications for a standard and nationwide healthcare infrastructure. The goal is to integrate information systems from different health providers and administrations (e.g., hospitals, laboratories, pharmacies, physicians, and government agencies) within each province, and then connect them to a nationwide healthcare network with standard data formats, communication protocols, and a unique health history file for each patient; where the health information is accessible ubiquitously, using common services according to different access privileges for patients and providers. Infoway's mission is to foster and accelerate the development and adoption of an interoperable Electronic Health Record (EHR) system [30].

National EHRs need a unique identifier for each record such as a national health record number which can be used to retrieve the patient's EHR.

3.3 Web-Enabled Decision Support System

Web services promote software portability and reusability in applications that operate over the Internet. They are a transition to service oriented, component-based, distributed applications. In other words, web services are applications implemented as Web based components with well-defined interfaces, which offer certain functionality to clients via the Internet [31].

Web-enabled DSSs are based on web services where users can access them through the Internet. All types of DSS can be deployed using Web technologies and can become Web-based DSSs. Figure 1 shows the interaction between the user and the decision support system through web services.



Figure 1. Web-enabled DSS

Web service architecture is built on open standards and vendor-neutral specifications i.e. they can be implemented in any programming language, deployed and then executed on any operating system or software platform [31].

4 The CMEDSS Process

The eventual network for EHR transactions is likely to be conducted over the World Wide Web, which is open, flexible and convenient [32]. For the purpose of being able to retrieve EHRs in mobility (i.e. in healthcare ambulance) and to centralize all the data and operations in one place and to reduce the load on client side (i.e. using thin clients), client-server architecture is the appropriate architecture for the proposed CMEDSS.

The CMEDSS is web-based which will be built on the server side of client-server architecture. The server side is located at a national healthcare center built around a country or several countries and keeps national EHRs for all people in that country.

At the server side, three-layer design is an effective approach to the development of robust and easy maintainable systems. This architecture is appropriate for CMEDSS that needs to support multiple user interfaces. The set of layers at the server side includes the following:

- Data layer - that manages stored data, usually in one or more databases. The data include EHRs and the knowledge base of the CMEDSS.

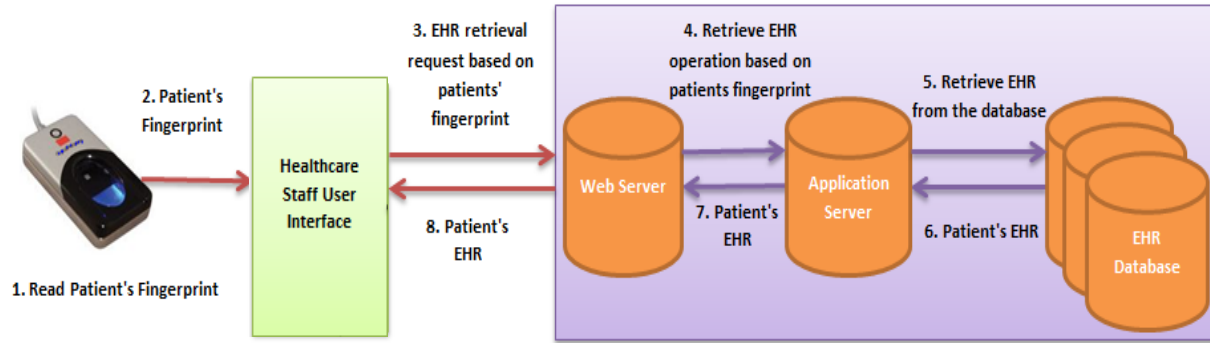


Figure 2. Real-time retrieval process of electronic health record based on the patient's fingerprint

- Business logic - (domain) layer that implements the rules and procedures of the business processing. It also includes the CMEDSS inference engine and the associated manipulation modules.
- View layer - that accepts input and formats and displays processing results.

As stated earlier, the national EHR needs a unique identifier. Using a national health record number to retrieve a patient's EHR is not suitable in emergency cases and especially in emergency ambulance care because the physicians have no means to know the patient's national health record number without being told. A better approach is to associate the patient's fingerprint with the EHR, so that in emergency cases, the patient's EHR can be retrieved easily by reading the patient's fingerprint using a fingerprint reader device.

Figure 2 shows the real time retrieval process of EHR. First, the healthcare staff uses a fingerprint reader device to read the patient's fingerprint. The patient's fingerprint is then entered into the emergency care (ambulance or department) computer. It is then used to retrieve the patient's EHR. The EHR is usually large and needs long time to be transferred to the emergency care computer; therefore a summary of the EHR can be requested to be transferred instead. The summary can be designed to include the standard and most important information about the patient (e.g., pre-existing medical conditions, allergies, blood type, etc.).

The CMEDSS core decision making processes have not been discussed yet. Figure 3 depicts the process of the CMEDSS. In order to get a decision from the CMEDSS, the physician inputs the patient's fingerprint and his current condition information into the physician user interface, and then submits this information to the server requesting a decision support.

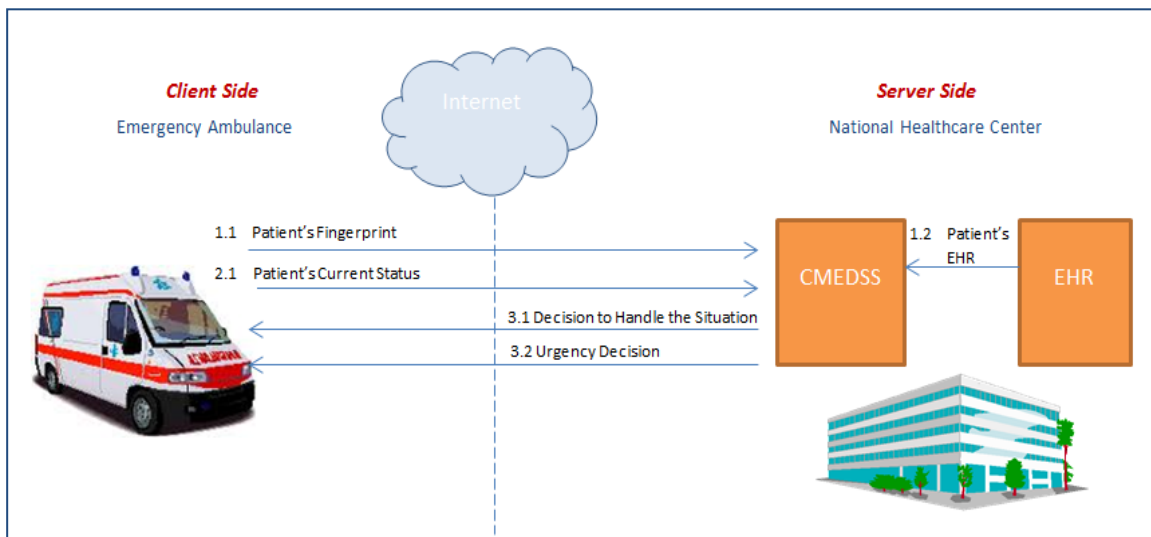


Figure 3. The CMEDSS Process

The server directs the request to the CMEDSS which is implemented internally at the server side. The CMEDSS retrieves the patient's EHR based on his fingerprint, extracts the important information from it, and use this extracted information along with the patient's current status information to output a decision regarding the patient's urgency level and a decision to handle the situation. The patient's urgency level determination is important in order to achieve an efficient appropriation of resources when the patient reaches an emergency department and improve the quality of care provided to the patient as recommended in [15, 16, and 17]. And the decision to handle the situation can assist physicians regarding the diagnosis in the emergency, critical care medicine, and pre-hospital treatment as proposed in [13]. The server then sends the CMEDSS output to the emergency physician user interface.

These operations will be highlighted in the following section which explains the structure and components of the CMEDSS and the responsibility of each component.

5 The Design of the CMEDSS

The design of the proposed CMEDSS is shown in Figure 4. The system has five components; the healthcare emergency staff user interface, the input module, the facts workspace, the inference engine and the knowledge base. These components are explained in detail in the following subsections.

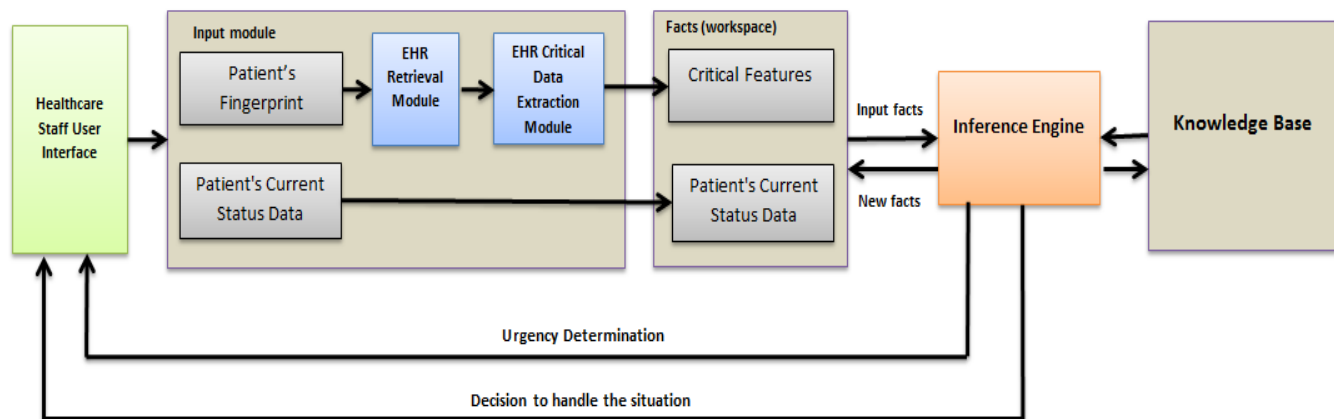


Figure 4. Structure of the proposed CMEDSS

5.1 Healthcare Emergency Staff User Interface

The user interface is the means of communication between a user and the CMEDSS problem-solving processes. The CMEDSS should include enhanced interfaces that enable automated data capture as opposed to

manual data entry and that “fit” with the healthcare staff’s emergency care processes as opposed to “getting in the way” of their time-critical work.

It has to be able to accept the queries or instructions in a form that the user enters and translate them into working instructions for the rest of the system. It also has to be able to translate the answers, produced by the system, into a form that the user can understand.

5.2 Input Module

The input module deals only with the information submitted by the user, it forms a gateway to the facts workspace. It accepts two inputs from the user, the patient’s fingerprint and his current status data. Then it uses the patient’s fingerprint to retrieve his EHR from the EHRs database, through the EHR retrieval module. The EHR is then passed to the EHR critical data extraction module, which extracts critical and important information from the EHR, this module is needed because the EHR is large and the CMEDSS does not need all information in it. The critical information needed by physicians needs to be determined by emergency and critical care experts (e.g. blood type, diabetic, heart disease, etc.), then defined in the EHR critical data extraction module. The input module then passes the extracted critical features regarding the patient and his current status data to the facts workspace.

5.3 Facts Workspace

The facts workspace is an area in the computer storage where CMEDSS stores the facts it has been given about situation and any additional information it has derived so far. It is also called blackboard, scratchpad or working storage. It includes the facts inserted by the input module

and the facts derived by the inference engine during the reasoning process.

5.4 Inference Engine

The heart of the CMEDSS is the inference engine. It accepts the input facts and chooses rules from the knowledge base to fire. Its implementation involves issues such as data structures, searching, sorting, pattern matching (recognition), and probability calculation. Many ready-made inference engines are available in the market and can be used for any intelligent DSS. Examples include OPS5, VP-Expert, EXSYS, KES, M.1, and Personal Consultant [33].

The inference engine produces two types of decisions, one regarding the urgency level of the patient status and one regarding a decision to handle the situation. The urgency level determination is important as proposed in [15, 16, 17, 18, 19, and 20] in order to ensure that the patient will receive medical intervention in an appropriate time frame when he reaches the hospital, it is critical for the health outcomes of the patient and also it reduces costs occurring from misappropriation of emergency resources. The decision to handle the situation, implements emergency medication standards to deal with the patient's situation including life support protocols, recommendations for data acquisition, diagnosis, proposing possible courses of action and therapy.

5.5 Knowledge Base

The expertise of the CMEDSS is represented in the knowledge base and the strength of the CMEDSS is reflected by the strength of the knowledge it possesses in its knowledge base. This includes best emergency care practices about emergency triage, diagnosis in the emergency, critical care medicine, and pre-hospital treatment of each specific health condition. All this knowledge must be taken from an expert and encoded into the knowledge base either by manual methods using a knowledge engineer as in [13] or using automatic knowledge acquisition methods like inductive learning, naïve Bayesian learning, generic algorithms learning, or artificial neural network learning.

6 Conclusions and Future Work

The use of decision support systems for enhancing quality and efficiency of medical decision-making has been flagged by many researchers. However, the greatest part of researches conducted around medical decision support systems were aimed to be used in clinics. Few researches proposed the use of decision support systems in emergency departments due to the greater challenges in emergency settings. And this paper is the first paper to introduce the

use of an emergency decision support system in emergency ambulances as well as in emergency departments.

This paper presents a high level design of a comprehensive medical emergency decision support system based on artificial intelligence and the patient's EHR. We proposed the use of the patient's EHR in the decision making process which is the major contribution of this paper (in addition to introducing the use of DSS in emergency ambulances). This system is expected to improve not only the quality of treatment provided to patients in pre-hospital emergency settings, but also the quality and timeliness of the emergency response when the patient reaches the emergency departments.

Future work includes the implementation, testing validation and refinement of the system.

7 Acknowledgement

This work is part of a two year research project which has been fully funded by a grant through King Abdul-Aziz City for Science and Technology (KACST) / National Plan for Science and Technology (NPST) in the Kingdom of Saudi Arabia. Grant number: 09-INF880-02

8 References

- [1] B. Schooley, T. Horan, and M. Marich. "User Perspectives on the Minnesota Inter-organizational Mayday Information System", in AMIS Monograph Series: Volume on Information Systems for Emergency Management, Van De Valle and Turoff, Eds.: IDEA Press, 2008.
- [2] J. Peters and Hall, B. "Assessment of ambulance response performance using a geographic information system," *Social Science and Medicine*, vol. 49, pp. 1551-1566, 1999.
- [3] M. Jason S. Shapiro, c, Joseph Kannry, MD, Andre W. Kushniruk, Gilad Kuperman, MD, PhD, e "Emergency Physicians. Perceptions of Health Information Exchange," pp. 700-705, 2007.
- [4] G. Mears, J. Ornato, and D. Dawson, "Emergency Medical Services Information Systems and A Future EMS National Database," in *Turtle Creek Conference III*, Dallas, TX, 2001.
- [5] T.A. Horan and B. Schooley, "Time-critical information services," *Commun. ACM*, vol. 50, pp. 73-78, 2007.
- [6] NHTSA, "Next Generation 9-1-1 System Concept of Operations (NG911 ConOps)," N. H. T. S. Administration, Ed., 2005.
- [7] D. N. Hatfield, "A Report on Technical and Operational Issues Impacting The Provision of Wireless Enhanced 911 Services," Federal Communications Commission, Washington, D.C. 2002.

- [8] L. K. Moore, "An Emergency Communications Safety Net: Integrating 911 and Other Services," Congressional Research Service, Washington, D.C. Feb 28, 2008.
- [9] Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J et al. (2005). "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review", *JAMA* 293 (10): 1223–38. doi:10.1001/jama.293.10.1223. PMID 15755945.
- [10] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, David F Lobach. (2005). "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success.". *BMJ*.
- [11] Wong HJ, Legnini MW, Whitmore HH. "The diffusion of decision support systems in healthcare: are we there yet?", *J Healthc Manag.* 2000 Jul-Aug;45(4):240-9; discussion 249-53.
- [12] Perreault L, Metzger J. "A pragmatic framework for understanding clinical decision support". *Journal of Healthcare Information Management.* 1999;13(2):5-21.
- [13] Mosshe Ben-Bassat, Richard W. Carlson, Venod K. Puri, Mark D. Davenport, John A. Schriver, Mohamed Latif, Ronald Smith, Larry D. Portigal, Edward H. Lipnick, and Max Harry Well. "Pattern-Based Interactive Diagnosis of Multiple Disorders: The MEDAS System", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, NO. 2, march 1980.
- [14] Li-Jen Chang and Martha Evens, Ph.D., David A. Trace, M.D., "A Knowledge Engineering System for MEDAS", Seventh Annual EEE Symposium on Computer-Based Medical Systems, IEEE, 1994.
- [15] Neha Padmanabhan, Frada Burstein, Leonid Churilov, Jeff Wassertheil, Bernard Hornblower, Nyree Parker, A "Mobile Emergency Triage Decision Support System Evaluation", Proceedings of the 39th Hawaii International Conference on System Sciences – 2006.
- [16] San Pedro J, Burstein F, Cao P, Churilov L, Zaslavsky A and Wassertheil J. "Mobile Decision Support for Triage in Emergency Departments". In Proceedings of the Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8.3 International Conference 2004, 714-723.
- [17] San Pedro J, Burstein F, Wassertheil J, Arora N, Churilov L and Zaslavsky A. "On the development and Evaluation of Prototype Mobile Decision Support for Hospital Triage". Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
- [18] Van der Loo RP, Van Gennip EMSJ, Bakker AR, Hasman A and Rutten FFH. "Evaluation of automated information systems in healthcare: An approach to classifying evaluative studies". *Computer Methods and Programs in Biomedicine*, 1995, 48, 45-52.
- [19] Victorian Department of Human Services. "Consistency of Triage in Victoria's Emergency Departments: Summary Report", Melbourne, Australia. 2001a.
- [20] San Pedro J, Burstein F, Cao P, Churilov L, Zaslavsky A and Wassertheil J. "Mobile Decision Support for Triage in Emergency Departments". Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS2004), 01 July 2004 to 03 July 2004, Monash University, Melbourne Vic Australia, pp. 714-723.
- [21] Webopedia, decision support system, http://www.webopedia.com/TERM/D/decision_support_system.html, April 2011.
- [22] Dhar V and Stein R. "Intelligent Decision Support Methods: The Science of Knowledge Work". Sydney: Prentice Hall, 1997.
- [23] Wikipedia, Intelligent decision support systems http://en.wikipedia.org/wiki/Intelligent_decision_support_systems, April 2011.
- [24] Power, D. J., "Decision Support Systems: Concepts and Resources for Managers", Westport, CT: Greenwood/Quorum, 2002.
- [25] Alter, S.L. "Decision Support Systems: Current Practice and Continuing Challenge". Reading, MA: Addison-Wesley, 1980.
- [26] Klein, M. and L. B. Methlie, "Knowledge-based Decision Support Systems with Applications in Business". Chichester, UK: John Wiley & Sons, 1995.
- [27] HIMSS - Electronic Health Record (EHR), http://www.himss.org/ASP/topics_ehr.asp, April 2011.
- [28] Tracy D Gunter, MD and Nicolas P Terry, LL.M., "The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions", *Med Internet Res.* 2005 Jan-Mar; 7(1): e3.
- [29] Kamran Sartipi, Mohammad H. Yarmand, and Douglas G. Down, "Mined-knowledge and Decision Support Services in Electronic Health", International Workshop on Systems Development in SOA Environments (SDSOA'07), opment in SOA Environments (SDSOA'07), IEEE, 2007.
- [30] Canada Health Infoway <http://www.infoway-inforoute.ca/>, April 2011
- [31] Yuri Boreish, "Web-Based Decision Support Systems As Knowledge Repositories for Knowledge Management Systems", http://www.ubicc.org/files/pdf/BoreishaUBICC_KE07_202_202.pdf, April, 2011.
- [32] Sydney Henrard, "Preliminary Instrumentation for the Efficient Use of Web-Based Electronic Health Records", Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04), IEEE, 2004.
- [33] C. W. Holsapple, Varghese S. Jacob, Andrew B. Whinston, "Operations research and artificial intelligence", Intellect Books, 1994.

BIOCAMP'11

Uncertain Gene Regulatory Networks Simplified by Gramian-Based Approach

Anke Meyer-Baese^a, Susanne Cappendijk^b and Fabian Theis^c

^a*Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, U.S.*

^b*Department of Biomedical Sciences, Florida State University, Tallahassee, FL 32306-4300, U.S.*

^c*Institute of Bioinformatics and Systems Biology, Helmholtz Center Munich, Ingolstädter Landstrasse 1 85764 Neuherberg, Germany*

Abstract

The complexity of gene regulatory networks described by coupled nonlinear differential equations is often an obstacle for analysis purposes. They are prone to internal parametrical fluctuations making thus robustness a crucial property of these networks to attenuate the effects of internal fluctuation. Therefore, the development of effective model reduction techniques for uncertain biological systems is of paramount importance in the field of systems biology. In this paper, we apply a Gramian-based approach for model reduction for gene regulatory networks based only on finding generalized Gramians and standard matrix transformations. The method is based on finding a generalized controllability and observability Gramian of the uncertain system and then based on a state transformation matrix a reduced-order representation. Under the assumption that the structured uncertainties are norm-bounded, we can prove that the reduced-order balanced system is also stable.

Key words: Gene regulatory network, uncertain system, model reduction

1 Introduction

Many gene regulatory networks are described by complex models which are difficult to analyze and also difficult to control. Analysis and synthetic design of such networks is very sensitive to parameter perturbations [1]. Errors in

Email address: ameyerbaese@fsu.edu (Anke Meyer-Baese).

parameters such as external perturbations and modeling errors are caused by data inaccuracies or computation errors. These perturbations can lead to location errors of equilibria, to instabilities, and even to spurious states [7]. Therefore, a rigorous understanding of the qualitative robustness properties of gene regulatory networks with respect to parameter variations becomes imperative [2]. On the other hand, order reduction may overcome some of the difficulties but at the price of a significant loss of accuracy. Therefore, a stringent need arises to analyze it such that it is made useful for many applications. The idea is to employ a model simplification that leads to a model of lower complexity, easier to handle, and to a simplified synthesis procedure for design problems. In addition, this simplification is reducing the computational complexity.

Balanced truncation is known as a popular method for model reduction since it is relatively simple and the quality of the reduced model is guaranteed. The interpretation of most balancing techniques is based on the concept of past and future energy. The most important contribution was the balancing for stable minimal linear systems [3]. It is based on a state-space point of view of employing the well-known observability and controllability Gramians and related to the past input energy (controllability) and future input energy (observability). The idea behind transforming a system into balanced form is to easily detect and remove a state component of the initial system to obtain a reduced-order model. The importance of a component is based on Hankel singular values which determine if the output energy of a certain component is small and thus difficult to observe and if the input energy to reach this state is large. While for linear systems finding a balancing coordinate transformation via solutions of the controllability and observability Lyapunov equations is quite easy, for nonlinear systems this equations are almost impossible to solve and thus balancing becomes in general not a simple task [5]. In a previous work [6], we applied a nonlinear model reduction technique for gene regulatory networks. However the very important concept of uncertainty paired with model simplification was not taken into account so far. We propose to apply and enhance the theoretical concepts from [8] to gene regulatory networks to obtain a stable model reduction under consideration of norm-bounded uncertainties. To the author's best knowledge, this method has not been applied so far to the analysis of gene regulatory networks.

The general kinetic equation describing the temporal evolution of the concentration for the j -th state and its output of a N -gene regulatory network is:

$$\begin{aligned} \dot{x}_i &= - \sum_{j=1}^N a_{ij}x_j + \sum_{j=1}^N b_{ij}x_ix_j \\ &+ \left(\sum_{j=1}^N c_{ij}x_j + \sum_{j=1}^N d_{ij}x_jx_i \right)u_i \\ y_i &= x_i \end{aligned} \tag{1}$$

where x_i is the current concentration state, y_i the current output of the gene regulatory network, and u_i is the external input, and m_{ij} a_{ij} , b_{ij} , c_{ij} and d_{ij} are the kinetic parameters associated with these reaction equations.

2 Global Asymptotic Stability Criteria for Quadratic Differential Equations

The general kinetic equation describing the temporal evolution of the gene regulatory networks (1) has a quadratic nonlinear term given as:

$$\dot{x}_i = - \sum_{j=1}^N a_{ij}x_j + \sum_{i=1}^N b_{ij}x_ix_j \tag{2}$$

In state space representation, we obtain the following general form:

$$\dot{x} = Ax + [B_1^T x, \dots, B_N^T x]^T x \tag{3}$$

where $A = a_{ij}$ and B_i^T is given as

$$B_i^T = \begin{pmatrix} 0 & \dots & 0 \\ b_{1i} & \dots & b_{Ni} \\ 0 & \dots & 0 \end{pmatrix} \tag{4}$$

A Lyapunov function for the above system is given as [4]

$$V = x^T Px, \quad P > 0 \quad P = P^T \tag{5}$$

with

$$A^T P + PA = -Q, \quad Q > 0, \quad Q = Q^T \tag{6}$$

guaranteeing thus the asymptotic stability of system (3) in the whole. Additionally, we need to require that $\dot{V} < 0$ for all $x \neq 0$. This leads to

$$\dot{V} = x^T(PA + A^T P)x + 2x^T P[B_1^T x, \dots, B_N^T x]^T x \quad (7)$$

\dot{V} is negative definite if and only if all the third-order terms it contains are identically zero, i.e.

$$x^T P[B_1^T x, \dots, B_N^T x]^T x = 0 \quad (8)$$

By choosing $Q = I$, we obtain assuming A is symmetric:

$$P = -\frac{1}{2}A^{-1} \quad (9)$$

The resulting stability condition for our system is:

$$\sum_{i=1}^N \tilde{a}_{ij} x_i^2 \sum_{j=1}^N b_{ij} x_j = 0 \quad (10)$$

where \tilde{a}_{ij} represent the elements of the inverse matrix.

3 Problem Statement

Notations:

$L_2^m = L_2^m[0, \infty)$ is the space of square integrable functions in R^m .

$\|\Delta\| = \sup_{z \in L_2^m[0, \infty), z \neq 0} (\|\Delta z\| / \|z\|)$ is the gain of an operator Δ in $\mathcal{L}(L_2^m)$

Δ^T is the adjoint operator of Δ if Δ is linear.

If $\Delta = \Delta^T$, then $\Delta < 0$ means that $x^T \Delta x < 0, \forall x \neq 0$ in R^m .

$\mathcal{L}(L_2^m)$ is the space of all linear bounded operators mapping from L_2^m to L_2^m .

$|\cdot|$ is the Euclidean norm in R^n .

M^T is the transpose of a complex matrix M .

$|z|_\Lambda^2 = z^T \Lambda z$ for $z \in R^m$ and a nonnegative matrix $\Lambda \in R^{m \times m}$.

State space representation of a transfer matrix is given as $G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] = C(sI - A)^{-1}B + D$

In the following, we will demonstrate the application of the model reduction based on balanced truncation.

For the sake of simplicity, we will consider a restricted state domain where the nonlinearity can be approximated by a linear function, $f(x_i) = x_i$.

$$\dot{x}_j = -l_j x_j + \sum_{i=1}^N D_{ij} x_i + \sum_{i=1}^p m_{ij} u_i \quad (11)$$

Thus, the system has a linear representation of the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) \end{aligned} \quad (12)$$

with $C = I$ and

$$A = D - L \quad \text{and} \quad B = M \quad (13)$$

It is assumed that the linear system is stable: $A = D - L$ is Hurwitz. We will assume that matrix D is a symmetric matrix.

Let us consider the uncertainty structure

$$\Delta^c = \left\{ \text{diag}(\Delta_1, \dots, \Delta_k) : \Delta_i \in \mathcal{L}(L_2^{h_i}), \Delta_i \text{ causal}, \|\Delta_i\| \leq 1 \right\} \quad (14)$$

resulting into the following uncertain gene regulatory network:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + E\zeta + Bu(t) \\ z(t) &= Kx(t) \\ y(t) &= Cx(t) \\ \zeta(t) &= \Delta z(t), \quad \Delta \in \Delta^c \end{aligned} \quad (15)$$

with $C = I$ and B, E, K are diagonal matrices. $x(t) \in R^n$ is the state, $u(t) \in R^m$ is the control input, $z(t) \in R^h$ is the uncertainty output, $y(t) \in R^l$ is

the measured output and $\zeta(t) \in R^h$ is the uncertainty input. We also have $h = h_1 + \dots, h_k$.

We thus obtain a nominal system as

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \left(\begin{array}{c|cc} A & E & B \\ \hline K & 0_{h \times h} & 0_{h \times m} \\ C & 0_{l \times h} & 0_{l \times m} \end{array} \right) \quad (16)$$

The uncertain system (15) is defined by a linear fractional transformation representation as $\mathcal{F}_u(M, \Delta) := M_{22} + M_{21}\Delta(I - M_{11}\Delta)^{-1}M_{12}$ if $I - M_{11}\Delta$ is non-singular.

We will define the following operators:

$$\begin{bmatrix} A_\Delta & B_\Delta \\ C_\Delta & 0 \end{bmatrix} = \begin{bmatrix} A + E\Delta K & B \\ C & 0 \end{bmatrix} \quad (17)$$

In the following, we will give the definition of robust stability.

Definition 1 (Robust Stability): The uncertain system (15) is robustly stable if $(I - M_{11}\Delta)^{-1}$ exists in $\mathcal{L}(L_2^h)$ and is causal for all $\Delta \in \Delta^c$.

The next lemma states a necessary condition for robust stability.

Lemma [8]: The uncertain system (15) is robustly stable if and only if there exists a $\Theta \in P_\Theta$ and $X > 0$ such that

$$A^T X + XA + K^T \Theta K + KE\Theta^{-1}E^T X < 0 \quad (18)$$

where

$$P_\Theta = \{\text{diag}(\theta_1 I_{h_1}, \dots, \theta_k I_{h_k}) : \theta_i > 0\} \quad (19)$$

is the positive commutant set corresponding to Δ^c .

We further introduce the generalized Gramians for the uncertain system from equation (15).

Definition: The matrices $S > 0$ and $P > 0$ are said to be generalized controllability or observability Gramians for the uncertain system (15) if the following inequalities hold:

$$\begin{aligned} \mathcal{A}_\Delta S + S \mathcal{A}_\Delta^T + \mathcal{B}_\Delta \mathcal{B}_\Delta^T &< 0 \quad \forall \Delta \in \Delta^c \\ \mathcal{A}_\Delta^T P + P \mathcal{A}_\Delta + \mathcal{C}_\Delta^T \mathcal{C}_\Delta &< 0 \quad \forall \Delta \in \Delta^c. \end{aligned} \quad (20)$$

As shown in [8], we can define the following algebraic Riccati inequalities for the uncertain system (15)

$$AS + SA^T + SK^T \Lambda_C K S + E \Lambda_C^{-1} E^T + BB^T < 0 \quad (21)$$

and

$$A^T P + PA + PE \Lambda_0^{-1} E^T P + K^T \Lambda_0 K + C^T C < 0 \quad (22)$$

with $S, P > 0$, $\Lambda_C^{-1}, \Lambda_0 > 0$ and $\Lambda_C, \Lambda_0 \in P_\Theta$.

Theorem: The following statements are equivalent assuming $K = E$:

- (i) The uncertain system (15) is robustly stable.
- (ii) The Riccati inequalities (21) and (22) admit a solution $S, P > 0$ for some $\Lambda_C, \Lambda_0 \in P_\Theta$.

Proof: We will prove the equivalence between (ii) and (i). We start from inequality (21) and we can easily show that inequality (21) holds with $X = S$, $\Lambda_C = \Theta^{-1}$ and $K = E$. The other inequality can be proven similarly as well as the equivalence between (i) and (ii).

Definition: An uncertain system of the form (15) is said to be balanced if it has generalized observability and controllability Gramians which are identical diagonal matrices.

The diagonal entries are called generalized Hankel singular values for the uncertain system.

We propose following the theoretical background in [8] a model reduction algorithm:

1. Solve the inequality system in (20) to obtain the generalized Gramians $S, P > 0$.

2. Balance S, P by choosing a state transformation matrix T such that

$$TST^T = (T^{-1})^TPT^{-1} = \text{diag}(\Sigma_1, \Sigma_2) = \text{diag}(\gamma_1, \dots, \gamma_n) \tag{23}$$

where $\gamma_1 \geq \dots \geq \gamma_d > \gamma_{d+1} \geq \dots \geq \gamma_n > 0$, $\Sigma_1 = \text{diag}(\gamma_1, \dots, \gamma_d)$ and $\Sigma_2 = \text{diag}(\gamma_{d+1}, \dots, \gamma_n)$.

3. Obtain the transformed nominal system as

$$M = \left(\begin{array}{c|cc} \bar{A} & \bar{E} & \bar{B} \\ \hline \bar{K} & 0_{h \times h} & 0_{h \times m} \\ \bar{C} & 0_{l \times h} & 0_{l \times m} \end{array} \right) \tag{24}$$

with $\bar{A} = TAT^{-1}$, $\bar{E} = TE$, $\bar{B} = TB$, $\bar{C} = CT^{-1}$ and $\bar{K} = KT^{-1}$.

The reduced order uncertain system of order d is defined as

$$M_r = \left(\begin{array}{c|cc} \bar{A}_r & \bar{E}_r & \bar{B}_r \\ \hline \bar{K}_r & 0_{h \times h} & 0_{h \times m} \\ \bar{C}_r & 0_{l \times h} & 0_{l \times m} \end{array} \right) \tag{25}$$

4. Represent the reduced dimension uncertain system as $\mathcal{G}_{r\Delta} = \mathcal{F}_u(M_r, \Delta)$, $\Delta \in \Delta^c$.

In the following, we will give a useful theorem without proof adapted from [8]:

Theorem: Consider a robustly stable uncertain system as given in (15) and suppose we can derive a reduced dimension uncertain system $\mathcal{G}_{r\Delta}$ based on generalized Gramians and state transformation. Then the system $\mathcal{G}_{r\Delta}$ is also balanced and robustly stable. We also have

$$\sup_{\delta \in [-1, 1]} \|\mathcal{G}_\Delta(s) - \mathcal{G}_{r\Delta}(s)\|_\infty \leq 2(\gamma_1^t + \dots + \gamma_q^t) \tag{26}$$

where γ_i^t denote the distinct generalized Hankel values of $\gamma_{d+1}, \dots, \gamma_n$.

Example: Consider the following uncertain system of the form (15) with $\Delta =$

$$\delta \in [-1, 1] \text{ and with } B = C = K = E = \text{diag}(1 \ 1) \text{ and } A = \begin{bmatrix} -9.7 & 0 & 0 \\ 1 & -1.7 & 0 \\ 0 & 1 & -2.7 \end{bmatrix}.$$

We choose $|\delta| = 0.3$. Based on the described balanced truncation procedure, we obtain the balanced Gramian $\Sigma = \text{diag}(0.33 \ 0.17 \ 0.05)$. A natural choice is to truncate the last state and keep the first two. The upper bound of the error is given according to (26) as $\sup_{\delta \in [-1, 1]} \|\mathcal{G}_\Delta(s) - \mathcal{G}_{r\Delta}(s)\|_\infty \leq 0.1$.

4 Conclusions

We present a model reduction of an uncertain gene regulatory network based on balanced truncation. The method is based on solving generalized Gramian inequalities and matrix transformations. We assume that structured uncertainty is norm bounded. When applied to linear systems, the reduced model corresponds to the usual balanced truncation of the system. A simple example is illustrating this novel approach of model reduction for gene regulatory networks.

References

- [1] R. Tanaka, H. Okano and H. Kimura (2006), Mathematical description of Gene Regulatory Units, *Biophysical Journal*, **vol. 91**, p. 1235-1247.
- [2] A. Meyer-Baese, C. Plant, S. Cappendijk and F. Theis (2010), Robust Stability Analysis of Multi-Time Scale Genetic Regulatory Networks under Parametric Uncertainties, *BIOCAMP 2010*, p. 854-863.
- [3] B. Moore (1981), Principal component analysis in linear systems: controllability, observability and model reduction, *IEEE Transactions on Automatic Control*, p. 17-32.
- [4] G. Lueders and K. Narendra (1972), Lyapunov functions for quadratic differential equations with applications to adaptive control, *IEEE Transactions on Automatic Control*, p. 798-801.
- [5] J. Scherpen (1993), Balancing for nonlinear systems, *Systems and Control Letters*, p. 143-153.
- [6] A. Meyer-Baese and F. Theis (2008), Gene Regulatory Networks Simplified by Nonlinear Balanced Truncation, *SPIE Symposium Computational Intelligence*, p. 6979C.
- [7] F. Ren and J. Cao (2008), Asymptotic and Robust Stability of Genetic regulatory Networks with Time-Varying Delays, *Neurocomputing*, **vol. 71**, p. 834-842.
- [8] L. Li and I. Petersen (2010) A Gramian-based Approach to Model Reduction for Uncertain Systems, *IEEE Transactions on Automatic Control*, vol. 55, p. 508-515.
- [9] L. Li and I. Petersen (2008) A Gramian-based Approach to Model Reduction for Uncertain Systems, *Proceedings of the 42nd IEEE Conference on Decision and Control*, p. 4373-4378.

Proposal of a Web Based Ambulance System in Saudi Arabia

H.Mirza¹, S.El-Masri²

¹College of Computer and Information Systems, King Saud University, Riyadh, Kingdom of Saudi Arabia

²College of Computer and Information Systems, King Saud University, Riyadh, Kingdom of Saudi Arabia

Abstract - *This paper proposes a new Ambulance System that automates all the processes of pre-hospital activities. Automation starts from allocating and dispatching the right ambulance, supporting the carried patient treatment by accessing the electronic health record, identifying the right hospital and communicating with the emergency department. The system has been developed by integrating several components. It has been found that the proposed web based ambulance system can provide a strong backbone for pre-hospital management process and gain competitive advantages. The system may reduce the response time and the human errors.*

Keywords: Ambulance, Dispatching, GPS, Emergency Department, Pre-hospital

1 Introduction

Pre-hospital patient treatment satisfaction has a huge impact on saving humans lives. Many studies wrote about the importance and possible advantages gained from reducing response time, and early specialized pre-hospital patient management. Moreover, it is realized that quick response time of pre-hospital patient management decrease the percentage of death and improves patient effect [5]. In a crowded area such as London, UK, it has been found that 49% of wounded people need 2 hours to reach a sufficient hospital care, 79% are victims of accidents in rural roads die in the accident place, and other 11% dies during their transportation to the hospital. And 8% of these accidents had chance of 50% to survive if enough pre-hospital management existed [10].

Saudi Arabia (SA) is a country in Middle East that has a large population, and thus crowded traffic in its main cities roads, about 26660857 populations was estimated in 2009 and, 27563432 in 2010. Riyadh as an example is the capital of SA and one of its main crowded cities. In 2009 it had about 42 of Ministry of Health Hospitals and 26 of private sector hospitals. Moreover, The Saudi Red Crescent Authority (SRCA)'s in 2009 had 274 first aid centers and 1097 ambulances [17]. Due to it's crowded roads many people become victims of car accidents and long response time. In Riyadh because of car accidents, in 2007 about 353 deaths

was reported, in 2008 about 357 and 266 deaths was reported in 2009 [18]. Thus, the critical role of pre-hospital treatment has a very important and critical impact on reducing the number of victims; their role could rescue many lives if it was fast and reliable.

Dealing with this amount of car accidents in such busy roads is a challenge for ambulances to arrive at the accident location in a short response time, and react to the patient with the best treatment. Therefore, this paper presents a solution to improve the management of ambulance dispatch and pre-hospital treatment with the fast response time and least amount of human mistakes for ambulance and hospital allocation process. The proposed Ambulance system automates the whole process (by using SOA, GPS and other technologies) from the time accident is reported by a caller to the time the patient is picked up and reaches the suitable hospital.

The paper is divided into four parts; first part includes a literate review of current systems that automates some of its processes, other research studies and human error and system failure. The second part describes the contribution which is a proposed Ambulance System, its components and how it functions. The third part includes the discussion and finally the conclusion and future work in last section.

2 Literature Review

Communication using computerized technology in many emergency medical systems (EMS) especially ambulance systems has been developed during the last two decades [4].

2.1 Current Systems

One of the first computerized systems was developed to manage the communications such as Computer Aided Dispatch (CAD) that was used in 1995 in Victoria-Australia. In 1998, this system was enhanced and introduced with a Medical Priority Consultant's Advanced Medical Priority Dispatch. It was one of the best emergency systems that provided to the hospital clinical information of the patients and included an automatic vehicle location system (AVL) in their ambulance [4].

Moreover, in 1998 due to the lack of communication between agencies, a project was developed to gain communication interoperability network by using different technologies. It is called Silicon Valley Regional Interoperability Project (SVRIP), that respond to emergency incidents with the nearest and most appropriate emergency response resource [6].

In 2007 CADIP (Computer Aided Dispatch Interoperability Project) was launched by the department of Homeland Security's Office for Interoperability and Compatibility (OIC). CADIP is created to solve the concerns of difficulty that occurs when an emergency response agency is trying to respond to multi-jurisdictional emergencies that are not linked to them. This in term happen when the time-consuming phone calls that is usually done to link such emergencies (incident) to the nearest resource is eliminated, and replaced by automatic dispatchers [5].

Recent emergency ambulance systems EAS have appeared with a good impact on health sector. The Victorian Ambulance Cardiac Arrest Registry (VACAR) is a leading system in cardiac arrest CA registries. This system has two parts of ambulance services. The first service is Metropolitan Ambulance Service (MAS), which uses a computerized, protocol-based dispatch system. And the second system is the Rural Ambulance Victoria (RAV) that uses a manual call talking and dispatch process [1]. Another area in Australia - north Victoria- employs a pre-hospital service known as Ambulance Service of New South Wales (ASNSW). This system dispatches its ambulances by CAD and provides them with Mobile Data Terminals (MDT) for messaging and Automatic Vehicle Location (AVL) for keeping track of their location. Moreover, for prioritization of dispatching tasks, a Medical Priority Dispatch System (MPDS) has been presented [2].

In 1977 Emergency Ambulance Services (EAS) was developed without the use of computerized system. Its emergency department doctors provided ad hoc advisory services, without a formal medical control. In 1989 EAS was attached to Singapore Civil Defense Force (SCDF) and stuffed with more specialized crew, but still without electronic communication technology. Singapore EAS continued developing until the pilot project HEAL (Hospital and Emergency Ambulance Link) was launched to improve data collection and communication. HEAL presents a wireless information technology system to support existing voice links between the ambulance crew and the emergency department (ED). This system include a touch screen with easy data entry, mobile computers to automatically capture vital signs and other medical data, then send them to the target hospital via warless communication network. Where these collected data creates an electronic pre-hospital record for the patient. And it uses a user-friendly client server application. HEAL is composed of: 1) Advanced patient details model; that capture

patients medical data and send it to ED, 2) Ambulance incident management module; that save and store all received records from the ambulances, 3) Drug request and authorization model; this supports the paramedics by physicians approved drugs, 4) Text communication module; this is responsible for message exchange between ambulance crew and ED staff. This system had a huge impact on the pre-hospital system quality; such as reducing the waiting time for critical care patients to be seen at the emergency department (ED) from 35 to 17 min. Also, the time spent by paramedics in the ED after handing over the patient to the ED staff was decreased from 15 to 8 min. therefore, HEAL showed the great possibility of electronic communication and data collection in the pre-hospital environment [3].

In Amsterdam EMS was employed to manage the pre-hospital care delivery, known as the Dutch EMS that depends on phone calls. This system is concerned with the dispatching and treatment level, and characterized as a nurse driven triage system. A dispatch center is responsible for receiving emergency phone calls on "112" phone number and redirect them to fire and police departments. This dispatch center doesn't have an automatically EMS. It updates the beds information by a computerized updates from the participating hospitals. The phone calls prioritizing are done by the dispatch nurse that dispatches the ambulance and present pre-arrival instructions. And the ambulance average arrival time to the incident is 8 min [6].

An ambulance with highly implemented design project is delivered as a partnership between the university of Texas Health Science Center at Houston, and Texas A&M University System and the U.S. Army Medical Research and Materiel Command, known as the Disaster Relief and Emergency Medical Service (DREAMSTM) project. This system use wireless internet access in its ambulances to gain a wide cellular coverage and transfer audio (voice), text communication, video, and vital signs, with a high quality transmission. It also, can create a real time communication between the moving ambulance emergency medical technicians and emergency room physicians in the hospital to share patient information. Transforming these several types of data can be achieved by wireless internet access that transfers data from the ambulance to the internet then to the emergency room. The ambulance communication system contains several third generation (3G) wireless cards system from different cellular service providers. This improves its ability to benefit from these distinct service providers by a wider coverage, and different technologies. It also, has the advantage of keeping the highest priority patient information inside the ambulance in the case of loosing communication [7].

Another system in managing pre-hospital health emergencies sector was developed by the Regional Health Information Network (RHIN) of Crete across the island, it's

known as the integrated pre-hospital emergency management system (PEMS). It provides several functionalities: ambulance tracking and route guidance, optimal resource management, management of emergency records, and real time patient multimedia data transmission and visualization. The ambulance tracking and route guidance functionality is achieved by a geographic information system (GIS), where each ambulance type, description, and current status (available, occupied, est.) are identified. These services were provided by adopting the GPS system, position monitoring technologies and intelligent route guidance [8].

LIFENET system that lay's in the hospital gives the EMS team ability to send data from their LIFEPAK (ambulance device) directly to the hospital. At the hospital, LIFENET receives alert and displays the patient data before he arrives. This gives the ability to the hospital to re-route the patient to another hospital if needed due to the received data, without an unnecessary stop at this hospital. Data transfer is done through wireless link via the internet. Clearly the system reduces time and improves efficiency. Moreover, it has many other capabilities such as storing information for a long time, etc. [12].

2.2 Novel Studies

New studies are competing and progressing toward advanced wireless communication in the health care sector. A novel study has been made at TEXAS AT DALLAS University, where a software system was developed. This is an online based software system that is accessed via internet connection. It interacts with several actors: accident caller for '911', an ambulance dispatcher and an emergency room. It is composed into five subsystems: 1) user interface subsystem, 2) main subsystem that is responsible for interactions between the UI and hospital subsystem, 3) hospital subsystem, 4) emergency subsystem, and 5) ambulance subsystem. The Ambulance Subsystem is in charge of any communications concerning ambulances. This includes ambulance List class and ambulance classes/objects which are responsible for using the specified methods in the related database class to insert, read, update and delete any related data. Each ambulance includes the software, and a set of hardware such as location tracking hardware, ambulance laptop, ambulance communication device (such as radio). The wireless communication between the dispatcher computers and the Ambulance Dispatch System (ADS) server is done via HTTP/HTTPS protocols by the web browser in their computers. Moreover, the ambulance uses Wi-Fi internet to communicate with ADS server. The ADS server includes a GPS mapping software, and communicates with Emergency Department Server (EDS) [9].

Another wireless acquisition system for ambulances is presented by (J. Liao, 2009 and others). This paper suggests equipment for the ambulance vehicle by wireless technologies to support pre-hospital treatment. The system has a wireless biomedical sensor network for collecting patient's physiological data. It also has multimedia interactive information, and a wireless transmitting backbone sensor network. The system can not only send text medical data in real time to the hospital, but it also presents a multimedia interactive communication including audio and video information [10].

Usually one of the main uses of GPS in ambulance vehicles is the Automatic Vehicle Location (AVL) technology to keep track of it. A newer AVL system has emerged; it's now integrated with an emergency vehicles display computer, or data terminal, records management and video systems. The data terminal's job is to state the ambulance engine status in real time. This can support dispatch centers with helpful updated information of their resources condition [11].

2.3 Human error and system failure

Information System (IS) failure can happen due to human development error or/and use error. A clear example of IS failure in ambulance system due to human error is the London Ambulance Service Computer Aided Dispatch (LASCAD) system. The system was put in use in October 1992, and failed after 2 weeks only. Almost 20-30 people have lost their lives due to LASCAD failure. Many argue on the causes of its failure; where some referred it to software engineer's inaccurate work on following the development methodology, lack of the system user's practice or any other reasons. The study of (P.Beynon-Davies, 1999) has identified two failure causes for LASCAD system; 1) error of use, 2) criticality. Error of use happened due to the weak relationship and communication between the managers and the workforce, and the resistance of the workforce to learn a new system. Criticality was another cause of failure via visibility issues. As the LAS problems reporting and the system safety nature were not sufficiently discussed [13][14].

(M.Hougham, 1996) have stated LASCAD system failure due to several reasons related to inaccurate information. These reasons are: 1) weak coverage of radio signals blocked the system from gaining all the data, and radio communication bottlenecks because of busy periods or when crew change shift and log in by their ambulance MDT, 2) ambulance crew mistaken pressing the right status button, due to the pressure of certain incidents and frustration feelings they get because of poor training, 3) insufficient personnel taking calls and lack interaction between operators and the system and its different parts, 4) and a single technical programming error related to full server memory space [14].

3 Proposed Ambulance System

According to earlier studies, Emergency Ambulance Systems are very critical that can save human's life or put it in danger. Where new technologies have emerged, still others are working to develop better systems. The proposed ambulance system is a wireless web based integrated system. It is designed to satisfy the highest effectiveness, efficiency, least response time, reduction of error and best quality decisions for resource allocation. The system is designed to eliminate any kind of human error during an emergency incident, by computerizing all the actions and functionalities and avoiding the human intervention as much as possible.

3.1 System components

The proposed Ambulance system is part of the "Comprehensive Medical Emergency System CME" [16] [15]. CME system is composed of five integrated subsystems; 1) Mobile device system, 2) Main Central System MCS (Ambulance dispatching system), 3) Ambulance system, 4) Online Health Record (OHR) system, 5) Hospital Emergency Department system (HEDS) [16], (see figure 1).

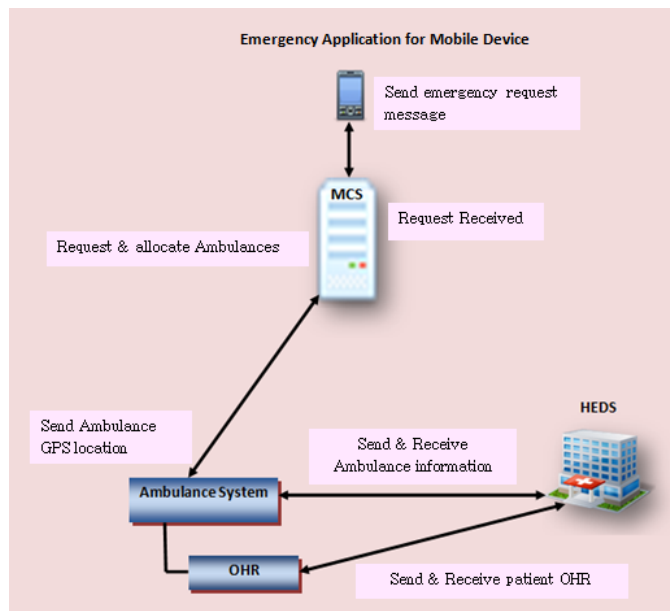


Fig 1 CME system components.

Each one of the CMS components has the following job:

Mobile device system: A mobile device system is an application that gives ability to an accident reporter in the accident area to notify the MCS with caller information (coordination) and, accident information (as the number of wounded people) [16].

MCS: After accident information has been captured in the MCS, the information will be processed, several interactions with an ambulance server that gets the ambulances GPS coordinates. Then allocation of the most suitable ambulance for the accident is based on several issues such as availability, nearest and shortest time to arrive (which is based on the navigation system map not the direct distance) [16].

OHR: The OHR is created to transfer patient's data to the hospital in real time (immediately) and using online (wireless) communication. The OHR system is a decision support system that can support the ambulance crew with the most proper treatment for the patient due to its electronic health record information and new accident information entered by the crew [15][16].

HEDS: The HEDS in each hospital has the ability to receive patient information entered by the coming ambulance crew. Moreover, the roadmap of the coming ambulance is automatically displayed on the HEDS screen [16].

Ambulance System: The proposed Ambulance System main goal is to gain reliable satisfied patient transformation from the accident to the proper hospital with minimal risk and without human intervention. Each ambulance vehicle must include several components; these are:

1. GPS navigation device installed in each vehicle.
2. MDT for messaging communication with the ED.
3. A unique website URL stored at its ambulance server (ambulanceID.AmulanceServer.com), several information is stored in this page:
 - Ambulance ID.
 - Ambulance status (on mission, available, broken)
 - Real time location (road map determined by GPS).
 - Time to arrive to accident (null)
 - Accident location (coordination)
 - Accident road map
 - Hospital name (null)
 - Hospital location (coordination)
 - Hospital road map
 - Time to arrive to hospital (null)
 - Time to complete mission (null).
 - Time.
 - Date.
4. A laptop that grants the crew access to the ambulance system (that communicates with MCS, OHR, and HEDS) via internet WIMAX wireless connection (using web browser).

3.2 System design

The Ambulance System is designed to communicate with each one of the CES components (accept Mobile Application) to satisfy its goal.

- 1.The MCS starts looking for the suitable ambulance car (available, nearest) by contacting the ambulance server to retrieve available cars locations.
- 2.MCS then matches the accident location with an appropriate ambulance.
- 3.MCS then sends a request message to the allocated ambulance car, with the accident location.
- 4.If the car accepts the mission, its status changes from available to on mission.
- 5.After, the ambulance reaches the accident and carry the patient, the ambulance crew starts entering the patient injury information into the OHR system.
- 6.This info will be sent to the Ambulance server with accident location.
- 7.Ambulance server then decides the appropriate hospital and sends the resulted hospital ID to the ambulance car automatically.
- 8.The ambulance car then, contacts the hospital by its ambulance system, (see figure 2).

3.3 System process

The communication process is divided into two phases:

3.3.1 Phase1: Ambulance versus MCS

The ambulance server and MCS is able to keep track of any ambulance vehicle via its GPS system.

After the accident information is captured by MCS, it will start searching for a matching ambulance to take the mission. This is done when the MCS interact with the ambulance server that

interacts directly with ambulance vehicle and retrieves its information.

The MCS triggers the ambulance server to retrieve all the available ambulance cars in the area. Each ambulance must have: one of the three vehicle statues; Available, non available, on mission. And real time coordination sited by its GPS.

When the MCS server finds the available nearest ambulance and defines its ID, it sends a request to the defined ambulance system.

The ambulance crew is then able to respond to the MCS call by the installed laptop touch screen, and choose accept, to confirm the mission acceptance (or reject otherwise). When it accepts the mission the vehicle status will be reset from available to on mission. But if the crew selects reject mission (for any reason); the MCS will start looking for another available nearest ambulance again. A road map to the incident location will be displayed on the ambulance system screen.

When the patient is inside the ambulance vehicle, the crew starts entering the patient injury information to the OHR system. This information and the accident location will be sent to the ambulance server.

3.3.2 Phase 2: Ambulance versus HEDS

The Ambulance Server already has the locations of all hospitals in all areas, and retrieved the accident location from the ambulance location (at the accident location).

The Ambulance server then matches the proper hospital due to three values:

- 1.Location (nearest to accident).
- 2.Specialty (according to the patient injury type).
- 3.Then it contacts the chosen hospital HEDS to check for bed availability to confirm the choice.

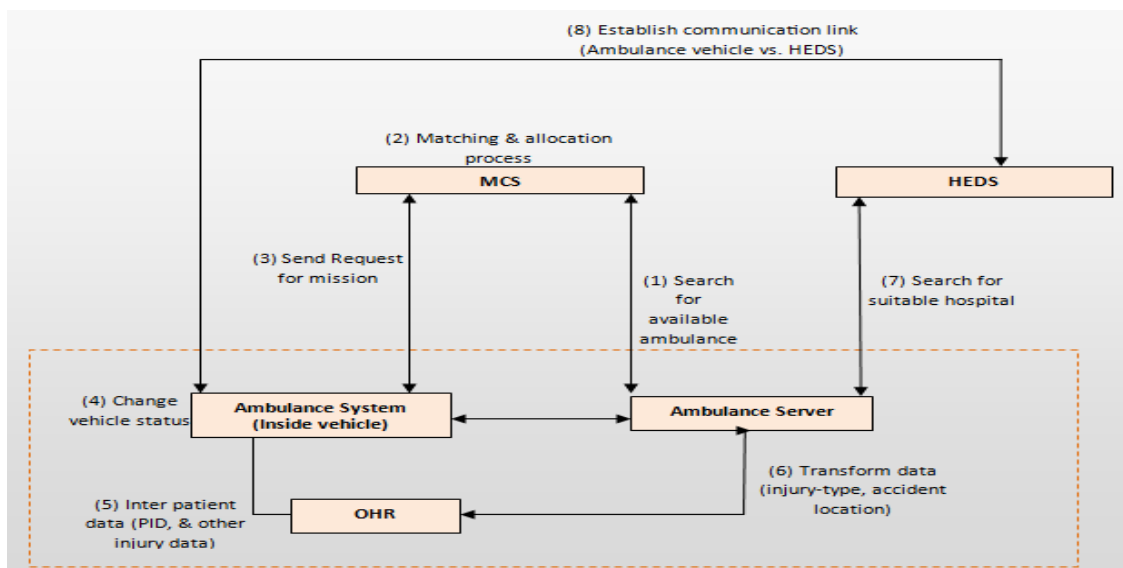


Fig 2 The proposed Ambulance System components interaction

Once the Ambulance car is allocated to the proper hospital, automatically a road map of that hospital will be shown on the ambulance system screen and contact it's HEDS by:

- Ambulance identification (ID) number will be displayed on the HEDS screen.
- The ambulance road map will be displayed on the HEDS screen.
- The ambulance arrival time is displayed on the HEDS screen with either one of three colors according to the Time Left to Arrive (TLA) to the hospital; as the ambulance alert colors:
 - Blue: 1 h => TLA > 30 min.
 - Orange: 30 min => TLA > 15 min.
 - Red: TLA <=15 min.
- Communication link is established between the ambulance and the chosen HEDS, to allow data, voice and text transformation.

Moreover, the OHR Allow the HEDS team to access online updated information for the coming patient during his delivery to the hospital. This is done automatically when the ambulance crew uses the OHR to transform patient identification number (PID) and injury information after the accident and any changes in his situation during the delivery.

It also, supports the ambulance crew with best decisions to handle patient injury, based on the system ability to get feedback from his medical record and make decisions upon it [16]. If the patient didn't have an OHR, then the ambulance crew will only send any patient identification (e.g. name and SSN, etc) and immediate injury information to the OHR (that will register him as new patient in the OHR automatically), and communicate with the HEDS for any support needed.

Furthermore, if there is no OHR established yet; the system can still function. So, the ambulance crew uses the Ambulance System to send any patient identification and immediate injury information, manually entered to transfer it to the HEDS.

After, few minutes (e.g. 4 min) from the ambulance arrival to the hospital location, automatically the ambulance status will be moved from on mission to available.

4 Discussion

In previous researches, many current emergency systems and novel studies are automating some of their functionalities and using wireless links in their communication. New researches are competing to develop advanced automated emergency systems due to its massive advantages. The proposed system has many features and advantages that can improve the processes performance and management of pre-

hospital tasks. Thus, the Ambulance system will increase the percentage of rescuing human's lives. These features and advantages can be summarized as:

- Decreasing human error, so then increasing system efficiency. This is achieved by automating the system functionalities and avoidance of human interaction.
- Reduces response time. This in term increase percentage of rescuing humans' life, by decreasing the time to complete the rescue mission. This is achieved from the system automated capability to decrease the time of allocating a proper ambulance car (nearest and available), and a proper hospital (nearest, specialty, and available bed).
- Lower cost and human effort. This is achieved on the long term after establishing the system. There will be less human effort needed to dispatch ambulances.
- Best resource allocation decisions are made. This is achieved by the system computation capability to decide the nearest and appropriate ambulance care and hospital for the mission.
- Best pre-hospital treatment decisions. This is achieved from the system capability to support the ambulance crew with the suitable treatment, based on the medical history (from OHR) and current situation.

Moreover, in the case of sudden change or degradation in patient situation during his transformation, the HEDS team will be aware of any such change; and thus be appropriately prepared.

5 Conclusion

In this paper, a new fully computerized ambulance system has been proposed. It has been designed to facilitate the dispatching, and through the GPS, identifying the nearest ambulance to the accident. All components of the system have been explained and advantages of the proposed system over current and existing systems have been shown. By reducing the response time and the communication human errors, and by accessing the electronic health record and communicating information to hospital, this may make the proposed system one of the most efficient systems comparing to others.

The future work will be focused on completing the development and evaluating the system in real life.

Acknowledgment

This work is part of a two year research project which has been fully funded by a grant through King Abdul-Aziz City for Science and Technology (KACST) / National Plan for Science

and Technology (NPST) in the Kingdom of Saudi Arabia. Grant number: 09-INF880-02.

6 References

- [1] Fridman, M, et al. (2007). A model of survival following pre-hospital cardiac arrest based on the Victorian Ambulance Cardiac Arrest Register. *Resuscitation*. 75 (2), 311-322.
- [2] Trevithick, S, Flabouris, A, Tall, G, and Webber, C.F. (2003). International EMS Systems: New South Wales, Australia. *Resuscitation*. 59 (2), 165-170.
- [3] Lateef, F. (2006). INTERNATIONAL EMS SYSTEMS The emergency medical services in Singapore. *Resuscitation*. 68 (3), 323-328.
- [4] El-Masri, S. (2005). MOBILE COMPREHENSIVE EMERGENCY SYSTEM USING MOBILE WEB SERVICES. *The Second International Conference on Innovations in Information Technology (IIT'05)*.
- [5] Computer-Aided Dispatch interpretability project Documentation of Regional Efforts, 2008, Homeland Security. <http://www.dhs.gov/index.shtm>. Last Accessed 9th March 2011.
- [6] Joe, E, et al. (2006). ANALYSIS AND APPLICABILITY OF THE DUTCH EMS SYSTEM INTO COUNTRIES DEVELOPING EMS SYSTEMS. *The Journal of Emergency Medicine*. 30 (1), 111–115.
- [7] Sahai, G. Goulart, A. and Wei Zhan Arnold, R. (2008). Optimal selection of wireless channels for real-time communication in ambulances. *IEEE Explore*.
- [8] Tsiknakis, M. Spanakis, M. (2010). Adoption of innovative eHealth services in prehospital emergency management: a case study. *Information Technology and Applications in Biomedicine (ITAB), 10th IEEE International Conference*.
- [9] Saracho, A. et al.. (2007). *Ambulance Dispatch System*. Available: http://www.google.co.uk/url?sa=t&source=web&cd=4&ved=0CDUQFjAD&url=http%3A%2F%2Fwww.utdallas.edu%2F~chung%2FCS6354%2FCS6354_U07_source%2FFantastic9%2FDeliverable3-SDD-Fantastic9.doc&rct=j&q=CS%206354%20. Last accessed 15th April 2011.
- [10] Liu, W. et al. (2009). Wireless Acquisition System for the Real Ambulance Field. : *Robotics and Biomimetics, 2008. ROBIO 2008. IEEE International Conference*.
- [11] U.S. Department of Homeland Security. (summer 2009). *Interoperability technology today a resource for the emergency response community*. Available: <http://www.safecomprogram.gov/NR/rdonlyres/BF225BD5-1541-41EAABAE2482E6932F5A/0/Summer2009InteroperabilityTechnologyToday.pdf>. Last accessed 20th April 2011.
- [12] Physio-Control. (2009). *LIFENET SYSTEM*. Available: http://www.physio-control.com/uploadedFiles/products/data-management/LIFENET_Brochure_3302845_B.pdf?n=7022. Last accessed 12th April 2011.
- [13] Beynon-Davies, P. (1999). Human error and information systems failure: the case of the London ambulance service computer-aided despatch system project. *Interacting with Computers*. 11 (6), 699–720.
- [14] Hougham, M. (1996). London Ambulance Service computer-aided despatch system. *International Journal of Project Management*. 14 (2), 103-110.
- [15] El-Masri S., “Mobile Comprehensive Emergency System using Mobile Web Services”. A book chapter, *Handbook of Research on Mobile Business: Technical, Methodological and Social perspective*, edited by B. Unhelkar. Idea Group, 1 (2005) 106-112
- [16] EL-MASRI, S. SADDIK, B. (2011). “Mobile Emergency System and Integration” 12th International Conference on Mobile Data Management 6-9 June, 2011, Luleå, Sweden
- [17] Central Department of Statistics and Information website. Available at <http://www.cdsi.gov.sa/socandpub/health>. Last Accessed 10th May 2011.
- [18] Riyadh Traffic website. Available at <http://www.rt.gov.sa/statistics.php?year=1430>. Last Accessed 10th May 2011.

BIOCAMP'11

Stability Analysis of Hybrid Stochastic Gene Regulatory Networks

Anke Meyer-Baese, Claudia Plant^a,
Jan Krumsiek, Fabian Theis^b, Marc R. Emmett^c and
Charles A. Conrad^d

^a*Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, U.S.*

^b*Institute of Bioinformatics and Systems Biology, Helmholtz Center Munich, Ingolstdter Landstrae 1 85764 Neuherberg, Germany*

^c*Ion Cyclotron Resonance Program, National High Magnetic Field Laboratory, Florida State University, Tallahassee, FL 32310-4005, and Department of Chemistry and Biochemistry, Florida State University, Tallahassee, FL 32306, U.S.*

^d*The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, U.S.*

Abstract

Gene regulatory networks (GRNs) represent complex nonlinear coupled dynamical systems that models gene functions and regulations at the system level. Previous research has described GRNs as coupled nonlinear systems under parametric perturbations without considering the important aspect of stochasticity. However, a realistic model of a GRN is that of a hybrid stochastic retarded system that represents a complex nonlinear dynamical system including time-delays and Markovian jumping as well as noise fluctuations. In this paper, we interpret GRNs as hybrid stochastic retarded systems and prove their asymptotical stability. The theoretical results are elucidated in an illustrative example and thus shown how they can be applied to reverse engineering design.

Key words: Genetic regulatory network, retarded systems, Markov chain, stochastic systems, p -th moment stability

Email address: ameyerbaese@fsu.edu (Anke Meyer-Baese, Claudia Plant).

1 Introduction

Gene regulatory combining a coupled dynamics of fast and slow states constitute an important class of biological networks [1]. Synthetic design of such networks is very sensitive to parameter perturbations. Errors in parameters such as external perturbations and modeling errors are caused by data inaccuracies or computation errors. These perturbations can lead to location errors of equilibria, to instabilities, and even to spurious states. Therefore, a rigorous understanding of the qualitative robustness properties of gene regulatory networks with respect to parameter variations on both a fast or slow time scale and under consideration of a transcriptional time delay [4] became imperative. In [2], the gene regulatory networks are formulated as coupled nonlinear differential systems operating at different time-scales under vanishing perturbations and time delays. In [3], gene regulatory networks are described as either two-time scale systems without delay [3] or as unperturbed systems [5].

It is well-known that molecules and reaction rates are subject to significant statistical fluctuations and especially gene regulation is an intrinsically noisy process due to intracellular and extracellular noise perturbations and environmental fluctuations. Additionally, the transition from one state to the next is based on certain transition probabilities forming a homogeneous Markov chain with finite state space. This aspect motivates the formulation of a stochastic model with Markovian switches to describe the dynamics of gene regulation. Previous work investigated genetic regulatory networks with parameter uncertainties and noise perturbations [7] or of Markov-type with delays and uncertain mode transition rates [8]. It is naturally to propose a more detailed model with delays that combines Markovian jumping and noise perturbations and analyze its dynamic behavior. In this paper, we analyze the robustness properties of gene regulatory networks, modeled by a system of competitive differential equations, from a rigorous analytic standpoint [6]. The network under study models the delayed nonlinear dynamics under consideration of Markovian jumping and noise perturbations.

2 Problem Statement

Gene regulatory networks represent circuits of genes that interact and regulate the expression of other genes by proteins. The change in expression of a gene is regulated by protein synthesis in transcriptional, translational and post-translational processes. Taking into account a transcriptional time delay [4] and the fact that mRNA typically decays much faster than the protein, we considered in a previous work [2] the gene regulatory network described by the following equation

$$\begin{aligned} \dot{M}_i(t) &= -a_i M_i(t) + \sum_{j=1}^n \tilde{w}_{ij} \tilde{g}_j(P_j(t - \rho)) + B_i \\ \dot{P}(t) &= -c_i P_i(t) + d_i M_i(t) \end{aligned} \tag{1a}$$

where $M_i(t), P_i(t) \in R$ are the concentrations of mRNA and protein of the i th node, respectively. The parameters a_i and c_i are the decay rates of mRNA and protein, respectively; d_i is the translation rate, $\tilde{g}_j(x) = \frac{\left(\frac{x}{\beta_j}\right)^{H_j}}{\left(1 + \left(\frac{x}{\beta_j}\right)^{H_j}\right)}$, $B_i = \sum_{j \in I_i} b_{ij}$ and I_i is the set of all the j which is a repressor of gene i , $\tilde{W} = (\tilde{w}_{ij}) \in R^{n \times n}$ is defined as follows

$$\tilde{w}_{ij} = \begin{cases} b_{ij}, & \text{if transcription factor } j \text{ is an activator of gene } i \\ 0, & \text{if there is no link from node } j \text{ to } i \\ -b_{ij} & \text{if transcription factor } j \text{ is a repressor of gene } i \end{cases} \tag{2}$$

Let $(M^{*T}, P^{*T})^T$ be an equilibrium point of the system (1a). By shifting the equilibrium of the system to the origin, we obtain a general formulation of the GRN as a nonlinear coupled system with both time-varying delays for feedback regulation $\rho_i(t)$ and translation $\sigma_i(t)$:

$$\begin{aligned} \dot{M}_i(t) &= -a_i M_i(t) + f_j(P_1(t - \rho_1(t)), \dots, P_n(t - \rho_n(t))) \\ \dot{P}(t) &= -c_i P_i(t) + d_i M_i(t - \sigma_i(t)) \end{aligned} \tag{3a}$$

We thus obtain $g_i(p_i(t)) = \tilde{g}_i(P_i(t) + P_i^*) - \tilde{g}_i(P_i^*)$. Because \tilde{g}_i is a monotonically increasing function with saturation, $g_i(\cdot)$ satisfies the sector condition $0 \leq \frac{g_i(x)}{x} \leq k_i$.

In terms of Hill function we obtain

$$\begin{aligned} \dot{M}_i(t) &= -a_i M_i(t) + \sum_{j=1}^n \tilde{w}_{ij} g_j(P_j(t - \rho_j(t))) \\ \dot{P}(t) &= -c_i P_i(t) + d_i M_i(t - \sigma_i(t)) \end{aligned} \tag{4a}$$

Further for simplicity, we will assume that all feedback regulation and translational delays are equal $\rho_1 = \dots, \rho_n = \rho$ and $\sigma_1 = \dots, \sigma_n = \sigma$. The above model can be formulated as a n -dimensional GRN

$$\begin{aligned}\dot{\mathbf{M}}(t) &= \mathbf{A}\mathbf{M}(t) + \mathbf{W}\mathbf{g}(\mathbf{p}(t - \rho(t))) \\ \dot{\mathbf{P}}(t) &= -\mathbf{C}\mathbf{P}(t) + \mathbf{D}\mathbf{M}(t - \sigma(t))\end{aligned}\quad (5a)$$

with $A = \text{diag}\{a_1, a_2, \dots, a_n\}$, $C = \text{diag}\{c_1, c_2, \dots, c_n\}$ and $D = \text{diag}\{d_1, d_2, \dots, d_n\}$.

3 Theoretical Concepts of Hybrid Stochastic Retarded Systems

In the following, we will introduce some notations and theoretical concepts from stochastic functional differential equation theory we will be using throughout this paper.

Notations:

$(\Sigma, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, P)$: complete probability space with a filtration $\{\mathcal{F}\}_{t \geq 0}$ that is right-continuous and \mathcal{F}_0 contains the P -null sets.

$B(t) = (B_1(t), \dots, B_m(t))^T$: m -dimensional Brownian motion defined on the probability space.

$|\cdot|$ is the Euclidean norm in R^n .

$\mathcal{C}([-\tau, 0]; R^n)$ with $\tau \geq 0$ denotes the family of all continuous R^n -valued functions ψ on $[-\tau, 0]$ with the norm $\|\psi\| = \sup\{|\psi(\theta)| : -\tau \leq \theta \leq 0\}$.

$\mathcal{C}_{\mathcal{F}_0}^b([-\tau, 0]; R^n)$ is the family of all \mathcal{F}_0 -measurable bounded $\mathcal{C}([-\tau, 0]; R^n)$ -valued random variables $\zeta = \{\zeta(\theta) : -\tau \leq \theta \leq 0\}$.

Let $r(t), t \geq 0$, be a right-continuous Markov chain on the probability space taking values in a finite space $S = \{1, 2, \dots, N\}$ with generator $\Gamma = (\gamma_{ij})_{N \times N}$ given by

$$P\{r(t + \Delta) = j : r(t) = i\} = \begin{cases} \gamma_{ij}\Delta + o(\Delta), & \text{if } i \neq j, \\ 1 + \gamma_{ij}\Delta + o(\Delta) & \text{if } i = j, \end{cases} \quad (6)$$

where $\Delta > 0$ and $\gamma_{ij} \geq 0$ is the transition rate from i to j if $i \neq j$ while $\gamma_{ij} = -\sum_{i \neq j} \gamma_{ij}$.

We also assume that the Markov chain $r(\cdot)$ is independent of the Brownian motion $B(\cdot)$. The sample pathes of $r(t)$ are right-continuous step functions with a finite number of simple jumps in any finite subinterval of $R_+ := [0, \infty)$.

In the following we describe a hybrid stochastic retarded system (HRRS) driven by continuous-time Markovian chains [6] used in stochastic modeling. Let such a n -dimensional HRRS be given as

$$dx(t) = f(x_t, t, r(t))dt + g(x_t, t, r(t))dB(t) \tag{7}$$

on $t \geq 0$ with initial data $x_0 = \{x(\theta) : -r \leq \theta \leq 0\} = \theta \in \mathcal{C}_{\mathcal{F}_0}^b([\tau, 0]; R^n)$ and with

$$\begin{cases} f & : & \mathcal{C}([-\tau, 0]; R^n) \times R_+ \times S \rightarrow R^n \\ g & : & \mathcal{C}([-\tau, 0]; R^n) \times R_+ \times S \rightarrow R^{n \times m} \end{cases}$$

with measurable functions with $f(0, t, i) = 0$ and $g(0, t, i) = 0$ for all $t \geq 0$. Thus, (7) has a trivial solution $x(t; 0) = 0$. $x_t = \{x(t + \theta) : -r \leq \theta \leq 0\}$ represents a $\mathcal{C}([-r, 0]; R^n)$ -valued stochastic process. We also assume that g, h are smooth functions such that (7) has only continuous solutions $x(t; \zeta)$ on $t \geq 0$.

Further $C^{2,1}(R^n \times R_+ \times S)$ is the family of all nonnegative functions $V(x, t, i)$ on $R^n \times R_+ \times S$ being twice continuously differentiable in x and once in t . With $V \in C^{2,1}(R^n \times R_+ \times S; R_+)$, we define an operator \mathcal{C} , from $\mathcal{C}([-\tau, 0]; R^n) \times R_+ \times S \rightarrow R$ by

$$\mathcal{L}V(x_t, t, i) = V_t(x, t, i) + V_x(x, t, i)f(x_t, t, i) \tag{8}$$

$$+ \frac{1}{2} \text{trace}[g^T(x_t, t, i)V_{xx}(x, t, i)g(x_t, t, i)] \tag{9}$$

$$+ \sum_{j=1}^N \gamma_{ij}V(x, t, j) \tag{10}$$

where

$$\begin{cases} V_t(x, t, i) & : = \frac{\partial V(x, t, i)}{\partial t} \\ V_x(x, t, i) & : = \left(\frac{\partial V(x, t, i)}{\partial x_1}, \dots, \frac{\partial V(x, t, i)}{\partial x_n} \right) \\ V_{xx}(x, t, i) & : = \left(\frac{\partial^2 V(x, t, i)}{\partial x_i \partial x_j} \right)_{n \times n} \end{cases}$$

We give a useful definition regarding the stability of HRRS.

Definition The trivial solution of (7) is

- a.) p th ($p > 0$) moment stable if, for every $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that

$$E|x(t; \zeta)|^p \leq \epsilon, \quad \forall t \geq 0$$

whenever $\|\zeta\|^p < \delta_0$;

- b.) globally p th moment asymptotically stable if it is p th moment stable and, moreover, for all $\zeta \in \mathcal{C}_{\mathcal{F}_0}^b([-\tau, 0]; R^n)$,

$$\lim_{t \rightarrow \infty} E|x(t; \zeta)|^p = 0 \quad (11)$$

As shown in [6], we immediately see that (global) p th moment (asymptotic) stability implies (global) stochastic (asymptotic) stability.

4 Asymptotic Stability Analysis of Hybrid Stochastic Retarded GRNs

The objective of this study is to discuss the stability properties of the hybrid stochastic retarded GRN. The analysis is based on a mathematical model and a rigorous analytic standpoint.

Based on the above model described by equation (4), we will now introduce noise perturbations and Markovian jumping parameters. As previously discussed, the parameters of the GRN may change randomly at discrete time instances or in other words, the GRN has finite modes and it can switch from one to another at different times determined by a Markov chain. Since the switching probabilities are not a priori known, the GRN can be modeled by a hybrid system. The system of the GRN has both continuous and discrete states which are described by a Markovian jumping system. We can rewrite the GRN from equation (5a) as a hybrid stochastic retarded system

$$dX(t) = F(X(t), t, r(t))dt + G(X(t), X(t - \rho(t)), X(t - \sigma(t)), t, r(t))dB(t) \tag{12}$$

with $X(t) = [M(t), P(t)]$, $Y(t - \rho(t)) = X(t - \rho(t)) = [0, \dots, 0, P(t - \rho(t))]$ and $Z(t - \sigma(t)) = X(t - \sigma(t)) = [M(t - \sigma(t)), 0, \dots, 0]$. X, Y and Z are $2n \times 1$ vectors. $G(X(t), X(t - \rho(t)), X(t - \sigma(t)), t, r(t))$ is the noise intensity function. The Markov chain $r(\cdot)$ is given as in (6) and $B(\cdot)$ is the Brownian motion.

We will make the following assumptions for computational simplicity without loss of generality.

Assumptions:

(A1) Let us assume that $\sum_{j=1}^N \gamma_{ij} = 1$.

(A2) The trace can be approximated as $\text{trace}[G^T(x, y, z, t, i) \cdot G(x, y, z, t, i)] \leq \rho_1|x(t)|^2 + \rho_2|y(t)|^2 + \rho_3|z(t)|^2$.

(A3) For the nonlinear term, we assume $F(x, y, z, t, i) \leq -\rho_4|x(t)|^2 + \rho_5|y(t)|^2 + \rho_6|z(t)|^2$.

Theorem: Let $p > 0$, $c_2 \geq c_1 > 0$ and $\lambda_{0i} \geq 0, \lambda_{1i} \geq 0, \lambda_{2i} \geq 0$ such that $\lambda_{0i} \geq \lambda_{1i} + \lambda_{2i}$ for all $1 \leq i \leq N$. Let $\lambda : R \times S \rightarrow R$ be a continuous nondecreasing function with respect to $s \in R$ for all $s \geq 0$ and $1 \leq i \leq N$. Moreover $\lambda(s, i)/s > 0$ for all $s > 0$ and $1 \leq i \leq N$. Assume that there exists a function $V \in C^{2,1}(R^n \times R_+ \times S; R^+)$ such that the following inequality [10,9]

$$c_1|X|^p \leq V(X, t, i) \leq c_2|X|^p, \quad \forall (X, t) \in R^n \times [-\tau, \infty) \tag{13}$$

is satisfied and, moreover, for all $1 \leq i \leq N$, let

$$E\mathcal{L}V(X, Y, Z, t, i) \leq -\frac{1}{2}E\lambda(c_1|X|^p,) \tag{14}$$

for all $t \geq 0$. For all $X, Y, Z \in R^n, t \geq 0$ and $1 \leq k \leq N$, we assume

$$\mathcal{L}V(X, Y, Z, t, i) \leq -\lambda_{0i} \max_{1 \leq k \leq N} V(X, t, k) \tag{15}$$

$$+ \lambda_{1i} \min_{1 \leq k \leq N} V(Y, t - \rho(t), k) \tag{16}$$

$$+ \lambda_{2i} \min_{1 \leq k \leq N} V(Z, t - \sigma(t), k)$$

$$- \lambda(\max_{1 \leq k \leq N} V(X, t, k), i) \tag{17}$$

then under the assumptions A1 to A3 the trivial solution of is globally p th moment asymptotically stable.

Proof:

In [6] was shown that the following inequalities are valid

$$\min_{1 \leq i \leq N} EV(Y, t - \rho(t), i) \leq \max_{1 \leq i \leq N} EV(X, t, i) + \frac{1}{2(1 + \lambda_{1i})} E\lambda(\max_{1 \leq i \leq N} V(X, t, i)) \tag{18}$$

and

$$\min_{1 \leq i \leq N} EV(Z, t - \sigma(t), i) \leq \max_{1 \leq i \leq N} EV(X, t, i) + \frac{1}{2(1 + \lambda_{2i})} E\lambda(\max_{1 \leq i \leq N} V(X, t, i)) \tag{19}$$

Based on the above assumptions A1 to A3 and the above inequalities (13) to (19), we obtain

$$\begin{aligned} E\mathcal{L}V(X, Y, Z, t, i) &\leq -\lambda_{oi} \max_{1 \leq k \leq N} EV(X, t, k) & (20) \\ &+ \lambda_{1i} \min_{1 \leq k \leq N} EV(Y, t - \rho(t), k) \\ &+ \lambda_{2i} \min_{1 \leq k \leq N} EV(Z, t - \sigma(t), k) \\ &- \lambda(\max_{1 \leq k \leq N} EV(X, t, k), i) \\ &\leq -\frac{1}{2} E\lambda(c_1|X|^p, i) & (21) \end{aligned}$$

for all $1 \leq k \leq N$. Since $\lambda(s, i)/s > 0$ for all $s > 0$ and $1 \leq k \leq N$, we can show that $E\lambda(c_1|X|^p, i) > 0$ if $E|X|^p > 0$ and this completes the proof.

Remark 1: By choosing as a Lyapunov function $V : R \times R_+ \times S \rightarrow R_+$ as $V(X, t, i) = X^2, \forall i$, we obtain as a consequence of the stability conditions based on the assumptions (A1) to (A3):

$$\begin{aligned} \rho_5 + \rho_6 &> 2\rho_4 - \rho_1 & (22a) \\ \rho_1 - 2\rho_4 &> \rho_2 + \rho_3 \end{aligned}$$

The derived conditions (22a) can be used in reversed engineering design.

Remark 2: The above Theorem was adapted from [6]. The derived theoretical concepts are illustrated in an example.

Example 1: Let us consider a two-gene Markovian model (5a) with $A = \text{diag}(1.4 \quad 1.52)$, $C = \text{diag}(1.4 \quad 1.32)$ and $D = \text{diag}(1 \quad 1)$ and $W = \begin{bmatrix} 1.2 & 1 \\ 1.2 & -1.3 \end{bmatrix}$

and $\zeta(t)$ being the Gaussian noise. Let $r(t)$ be a right-continuous Markov chain taking values in $S = 1, 2$ with $\Gamma = (\gamma_{ij})_{2 \times 2} = \begin{bmatrix} -1.68 & 1.68 \\ 1.49 & -1.49 \end{bmatrix}$. Let us assume that we want to determine the parameters ρ_1, ρ_2 and ρ_3 . Based on the conditions (22a) and the Theorem, we can derive the inequalities $\rho_1 > 0.74$ and $\rho_1 - 3.04 > \rho_2 + \rho_3$ to be fulfilled in order to ensure the stability of system (5a).

In summary, we have shown that the most detailed model of GRN known yet and described by a hybrid stochastic retarded system is asymptotically stable.

5 Conclusion

We analyzed the dynamical behavior of genetic regulatory networks subject to noise perturbations and time-delays, and with both continuous and discrete states described by Markovian jumping systems based on the theory of hybrid stochastic retarded systems. The proposed model represents the most complex GRN model known so far in the literature. We assumed that the nonlinear nominal system and the noise intensity are bounded and that the Markov chain is independent of the Brownian motion. Specifically, we applied these theoretical concepts to study asymptotic stability properties of gene regulatory networks. In this sense we established stability results for the perturbed genetic regulatory network and determined the conditions that ensure the existence of globally p th moment asymptotically stable equilibria of the perturbed system. A sufficient condition for the nonlinear part and the noise intensity function are derived. The established results have potential application for reverse engineering and robust biosynthetic gene regulatory network design.

References

- [1] R. Tanaka, H. Okano and H. Kimura (2006), Mathematical description of Gene Regulatory Units, *Biophysical Journal*, **vol. 91**, p. 1235-1247.
- [2] A. Meyer-Baese, C. Plant, S. Cappendijk and F. Theis (2010), Robust Stability Analysis of Multi-Time Scale Genetic Regulatory Networks under Parametric Uncertainties, *BIOCAMP 2010*, p. 854-863.
- [3] M. Simpson, C. Cox and G. Sayler (2003), Frequency Domain Analysis of Noise in Autoregulated Gene Circuits, *Proceedings of National Academy of Sciences*, **vol. 100**, p. 4551-4556.
- [4] N. Monk (2003), Oscillatory Expression of Hes1, p53 and NF-kB Driven by Transcriptional Time Delays, *Curr. Biol.*, **vol. 13**, p. 1409-1413.
- [5] F. Ren and J. Cao (2008), Asymptotic and Robust Stability of Genetic regulatory Networks with Time-Varying Delays, *Neurocomputing*, **vol. 71**, p. 834-842.
- [6] L. Huang, X. Mao and F. Deng (2008) Stability of Hybrid Retarded Systems, *IEEE Transactions on Automatic Control*, p. 3413-3420.
- [7] P. Li, J. Lam and Z. Shu (2008) On the Transient and Steady-State Estimates of Interval Genetic Regulatory Networks, *IEEE Transactions on Systems and Cybernetics, part B*, p. 336-349.
- [8] J. Liang, J. Lam and Z. Wang (2008) State Estimation for Markov-Type Genetic Regulatory Networks with Delays and Uncertain Mode Transition Rates, *Physics Letters A*, p. 4328-4337.
- [9] Ali Saberi und Hassan Khalil (1984), Quadratic-type functions for singularly perturbed systems, *IEEE Transactions on Automatic Control*, p. 542-550.
- [10] M. Vidyasagar (1993), Nonlinear Systems Analysis, *Prentice Hall*.

Towards Integrating National Electronic Care Records in Saudi Arabia

Mohammed Alnuem, Samir EL-Masri, Ahmed Youssef, Ahmed Emam
Department of Information Systems, King Saud University, Riyadh, KSA

Abstract - *The importance of sharing and integration of patient health records that are dispersed and distributed on many healthcare organizations has pushed many countries to work hard towards achieving this objective. In this paper, a model of integration has been proposed, to share at least a brief summary of important information about the patient health record from many healthcare providers. The challenges of this integration are numerous, although the focus in this paper is on three; integration issues, security, and uniqueness of the patient identifier. In this regard, a centralized summary healthcare record has been proposed. That summary will contain an integrated summary of the patient health record collected from all encounters records at different hospitals and clinics.*

Keywords: Electronic Summary Care Record (SCR), Universal Patient Identifier (UPI).

1 Introduction

Saudi Arabia is one of the biggest economies in the Middle East. Due to the increase in the healthcare demand and the relatively wealthy population, latest reports show that spending in the healthcare in Saudi Arabia has increased to more than 16 billion USD in the public sector the last year. On the other hand the cost of healthcare is relatively low due to the high availability of medical care staff from nearby Arab countries and the Far East. This resulted in a unique situation where health care became very accessible through thousands of small-midrange private health care centers. Add to this the limits found in the public GPs like the lack X-Rays and advanced laboratory facilities. This resulted in a situation where many Saudis have their health care in many different private health care centers during their life. Due to this the patient health record may become fragmented in many different clinics and hospitals.

This is an important challenge that may face Saudi Arabia health system in particular. Hence adopting a unified electronic health record for patient may be a challenge. In this work we will study the possibility of having an integrated electronic summary Care record instead of having a complete electronic health record which may not be practical in the Saudi case because of the expected high level of records per patients.

2 Saudi Health Care System

In Saudi Arabia, the healthcare system can be classified as a national healthcare system and the private healthcare sector. National healthcare system provides health care services through a number of government agencies. On the other hand, there is a growing role and increased participation from the private sector in the provision of health care services. The Ministry of Health (MOH) is the major government agency entrusted with the provision of preventive, curative and rehabilitative health care for the Saudi Arabia's population. The Ministry provides primary health care (PHC) services through a network of health care centers throughout Saudi Arabia. The MOH is considered the lead Government agency responsible for the management, planning, financing and regulating of the health care sector. The MOH also undertakes the overall supervision and follow-up of health care provided by the private sector. There are also some other mini national health services that provide healthcare services for their sectors such as: the Ministry of Defense and Aviation (MODA) hospitals, the Ministry of Interior (MOI) hospitals, the Saudi Arabian National Guard (SANG) hospitals, universities' hospitals, The Royal Commission for Jubail and Yanbu hospitals and clinics, King Faisal Specialist Hospital

and research center, King Khalid Eye Specialist Hospital, and so on [9].

In Saudi Arabia, There are numerous of hospitals and clinics either public or private. For public hospitals, there are 240 hospitals around Saudi Arabia, 39 of them are in Riyadh. For private hospitals, there are 327 hospitals around Saudi Arabia, 230 of them are in Riyadh. For public clinics, there are 1690 clinics in Saudi Arabia, 361 of them in Riyadh. For private clinics, there are 620 clinics in Saudi Arabia, 205 of them in Riyadh [9].

Table 1: Number of Hospitals and Clinics in SA

	Public	Private
Hospital in SA	240	327
Clinic in SA	1690	620
Hospital in Riyadh	39	230
Clinic in Riyadh	361	205

3 Electronic Summary Care Record

Electronic Summary Care Record (SCR) extends the concept of digital health summaries to create an updated and centrally stored patient's summary record, extracting key data from local systems after each encounter [1]. SCR is formed from files of the same patient and belongs to different hospitals within the country. The record should contain an encounter, admission, discharges, electronic clinical records, medications etc. SCR should be shared and accessible across the hospitals and clinics taking into account the security rules, regulations and all medical application international standards.

The main benefits of having a shared SCR can be summarized as following:

- Reduce/Eliminate the time usually needed to transfer physical copies of patient data between hospitals.

- Provide more information about the patient condition from different sources which will increase diagnosis accuracy.
- Reduce the cost in terms of time and diagnosis.
- Reduce the medical errors and hence increase the healthcare quality.
- Help producing healthcare statistics (medical and clinical informatics) which plays important role in developing healthcare strategies and planning future improvements and extensions in healthcare systems.

Many problems occur in Saudi hospitals through transferring patients' summary records from one healthcare organization to another which result in affecting the quality of the outcomes of the treatment. In order to solve these problems, a solution to integrate the hospitals systems is proposed to be implemented in Saudi Arabia that helps to increase healthcare integration and quality.

4 Related work

Three main challenges have been identified in order to create integrated summary healthcare record in Saudi hospitals: Integration, Security and the need to have a unified patient identifier (UPI).

4.1 Integration

In usual cases the patient can have multiple electronic health records (EHR) in many hospitals recorded accordingly with their medical encounters. EHR needs to be integrated among the hospitals in order to obtain a total overview of a patient's health-history. Many countries in the world are seeking to integrate and communicate their patient information among their hospitals in order to help them to improve the quality of healthcare outcomes. Some of these countries are Canada, Australia, England, USA, India, and Korea. Canada, Australia and England have the development of national healthcare strategies. Electronic Health Record (EHR) is considered as the main component of the healthcare infrastructure. Some obstacles have been discussed such as politics, geographies, population density [2].

In England, building national Dispersed Electronic Health Record (DEHR) is proposed to be a solution to integrate the hospitals systems in England. England is divided into five different geographical areas called "clusters". Each cluster represents one database and the database could be divided into more than one instance. The National Care Record Service (NCRS) is the existing EHR project which allows the authorized people to access the patient record 24 hours a day, seven days a week [2]. NCRS based on two components: Detailed Care Record and national Summary Care Record (SCR). Detailed Care Record is used inside local healthcare where the patient care is happen [2]. The national SCR extends the concept of digital health summaries to create an updated and centrally stored patient's summary record, extracting key data from local systems after each encounter [1] such as an encounter, admission, discharges, electronic clinical records, clinical messaging etc [1]. It can be easily extracted from the hospital systems and loaded to a central database called "Spine" using Dispersed Electronic Health Record (DEHR). Spine stores the important patient records for all England's 50 million population [2]. SCR is used instead of dispersed electronic health record because a dispersed electronic health record will take a long time to be built. Uncertainty about the quality and provenance of SCR data raises concerns about patient safety, as key data may be absent and old data may persist, partially because of a lack of ownership of the summary [1].

In Australia, HealthConnect is the national Australian EHR service which involves the collection storage and sharing of patients' information in summary. HealthConnect aimed to improve the healthcare outcomes by increasing quality and enhance patient safety. The components of HealthConnect model are a series of event summaries which contain key information about specific healthcare event such as allergies, diagnosis, medications, referrals, and EHR lists which will be extracted from the event summaries. Therefore, predefined HealthConnect views are available to access these stored event summaries. Each HealthConnect electronic health record would be stored in two locations: a HealthConnect Record System (HRS) and the National Data Store. HRS is used to process the event summaries and transactions while National Data Store preserves copies of EHR [2].

In India, Distributed Infrastructure for Global EHR Technology (DIGHT) project was built to integrate electronic health record in India. Scalability, reliability and high availability were the most challenges of DIGHT project. Some requirements of EHR storage have been implemented in order to meet the challenges such as high data availability in terms of hardware and software, high performance which will ensure the system can work effectively any time and data security which protect the patient data from any unauthorized access by using data replication algorithm. Using central storage could be a solution but it degrades the high availability and performance. However, clustering technology was used in DIGHT project as a solution [5].

In Korea, National e-health project was built to integrate electronic health record among the hospitals. Many positive outcomes were achieved: improving transparency and effectiveness, enhancing accessibility and quality, strengthening quality and satisfaction of patients, reducing medical expenses, management rationalization of healthcare organization, and enhancing accountability through public healthcare inspection system. Some policies were applied in order to prevent the patient data access from any unauthorized access [6].

Some challenges have arisen during the integration process in the data heterogeneity [4]. There are two different types of problems that have to be addressed to make the patients' data consistent in order to share the EHR between multiple Database Management System (DBMS). First heterogeneity problem is on DBMS level which is different hospital use different DBMS. Therefore, traditional database normalization ACID (Atomicity, Consistency, Isolation and Durability) properties could be missed across the hospitals. Relaxed ACID properties were proposed to be solution. Second heterogeneity problem is on electronic health record level which is EHR incompatibility between different hospitals. No solution was proposed for this problem [4].

4.2 Security

Patient data privacy and confidentiality are considered the most important issues when exchanging

and sharing relevant patient data among multiple systems. Secure Dispersed Electronic Health Record based on cryptographic constructions was proposed to address these concerns in order to be accepted by the patient [3]. In order to protect EHR, there should be some policies, regulations, and agreements that the patients, physicians, and the other stakeholders agree on. Therefore, EHR will be protected against any illegal use [7].

Some of the agreements that used in Electronic Health Record in Serbia (EHR-S): (a) patient consent to access EHR-S agreement which let the patient sign on the agreement, (b) medical institution / medical practice access to EHR-S which allow authorized medical practitioners to update, request, and receive up-to-date EHR in timely and secure manner, (c) hospital access to EHR-S which allow authorized physicians and pharmacists to update, request, and receive up-to-date EHR in timely and secure manner, and (d) Emergency Department (ED) which has an access to the patient medical history to help the patients when they arrive to ED [7].

4.3 Universal patient identifier

Universal Patient Identifier (UPI) was proposed to address the patient uniqueness issues. It consists of four parts: birth date code with 7 digits, geographical code with 6 digits, and sequence code with 5 digits to identify people born on the same date and geographical area, and single check digit. Birth date code can be divided into three codes as: day (1-31) with 2 digits, month (1-12) with 2 digits and year (0-99) with 3 digits. Geographical code can be divided into two codes: Latitude code (0-180) with 3 digits and Longitude code (0-360) with 3 digits. Hospital code could be part of sequence code.

For example, a person born on 1 March 1993 in Minneapolis, MN the code has the appearance: 9930301^044237^00047^2 [8].

5 SCR-SA

The authors have done a survey on the main hospitals in Saudi Arabia such King Faisal specialized hospital, King Fahad Medical City, King Khalid University Hospital and found no evidence of any kind

of integration between them. In order to integrate the electronic health records among Saudi Arabia's hospitals we identified two main requirements which are missing in Saudi Hospitals:

- Patient unique identifier.
- Summary care record.

These two requirements need to be considered and implemented on each of hospital systems as minimum requirement for integration.

Identifying the uniqueness of patient is a major concern in national SCR. Patient unique identifier in Saudi Arabia could be national ID for citizens and resident permit (known as Iqama) number for foreigners. But all the hospitals are not considering the national ID or Iqama number as a unique identifier and they have their own unique identifier such as medical record number. Therefore, implementing a national ID or Iqama number as a patient unique identifier across the hospitals may face many challenges. However, the Universal Patient Identifier (UPI) is proposed to be used as the patient identifier in Saudi Arabia. UPI has been suggested to be consisted of three parts as: Birth Date in Gregorian with eight digits (e.g. 19800126), Region Code with two digits (01-13) because of the thirteen regions in Saudi Arabia, sequence with three letters (e.g. XWU) to identify people born in the same date and region. Therefore, the length of the suggested UPI is thirteen characters. E.g. 1980012604XWU (Birth date: 19800126, Region Code: 04, Sequence: XWU).

The Summary Care Record (SCR) is an electronic health record summary of the patient such as an encounter, admission, discharges, electronic clinical records, medications etc. Each hospital must provide the patients' SCR in order to be easily extracted and loaded to a central database (see Figure 1). SCR should be protected; only the authorized people can access the part that they need only. So, there are some rules, regulations, and policies should be applied on SCR-SA in order to protect the data.

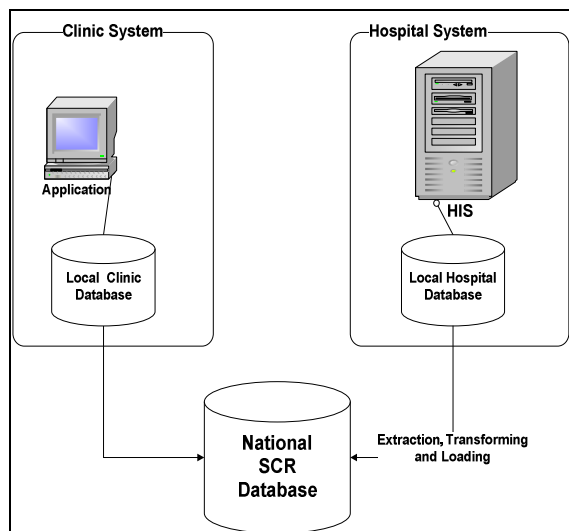


Figure 1: General Structure of SCR System

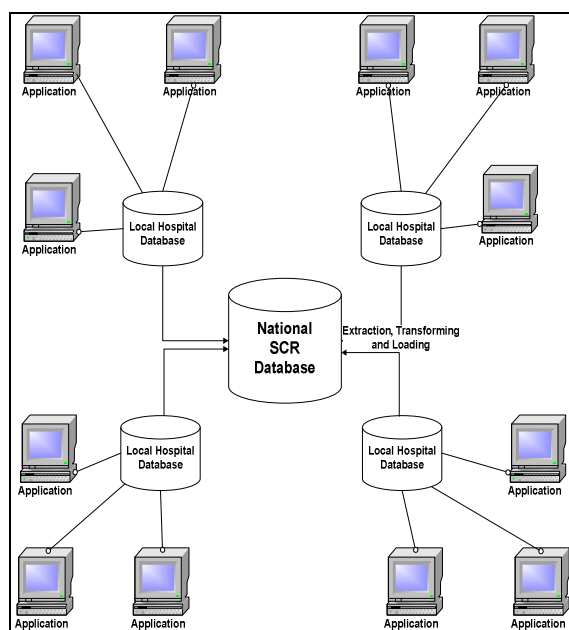


Figure 2: Detailed Structure of SCR System

In Figure 2, we suggest to have a centralized national SCR database to be the center for the patients' summary data in Saudi Arabia. In order to link the hospitals and clinics to the centralized database, there should be an Extraction, Transforming, and Loading (ETL) channel between the centralized national SCR database and the client because the different hospitals and clinics use different systems and DBMS.

In order to inquire the SCR, we need to follow our proposed procedure. The procedure (see Figure 3) shows that the hospital checks whether the patient's record available in the local hospital database to be fetched, or it checks national summary care record centralized database to get the SCR. A new record will be created in both local hospital database and national SCR database if the record is not there.

SCR is believed be enough for a healthcare professional to make a decision in many cases. However in case more details about a particular medical encounter, lab or radiology results are needed, the system can retrieve them from that particular hospital.

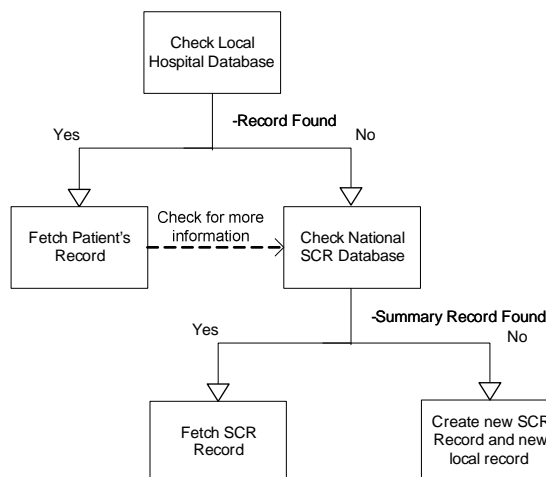


Figure 3: SCR Inquiry Procedure

6 Conclusion and future work

By applying DEHR-SA between the Saudi hospitals and clinics, we obtain a lot of benefits for hospitals, patients and ministry of health. This paper showed the elements, components, methodologies and approaches to the proposed system. The paper has showed the importance and the value added of the Summary Care Record and how it can be implemented in addition to the issues of the unique patient identifier and security. In the future, we will investigate more in depth information about the Saudi hospital systems and to make a survey to have their specifications that might help to fully integrate the Saudi hospitals systems.

Acknowledgement

This work is part of a research project funded by a grant through KACST/National Plan for Science and Technology in the Kingdom of Saudi Arabia. Grant number: 09-INF880-02.

7 References

- [1] Enrico Coiera, "Do we Need a National Electronic Summary Care Record?", *The Medical Journal of Australia*, 2011.
- [2] Jalal-Karim and W. Balachandran, "The National Strategies for Electronic Health Record in three developed countries: General Status", Brunei University, Cleveland Road, 2008.
- [3] Jinyuan Sun and Yuguang Fang, "Cross-Domain Data Sharing in Distributed Electronic Health Record Systems", *IEEE*, 2010.
- [4] Lars Frank and StigKjcer Andersen, "Evaluation of Different Database Designs for Integration of Heterogeneous Distributed Electronic Health Records", *IEEE*, 2010.
- [5] Jim Dowling and SeifHaridi, "Developing a Distributed Electronic Health-Record Store for India", 2009.
- [6] HeeJeong Cheong, Na Yoon Shin and YounBaekJoeng, "Improving Korean Service Delivery System in Health Care: Focusing on National E-health System", *International Conference on eHealth, Telemedicine, and Social Medicine*, 2009.
- [7] Srdjan B Stakic and Nada Teodosijevic, *Agreements Based Distribution of Responsibilities in National Electronic Health Records System*, *IEEE*, 2010.
- [8] Paul C. Carpenter and Christopher G. Chute, "The Universal Patient Identifier: A Discussion and Proposal", *Division of Endocrinology and Section of Medical Information Resources*, 1994.
- [9] Ministry of Health Portal, Kingdom of Saudi Arabia, "Directory of medical centers", [Online]. Available:<http://www.moh.gov.sa/en/eServices/Directory/Pages/MedicalCentersService.aspx>. [Accessed: May 23, 2011].

Dynamics of HIV-1 Associated Kaposi Sarcoma During HAART Therapy

Frank Nani and Mingxian Jin

Department of Mathematics and Computer Science
Fayetteville State University, Fayetteville, NC 28301, USA

Abstract –The techniques of mathematical modeling and investigative computer simulations are used to study the qualitative aspects of the patho-physiodynamics of HIV-1 associated Kaposi sarcoma (KS) during Highly Active Anti-Retroviral Therapy (HAART) of AIDS. Using a system of non-linear deterministic differential equations, the model incorporates the biologically measurable and clinically relevant immunological interactions and parameters. In particular, the computer simulations elucidate the role of $CD8^+$ T lymphocyte in the annihilation and persistence of Kaposi sarcoma during HAART.

Keywords: Kaposi sarcoma, mathematical modeling, HAART efficacy, computer simulations, persistence of Kaposi Sarcoma

1 Introduction

Human Herpes Virus 8 (HHV8) acts in association with HIV-1 to induce lympho-proliferation and Kaposi sarcoma (KS) in AIDS patients. The clinical and histo-pathological aspects of KS have been documented by Kemény et al. [3], Lesbordes et al. [5], and Zhu et al. [11].

The role of $CD8^+$ T lymphocytes in regulating the growth of KS has been investigated by Li et al. [6] and Stebbing et al [8]. The use of adoptive immunotherapy with activated autologous $CD8^+$ T cells with interleukin-2 infusion in treatment of AIDS was described in a paper by Klimas et al. [4], Touloumi et al. [9], and Urassa et al. [10]. The patho-physio-dynamics of KS during HAART has been clinically investigated by Bihl et al. [1], and Dupont et al. [2].

In the current research, we shall present a mathematical model of the patho-physio-dynamics of KS associated with AIDS during HAART. This paper is extension of our earlier mathematical model on HIV-1 AIDS dynamics during latency phase [7]. Investigative computer simulations will be used to elucidate the effect of adoptive transfer of $CD8^+$ T cells on Kaposi sarcoma dynamics during HAART. This research is one of the major attempts to construct a clinically plausible mathematical model which incorporates HAART therapy, HIV-1 induced AIDS dynamics, and Kaposi sarcoma.

2 Parameters

In this section, the model parameters, constants, and variables are presented as modified from [7].

- x_1 : the number density of non-HIV-1-infected $CD4^+$ helper T-lymphocytes per unit volume at any time t
- x_2 : the number density of HIV-1 infected $CD4^+$ helper T-lymphocytes per unit volume at any time t
- x_3 : the number density of HIV-1 virions in the blood plasma per unit volume at any time t
- x_4 : the number density of HIV-1 specific $CD8^+$ cytotoxic T-lymphocytes per unit volume at any time t
- x_5 : the concentration of drug molecules of the HAART treatment protocol at any time t
- x_6 : The number of Kaposi sarcoma cancer cells in the AIDS patient at any time t during HAART
- S_1 : rate of supply of un-infected $CD4^+$ T_4 -lymphocytes
- S_2 : rate of supply of latently infected $CD4^+$ T_4 -lymphocytes
- S_3 : rate of supply of HIV-1 virions from macrophage, monocytes, microglial cells and other lymphoid tissue different from T_4 -lymphocytes
- S_4 : rate of supply of $CD8^+$ T_8 lymphocytes from the thymus
- D : rate of HAART drug infusion by transdermal delivery
- a_i, b_i : constant associated with activation of lymphocytes by cytokine interleukin-2 (IL-2) ($i = 1, 2, 3, 4$)
- α_i : constant associated with HIV-1 infection of $CD4^+$ T_4 helper cells ($i = 1, 2, 3$)
- β_1 : the number of HIV-1 virions produced per day by replication and budding in $CD4^+$ T_4 helper cells
- β_2 : rate constant associated with replication and “budding” of HIV-1 in syncytia $CD4^+$ T_4 helper cells per day per microliter (μl) and released into the blood plasma
- β_3 : the number of HIV-1 virions produced per day by replication and “budding” in non-syncytia $CD4^+$ T_4 helper cells and released into the blood plasma
- η_i : constant depicting the rate of which HIV-1 virions incapacitate the $CD8^+$ T_8 cytotoxic cells ($i = 1, 2$)
- (σ_0, λ_0) : Michaelis-Menten metabolic rate constants associated with HAART drug elimination
- (σ_i, λ_i) : Michaelis-Menten metabolic rate constants associated with HAART drug pharmacokinetics ($i = 2, 3$)

(σ_4, λ_4) : Michaelis-Menten metabolic rate constants associated with cytolytic action of $CD8^+$ against Kaposi Sarcoma cancer cells

γ_4 : constant depicting the cytolytic efficacy of $CD8^+$ T cells against Kaposi sarcoma cancer cells

ξ_i : cytotoxic coefficient where $0 \leq \xi_i \leq 1$ ($i = 2, 3$)

q_i : constant depicting competition between infected and un-infected $CD4^+$ T₄ helper cells ($i = 1, 2$)

k_i : constant depicting degradation, loss of clonogenicity or "death" ($i = 1, 2, 3, 4$)

e_{i0} : constant depicting death or degradation or removal by apoptosis (programmed cell death) ($i = 1, 2, 3, 4$)

K_i : constant associated with the killing rate of infected $CD4^+$ T₄ cells by $CD8^+$ T₈ cytotoxic lymphocytes ($i = 1, 2$)

All the parameters are positive

c_i : kinetic constants depicting logistic tumor growth for Kaposi sarcoma

3 Model Equations

The following system of non-linear deterministic ordinary differential equations models the patho-physiological dynamics of HIV-1 induced AIDS virions and associated Kaposi sarcoma cancer cells, $CD4^+$ (infected and non-infected) T cells, and $CD8^+$ T cells during HAART therapy.

$$\begin{cases} \dot{x}_1 = S_1 + a_1 x_1^2 e^{-b_1 x_1} - \alpha_1 x_1 x_3 - q_1 x_1 x_2 - k_1 x_1 - e_{10} \\ \dot{x}_2 = S_2 + a_2 x_1 x_2 e^{-b_2 x_2} + \alpha_2 x_1 x_3 - q_2 x_1 x_2 - k_2 x_2 - \beta_1 x_3 \\ \quad - K_1 x_2 x_4 - e_{20} - \frac{\xi_2 \sigma_2 x_2 x_5}{\lambda_2 + x_5} \\ \dot{x}_3 = S_3 + \beta_2 x_2 x_3 + \beta_3 x_3 - \alpha_3 x_1 x_3 - \eta_1 x_3 x_4 - k_3 x_3 - e_{30} \\ \quad - \frac{\xi_3 \sigma_3 x_3 x_5}{\lambda_3 + x_5} \\ \dot{x}_4 = S_4 + a_4 x_1 x_4 e^{-b_4 x_4} - K_2 x_2 x_4 - \eta_2 x_3 x_4 - \gamma_4 \frac{\sigma_4 x_4 x_6}{\lambda_4 + x_4} \\ \quad - k_4 x_4 - e_{40} \\ \dot{x}_5 = D \left| \sin nt \right| - \frac{\sigma_0 x_5}{\lambda_0 + x_5} - \frac{\sigma_2 x_2 x_5}{\lambda_2 + x_5} - \frac{\sigma_3 x_3 x_5}{\lambda_3 + x_5} \\ \dot{x}_6 = c_1 x_6 - c_2 x_5^2 - \frac{\sigma_4 x_4 x_6}{\lambda_4 + x_4} \\ x_i(t_0) = x_{i0} \quad \text{for } i = \{1, 2, 3, 4, 5, 6\} \end{cases} \quad (3.1)$$

4 Simulation results and discussion

A brief summary of the simulation results will be presented in this section. Figure 1 and Figure 2 correspond respectively to hypothetical HIV-1 KS patient's physiological parametric configurations P_1 (Table 1) and P_2 (Table 2).

(i) Hypothetical clinical case #1 [Figure 1, P_1]:

It is observed that HAART treatment successfully annihilates the HIV-1 virions in the blood plasma and reduces the number density of HIV-1 infected $CD4^+$ T cells, whereas the non-infected $CD4^+$ T cells proliferate to clinically efficacious levels. On the other hand, the HIV-1 specific $CD8^+$ T cells are eliminated and consequently the Kaposi sarcoma proliferates out of control.

(ii) Hypothetical clinical case #2 [Figure 1, P_1']:

In this scenario, the physiological parametric configuration is the same as that of P_1 except that there is an adoptive transfer of 2000 units of ex-vivo interleukin-2 activated $CD8^+$ cytotoxic T cells. In P_1' , the S_4 value is now assigned to a value of 2000 instead of 10 as in P_1 . The therapeutic outcome is clinically efficacious because the Kaposi sarcoma is annihilated.

(iii) Hypothetical clinical case #3 [Figure 2, P_2]:

This scenario discusses the effect of HIV-1 latent viral reservoirs on the treatment outcome. In particular, S_3 is set to a value of 1000, depicting the influx of 1000 HIV-1 virions from reservoirs such as microglial cells, macrophages and dendritic cells. It is observed that even though the HAART dose rate D is increased to 4000 units, there is a subsequent therapeutic failure because the non-infected $CD4^+$ cell number plummets as HIV-1 virions overwhelm the immune system. On the other hand, the adoptive transferred 2000 units of $CD8^+$ cells are able to keep the Kaposi sarcoma cancer cells under the clinically detectable level of 1000 cells.

(iv) Hypothetical clinical case #4 [Figure 2, P_2']:

The physiological parametric configuration is the same as that of P_2 except for the fact that the HAART drug dose rate D is increased to 5000 units, and the non-infected $CD4^+$ T cells (x_1) are given an extra boost of interleukin-2 (IL-2) dose and as such the value of a_1 is now 0.45. The outcome is clinically efficacious because the plasma HIV-1 virions (x_3), the HIV-1 infected $CD4^+$ T cells (x_2), and the KS cancer cells are kept under the clinically detectable level of 1000 cells, whereas the non-HIV-1 infected $CD4^+$ T cells (x_1) repopulate to clinically efficacious level.

5 Summary

Our research can be summarized in the following statements:

- (i) It is possible for HAART therapy to annihilate the HIV virions without necessarily eliminating KS.
- (ii) Adoptive transfer of $CD8^+$ T cells at a predetermined dose rate can annihilate KS cancer cells.

(iii) It will require both HAART and adoptive transfer CD8⁺ T cells incubated with IL-2 to decimate both HIV-1 virions and the Kaposi sarcoma cancer cells.

TABLE 1. Hypothetical AIDS Patient Parametric Configuration P_1

$S_1 = 800$ /day/ μ $a_1 = 0.15$ /day/cell/ μ $b_1 = 0.01$ /cell/ μ $\alpha_1 = 0.5$ /day/virions/ μ $k_1 = 0.0005$ /day/ μ $q_1 = 0.00045$ /day/ μ /cell $e_{10} = 0.0025$ cells/day/ μ $x_{10} = 500$ cells/ μ	$S_2 = 800$ /day/ μ $a_2 = 0.11$ /day/cell/ μ $b_2 = 0.004$ /cell/ μ $\alpha_2 = 0.5$ /day/virions/ μ $k_2 = 0.005$ /day/ μ $q_2 = 0.00001$ /day/ μ /cell $\beta_1 = 1.5$ virions/CD4 ⁺ /day $K_1 = 0.0001$ /day/ μ $e_{20} = 0.0005$ cells/day/ μ $x_{20} = 400$ cells/ μ	$S_3 = 10$ /day/ μ $\beta_2 = 0.0085$ virions/CD4 ⁺ /day/ μ $\beta_3 = 2.75$ virions/CD4 ⁺ /day $\alpha_3 = 0.027$ /day/virions/ μ $k_3 = 0.0001$ /day $e_{30} = 0.0001$ /day $\eta_1 = 0.055$ $\xi_2 = 0.85$ $\xi_3 = 0.0001$ $x_{30} = 1000$ virions/ μ	$S_4 = 10$ /day/ μ $a_4 = 0.35$ /day/cell/ μ $b_4 = 0.01$ /cell/ μ $K_2 = 0.0024$ /day/ μ $k_4 = 0.08$ /day/ μ $e_{40} = 0.0002$ cells/day/ μ $\eta_2 = 0.055$ $\gamma_4 = 0.15$ $x_{40} = 1500$ cells/ μ	$D = 4000$ units $\sigma_0 = 0.5$ mg/day $\sigma_2 = 30$ mg/day $\sigma_3 = 5$ mg/day $\lambda_0 = 5$ mg/L $\lambda_2 = 10$ mg/L $\lambda_3 = 0.015$ mg/L $x_{50} = 1500$ cells/ μ $n = 5$	$c_1 = 6.405$ $c_2 = 0.00075$ $\sigma_4 = 7$ mg/day $\lambda_4 = 5.5$ mg/L $x_{60} = 2500$ cells
---	--	--	---	--	--

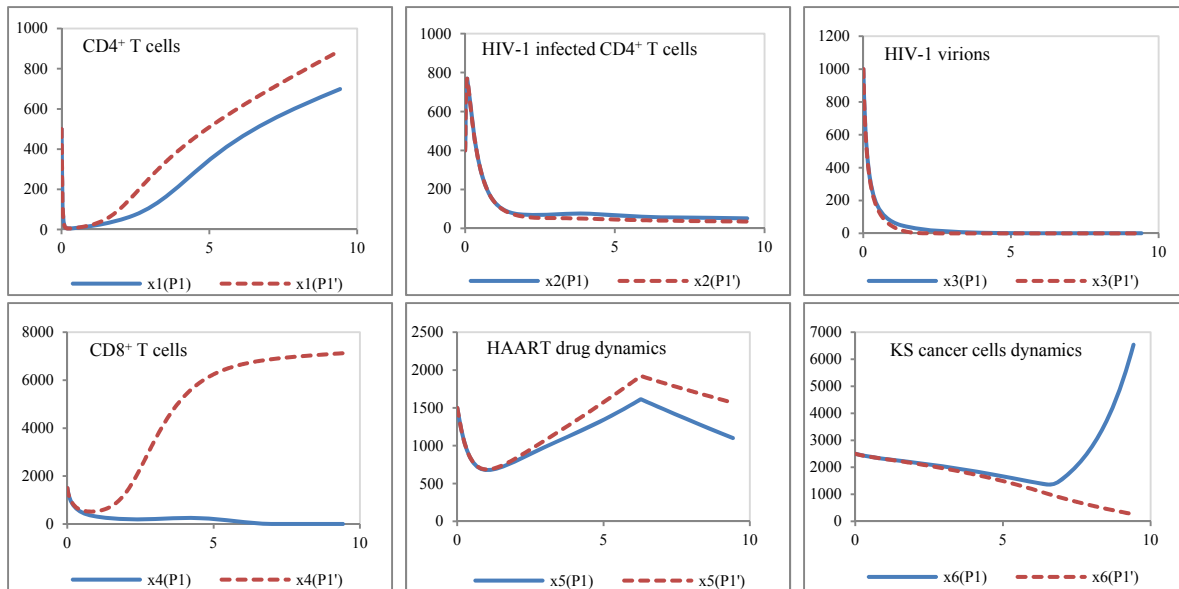


Figure 1 Simulation results using parametric configurations P_1 vs. P_1' (P_1' is the modified P_1 : same as P_1 except $S_4 = 2000$. The time axis unit is months.)

TABLE 2. Hypothetical AIDS Patient Parametric Configuration P_2

$S_1 = 800$ /day/ μ $a_1 = 0.15$ /day/cell/ μ $b_1 = 0.01$ /cell/ μ $\alpha_1 = 0.5$ /day/virions/ μ $k_1 = 0.0005$ /day/ μ $q_1 = 0.00045$ /day/ μ /cell $e_{10} = 0.0025$ cells/day/ μ $x_{10} = 500$ cells/ μ	$S_2 = 800$ /day/ μ $a_2 = 0.11$ /day/cell/ μ $b_2 = 0.004$ /cell/ μ $\alpha_2 = 0.5$ /day/virions/ μ $k_2 = 0.005$ /day/ μ $q_2 = 0.00001$ /day/ μ /cell $\beta_1 = 1.5$ virions/CD4 ⁺ /day $K_1 = 0.0001$ /day/ μ $e_{20} = 0.0005$ cells/day/ μ $x_{20} = 400$ cells/ μ	$S_3 = 1000$ /day/ μ $\beta_2 = 0.0085$ virions/CD4 ⁺ /day/ μ $\beta_3 = 2.75$ virions/CD4 ⁺ /day $\alpha_3 = 0.027$ /day/virions/ μ $k_3 = 0.0001$ /day $e_{30} = 0.0001$ /day $\eta_1 = 0.055$ $\xi_2 = 0.85$ $\xi_3 = 0.0001$ $x_{30} = 1000$ virions/ μ	$S_4 = 2000$ /day/ μ $a_4 = 0.35$ /day/cell/ μ $b_4 = 0.01$ /cell/ μ $K_2 = 0.0024$ /day/ μ $k_4 = 0.08$ /day/ μ $e_{40} = 0.0002$ cells/day/ μ $\eta_2 = 0.055$ $\gamma_4 = 0.15$ $x_{40} = 1500$ cells/ μ	$D = 4000$ units $\sigma_0 = 0.5$ mg/day $\sigma_2 = 30$ mg/day $\sigma_3 = 5$ mg/day $\lambda_0 = 5$ mg/L $\lambda_2 = 10$ mg/L $\lambda_3 = 0.015$ mg/L $x_{50} = 1500$ cells/ μ $n = 5$	$c_1 = 6.405$ $c_2 = 0.00075$ $\sigma_4 = 7$ mg/day $\lambda_4 = 5.5$ mg/L $x_{60} = 2500$ cells
---	--	--	---	--	--

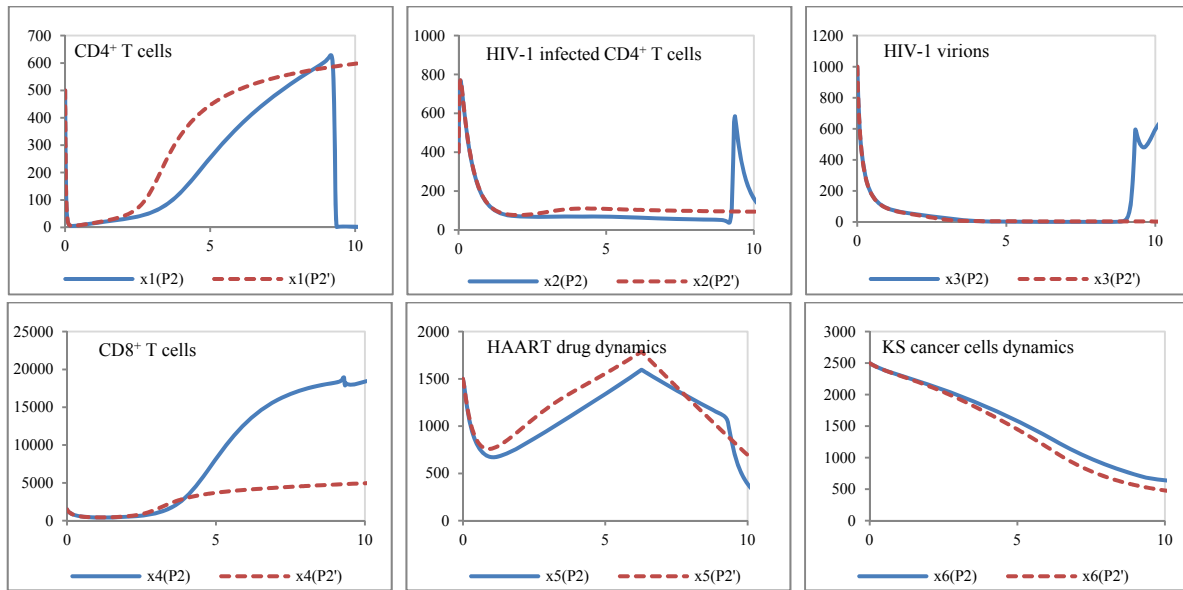


Figure 2 Simulation results using parametric configurations P_2 vs. P_2' (P_2' is the modified P_2 : same as P_2 except $a_1=0.45$, $D=5000$. The time axis unit is months.)

6 References

- [1] F. Bihl, et al, "Kaposi's sarcoma-associated herpes virus-specific immune reconstitution and antiviral effect of combined HAART/chemotherapy in HIV clade C-infected individuals with Kaposi's sarcoma", AIDS (London, England), 21(10), pp. 1245-1252, June 2007
- [2] C. Dupont, E. Vasseur, et al., "Long-term efficacy on Kaposi's sarcoma of highly active antiretroviral therapy in a cohort of HIV-positive patients. CISH 92. Centre d'information et de soins de l'immunodéficience humaine", AIDS, 14(8), pp. 987-93, May 26, 2000
- [3] L. Kemény, et al., "Human herpes virus 8 in classic Kaposi sarcoma", Acta Microbiologica et immunologica Hungarica, 43(4), pp.391-395, 1996
- [4] N. Klimas, et al., "Clinical and immunological changes in AIDS patients following adoptive therapy with activated autologous CD8 T cells and interleukin-2 infusion", AIDS, 8(8), pp.1073-81, Aug. 1994
- [5] J. L. Lesbordes, et al., "Clinical and histopathological aspects of Kaposi's sarcoma in Africa: relationship with HIV serology", Annales de l'Institut Pasteur, Virology, 139(2), pp.197-203, Apr-Jun 1988
- [6] T. Li, Z. Qiu, A. Wang, and R. Sheng, "T-lymphocyte immune in HIV-infected people and AIDS patients in China", Zhonghua Yi Xue Za Zhi, 82(20), pp.1391-5, Oct 2002
- [7] F. Nani and M. Jin, "Mathematical modeling and simulation of latency phase HIV-1 dynamics", Int'l Conf. Bioinformatics and Computational Biology (BIOCOMP'10), vol. II, pp. 428-434, July 2010
- [8] J. Stebbing, A. Sanitt, A. Teague, T. Powles, M. Nelson, B. Gazzard, M. Bower, "CD8 count measurement has independent prognostic significance in individuals with AIDS-Kaposi sarcoma", Journal of Clinical Oncology, 2007 ASCO Annual Meeting Proceedings Part I. Vol 25, No. 18S (June 20 Supplement):20500, 2007
- [9] G. Touloumi, et al., "The role of immunosuppression and immune-activation in classic Kaposi's sarcoma", International Journal of Cancer, 82(6), pp. 817-21, 1999
- [10] W. K. Urassa, et al., "Immunological profile of endemic and epidemic Kaposi's sarcoma patients in Dar-es-Salaam, Tanzania", International Journal of Molecular Medicine, 1(6):979-82, 1998
- [11] B. Zhu, N.P. Wu, S. Hoxtermann, A. Bader, and N. Brockmeyer, "Immune activation in AIDS related Kaposi's sarcoma", Zhejiang da xue xue bao = Journal of Zhejiang University, Medical science, 32(2), pp. 101-103, Apr 2003

A Proposal of Clinical Decision Support system Architecture for Distributed Electronic Health Records

Shaker H. El-Sappagh^{1,2}, Samir El-Masri³

¹College of Science, King Saud University, Saudi Arabia

²Faculty of Computers and Information, Minia University, Egypt

³Department of Information Systems, College of Computer and Information Sciences, King Saud University, Saudi Arabia

Abstract - *Improving the quality of healthcare, reducing medical errors, guarantying the safety of patients are the most serious duty of the hospital. Electronic Health Record (EHR) was introduced to achieve these goals. HER has a very large data source which can guide and improve the clinical decision making process. In this paper, we will propose a distributed Clinical Decision Support System (CDSS) architecture which satisfies the compatibility, interoperability, and scalability objectives of EHR. The proposed framework will take advantages of EHR, data mining techniques, clinical databases, domain experts' knowledge bases, and available technologies and standards to provide decision making support for the healthcare personnel.*

Keywords: Data Mining; Knowledge Management; Clinical Decision Support Systems (CDSS); Electronic health record; Health informatics.

1 Introduction

Healthcare faces multiple problems, including high and rising expenditures, inconsistent quality, and gaps in care and access. Because of this, health care services represent a major portion of the government spending in most countries [1].

Healthcare information technology, especially EHRs, have been thought to be possible solution to healthcare problems. EHRs help administrators, physicians, nurses, researchers and healthcare personnel. EHR provides a complete, integrated and consistent view about patient conditions. However, the volume of data is very large and is increasing continuously. Healthcare personnel need to take all of the patient medical history in to consideration; they also need to connect this information together and take advices from domain experts. This huge amount of data cannot benefit physician without a helping automated system. This system can analyze these data, connect it together, integrate it with knowledge from domain expert, and search for needed knowledge - if it is possible - in

other connected systems. This system is CDSS. In this paper, we tried to build a complete architecture for this system. The proposed model will take an order and initial diagnose from healthcare personnel and provide a decision support in understandable form based on existing knowledge. The system will integrate off-line standardized knowledge bases from domain experts with online knowledge extracted continuously from EHR and clinical databases and provide applicable decisions support. The paper is organized as follow. Section 2 discusses related work. Section 3 explains the research problem. In section 4 we define CDSS. The proposed framework for CDSS is discussed in section 5. The conclusion is shown in section 6.

2 Related Works

2.1 EHR standards

Many organizations provide EHR standards that standardize the structuring, implementing, sharing, integration and interoperability in EHR environment. Some of them are ISO, CEN, CFR, ASTM, HL7, NEMA, ONCHIT, etc. Also, coding systems are critical to build a shared EHR because the new environment connects heterogeneous systems each with different terminologies. Some organizations that provide these standards are AMA, IHTSDO, CMS, WHO, etc.

2.2 Data mining and Artificial Intelligence (AI)

Applying data mining and AI techniques on EHR data has many opportunities to improve the delivery, efficiency, and effectiveness of health care [12][13], such as operations management, preventive healthcare, chronic disease treatment and prevention, association analysis, evidence-based treatment, population tracking, etc. If CDSS depends only on the Knowledge Base (KB) derived from knowledge expert, then it will be inactive and not applicable. EHR contains a very large and historical dataset which change continuously and

contains useful hidden knowledge. As a result, data mining and AI services should be embedded in the active CDSS system to continuously update the CSDD's knowledge base by the most recent patterns from EHR and clinical databases.

2.3 Knowledge representations in medical domain

Because there are many sources and uses for medical knowledge, many international methodologies and standards for representing medical and healthcare body of knowledge are integrated. Clinical workflows (clinical guidelines) are used to represent human-based medical knowledge through rule-based or flow-based guideline techniques. Furthermore, mined knowledge can be automatically extracted from clinical databases and/or EHR through data mining and AI techniques to be incorporated into human-generated knowledge in order to enhance their decision-making processes.

Both types of knowledge can be represented as logical conditions, rules, graphs/networks, or structural representations [5]. Predictive Model Markup Language (PMML) and GLIF (Guide Line Interchange Format) are examples of knowledge representation languages which are used to acquire and integrate knowledge. Also there are many tools for knowledge acquisition and representation as Unified Medical Language System (UMLS) [6], Protégé [7], GLARE [8], PROforma [9], Asbru [10].

2.4 Service Oriented Architecture (SOA)

SOA has been widely adopted to solve the interoperability of the involved heterogeneous distributed EHR systems [2][3]. It plays a key role in the integration of heterogeneous systems by the means of services that represent different system functionality, independent from the underlying platforms or programming languages, and interact via messages exchange. *Web services* also play critical role in systems interoperability.

Web services technology is defined as a systematic and extensible framework for application-to-application interaction built on top of existing web protocols. These protocols are based on XML [11] and include: Web Services Description Language (WSDL) to describe the service interfaces, Simple Object Access Protocol (SOAP) for communication between web services and client applications, and Universal Description, Discovery, and Integration (UDDI) to facilitate locating and using web services on a network [4].

3 The Research Problem

Building CDSS will improve the quality and efficiency of healthcare [17]. These systems will be more practical when coupled with Computerized Physician Order Entry (CPOE). It contains a set of knowledge bases (one in each hospital) extracted off-line from domain experts. If CDSS only depends on these knowledge bases it will be inactive and will become not applicable. The solution is to continually update these knowledge bases to make CDSS more active. At each site, new knowledge will be discovered and added to knowledge base from (1) new expert knowledge discovered by research, (2) data mining engine connected to local EHR and clinical databases. This action will make CDSS more active by including the most recent knowledge from active databases. Because knowledge base must be in specific domain such as heart diseases, the proposed framework will be distributed with co-operative and integrated knowledge bases. Each knowledge base in each hospital will be in specific domain. At each hospital, CDSS will build patient profile from patient's medical history and current diagnose, and it will use its local knowledge base to make decision. If CDSS cannot take decision by using its local knowledge base, it can send some data to other sites to consult its specialized knowledge bases. Other sites will response by some knowledge that helps CDSS to make more accurate decision. The goal of this paper is to propose a distributed CDSS framework that achieves (1) build co-operative knowledge bases from different domain experts' knowledge and most recent academic researches, (2) Standardize knowledge into XML format before storage, (3) connect data mining engine to EHR and clinical databases to continuously mine the most recent and applicable knowledge and adds it to local knowledge base, (4) CDSS can consult specialized knowledge bases in other institutions for other relevant knowledge, (5) before starting to take decision, CDSS collects all patient EHRs from all sites, integrates it with current diagnose, standardizes it and enters it to the inference engine, (6) assure interoperability by converting all patient data and knowledge in to standard XML format. This way, we build a complete, interoperable, active, distributed and continuously learning CDSS system.

4 Clinical Decision Support System

CDSS are interactive computer programs which are designed to assist physicians and other health professionals [14]. It helps in drug prescription, diagnosis and disease management to improve services and reduce risks and errors. It can check for patient drug allergies, compare drug and laboratory values, evaluate the potential for drug-drug interactions,

suggest drug alternatives, block duplicate orders, suggest drug doses, routes, and frequencies and provide recommendations. Also, CDSS can provide clinical knowledge and best practice standards and guidelines for inexperienced physicians. CDSS must be integrated with EHR and CPOE system which is connected to other HISs (laboratory, radiology, billing, etc). The basic components of a CDSS include medical knowledge base and an inference mechanism (usually a set of rules derived from the experts and evidence-based medicine) and implemented through medical logic modules based on a language such as Arden syntax or using artificial neural network as in figure 1 [15].

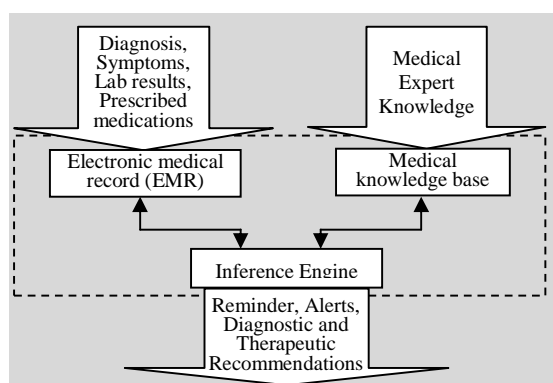


Figure 1: General model of CDSS

CDSS provides recommendations based on the available patient specific data (EHR) and medical facts (knowledge base). It has ten levels of automation ranging from L1 where all decisions made by humans to L10 where computer makes all decisions.

The EMR is continuously updated, so the knowledge bases must be continuously updated by discovered knowledge from domain expert and discovered knowledge from EMR.

5 Proposed CDSS Framework

We assume that EHR architecture and connectivity exists, and we will integrate the distributed CDSS architecture with it. The proposed architecture of CDSS is independent. It does not depend on and does not affect by the architecture of EHR or HIS. Moreover, the architecture is scalable. We can add any number of knowledge bases, EHRs, or clinical databases to the architecture using available standards and technologies. Previous CDSS was separate system from the healthcare systems. This way, it will require physician to manually activate it, log in to it, and reenter redundant patient data. This process will make CDSS not applicable and waste time. Also CDSS will depend on the entered data which may be inadequate or contain

errors. The needed CDSS will be directly integrated with the healthcare system's CPOE component, it will be activated automatically, collect the needed data from patient order, ask for unknown parameters, and make recommendation on time. Figure 2 [16] show the three phases in the decision making process.

Phase 1 (knowledge preparation) uses data mining techniques to extract knowledge from electronic healthcare data and store it in knowledge base. *Phase 2* (knowledge interoperability) takes the patient data that need decision making and translate it in to standard XML form (CDA) and make PMML encoding of the knowledge from knowledge base (KB). *The last phase* takes the previous standardized data and knowledge and makes decision. Figure 3 shows our proposed CDSS framework. It will operate as follows:

5.1 Knowledge Bases Building

The first step is to build the initial KBs. Constructing KBs of the CDSS is a crucial task that determines the success of the CDSS in general [19]. The goal is to collect the medical knowledge from the relevant sources (domain expert, EHR and/or clinical databases, and research), systemize it and represent it in a formal human understandable and computer-interpretable manner. In this framework the three services or components responsible for generating and standardizing knowledge to populate the standard XML KBs are:

- 1- *Knowledge Extraction Module (KEM)*, it is responsible for extract knowledge from domain expert. There are many ways to represent this knowledge.
- 2- *Data Mining Engine (DME)*, it is responsible for mine both EHR and clinical databases.
- 3- *PMML Encoding Module (PEM)*, it employs PMML to encode the generated knowledge in to standard XML based document to achieve interoperability goal between knowledge discovered from different HISs and knowledge from domain experts. The XML schema for each document describes the input data items, data mining algorithm specific parameters, and the final mining results.

The challenges in constructing and maintaining the knowledge bases are numerous. Firstly, for specific domain in one hospital, the KB is built from domain expert's knowledge off-line. KEM can take variety of methods and techniques to build knowledge base [5]. KEM Then passes the knowledge to PMML encoding module to translate it to XML form. In each hospital, CDSS will have specialized KB according to the field of the hospital, and the distributed framework will make these KBs co-operative.

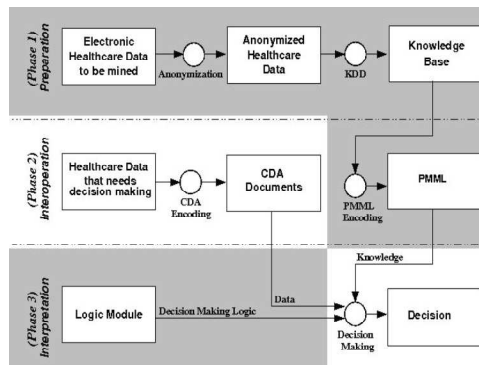


Figure 2: Health care Knowledge management framework. The shaded areas designate off-line parts.

Assuring that KBs are up to date is critical to make CDSS active and continuously learning and therefore applicable, and this can be achieved by:

- (1) Continuously update KBs by new domain expert or research knowledge,
- (2) Applying data mining techniques and algorithms on local clinical databases and EHR to discover hidden and non-trivial patterns and update the KB by these results.

This way, the CDSS will provide the most up to date and the most applicable knowledge. DME has two processes:

- (1) Data Preparation Engine which identifies the task relevant data from clinical databases and HER after triggers from there sources to apply data mining process, removes the healthcare data attributes that can identify a patient or reveal their private data (Anonymization) and performs data selection, cleaning and transformation.
- (2) KDD which performs the actual data mining operations. Finally, the results are assessed in terms of usefulness, validity, and understandability.

EHR is important source for medical knowledge. It contains a longitudinal and history of patient clinical and diagnostic data. This makes EHR a good place for applying data mining and AI techniques. Also EHR attributes are selected carefully which add another advantage. This process is continuous because EHR is updated continuously. Any update to EHR will trigger the DME to discover new knowledge then pass it to PEM to standardize and store it in KB.

Another source of knowledge is the clinical databases because it contains detailed data about patient. The DME is triggered to search for new knowledge in the updated databases as with HER. This way we assure that KB contains the complete, most recent, accurate and applicable knowledge.

The domain expert knowledge and the data mining discovered knowledge are passed to PEM module to be

standardized in XML format and stored in knowledge base.

5.2 CDSS Supporting CPOE

Healthcare personnel use the CPOE for prescription. Previously, the Health Information System (HIS) was depending on the paper-base prescribing or poor or unstructured notes in a separate system (order entry system). Also, the order entry system was collecting only administrative data not medical, clinical or diagnostic data. The needed system will use electronic prescribing system which allow the writing of e-prescribe. Additionally, human errors and mistakes are expected when writing the prescription. With the existence of CDSS integrated with CPOE system, CDSS will not only provide recommendations for treatment, but it also can check for errors or shortage of data and notify physician before proceeding with decision support. There are many methodologies for building user interface for CPOE. It may be a series of questions and answers [18]. Another methodology uses the standard paper-base forms to build data entry templates and adds features relevant to decision support. Web-based order entry forms also can be used.

5.3 Framework Execution Steps

After building knowledge bases the CDSS is now ready to guide and help healthcare personnel. The execution of this framework will work as follow:

- 1- In an on-line operation, Healthcare personnel enters patient Universal ID (UID) which identify the patient nation-wide, and enters subject data or current diagnose (i.e., healthcare data that needs decision making).
- 2- UID passes to HRS (History Retrieval Service) in the local hospital, and travels via a secure network channels to all hospitals.
- 3- Each HRS in each hospital checks whether this patient has an EHR in its hospital or not.
- 4- If the patient has no record then the service returns message indicating that, else there are many methodologies for implementing the service to retrieve patient record. It may be implemented to retrieve the last N visits, visits within specific period, specific disease's related data, etc.
- 5- The returned records will be collected and filtered by Accumulator and Filter service which produce the patient profile.

6- Patient profile is integrated with the current diagnose and entered to CDA Encoding service which standardize the patient medical and diagnoses data into standardized XML-based CDA.

7- The encoded PMML knowledge from local KB and CDA document from CPOE provides the interoperability of knowledge and data in our framework in the sense that CDSS will be independent of the proprietary data format of the involved healthcare providers.

Now we have a complete view about patient's current and previous conditions.

8- The encoded patient profile enters as input to local Knowledge Engine (KE) which make inference of diagnose, determine the correct medicines, etc, as discussed in section 4.

9- KE can be programmed by any AI methodology as artificial neural network. It can access, query, and interpret the data and knowledge that flow from CPOE and KB respectively. Decision making is carried out in 3 main steps, retrieving the right data fields from the data source (CDA); applying the knowledge's models to the data; and eventually taking an action or a set of actions based on the results of this application. For example, if the module was invoked at a decision step in a guideline, it may branch to a specific path; or it may simply display the results in the form of a reminder or an alert.

10- According to the complexity of the problem and according to the specialization of KE, KE may need to consult the other site's KE of its problem if it has shortage in available knowledge or if it is not specialized in this problem. This is done by sending the CDA or specific fields from its site to all or set of other sites that use the same technologies, interfaces, standards, services, and terminologies. All of the helping KEs determine the relevant knowledge and send it to the requesting KE.

11- This way KE will take decision based on the initial physician diagnose, EHRs, and knowledge from its local KB other KBs. Also, it will use KB which contains the most recent knowledge. This way we ease the process of developing KBs because each KB will be specialized in specific domain and KEs will co-operate or consult each other according to patient profile to make the most accurate decision.

12- The final results of the KE will be displayed to healthcare personnel by the Knowledge Representation (KR) module. It will be used to communicate the final results to physician. According to the level of automation in CDSS, the KR may:

- 1- Display recommendation in the form of images, texts, sounds, videos, etc.
- 2- Require physician's decision about the final diagnose and actions. The physician has the choice to refuse, alter or accept the given support. If the physician accepts the support, CDSS will send an order to: the pharmacy to prepare the medicine and give it the treatment policy, laboratory system to prepare for specific tests, radiology system to be ready for some rays tests, etc.
- 3- Request additional data to be entered again into CPOE.

13- The CDSS may make many diagnoses with different probabilities and physician can choose the best. Also, data mining and machine learning can predict the likelihood for any future problem in health of community.

We expect that this framework will provide the most accurate and applicable decision support, and will achieve great integration between HIS and decision support processes. Also, the proposed model is fully automated. The physician only enters the patient UID and initial diagnoses, and CDSS returns decision supports. Moreover, the architecture is component-based. Each component of architecture is pluggable and reusable.

6 Conclusion

In this paper, we proposed a novel knowledge management framework for distributed health care systems that incorporate the knowledge extracted by data mining techniques with knowledge from domain experts with EHR data into health care information systems for decision making support. The model successfully integrates CDSS into the workflow of the HIS. This process fast the physician operations and reduce the level of error. We use many standards as CDA and PMML to achieve the system interoperability and integration to enhance healthcare decision making environment. Our model is fully automated. It only needs the patient universal ID and the physician's initial diagnose, then the system collects all patient EHRs from all hospitals, standardize it, and introduce it to the Knowledge Engine which make intelligent decision support.

This model depends on a set of knowledge bases located in different hospitals. Each is specialized in specific domain and the distributed CDSS architecture facilitates the integration and cooperation of KEs in case of patients who have complex medical or diagnostic problems. KE may send the patient profile or specific data to other sites for consultation.

This model also assures that knowledge base is continuously up to date to allow the CDSS to produce

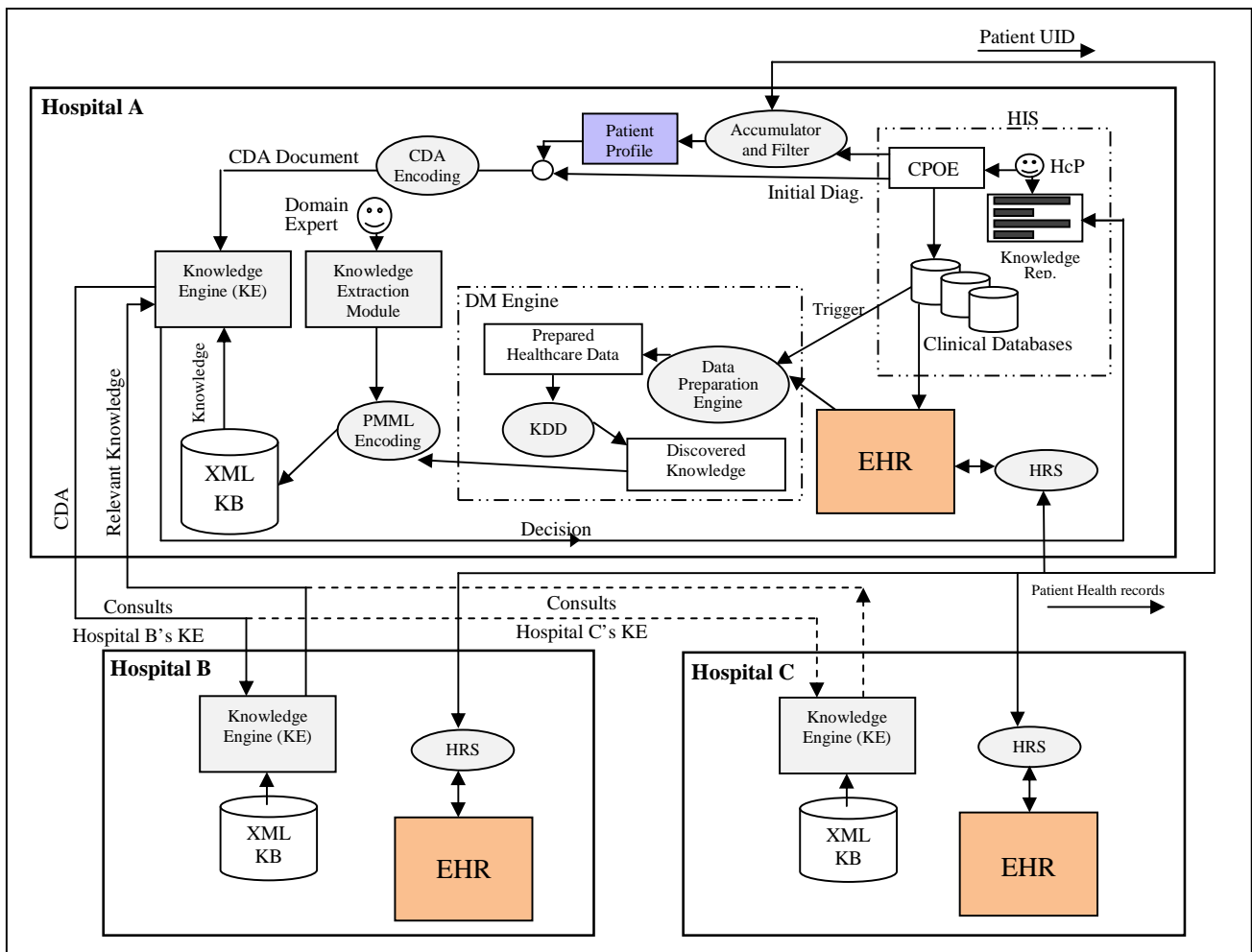


Figure 3: Proposed Distributed CDSS architecture

an applicable recommendations and actions. If the result recommendations are not represented to physician with a correct way, then it will have less benefit. As a result of that, the model has a module to represent results from KE in a meaningful way which allows physicians to make fast and accurate decisions. The next step is to implement this framework.

7 References

- [1] Canadian institute for health informatics (cihi) project. Available: <http://www.cihi.ca/>.
- [2] Hahn, C., Jacobi, S., Raber, D.; "Enhancing the Interoperability between Multiagent Systems and Service-Oriented Architectures through a Model-Driven Approach"; Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conf.; Volume: 2; Page(s): 415; 2010.
- [3] Maciel, R.S.P.; David, J.M.N.; "WGWSOA: A Service-Oriented Middleware Architecture to support Groupware Interoperability"; IEEE 11th International Conference in Computer Supported Cooperative Work in Design (CSCWD); Page(s): 556; 2007.
- [4] Jean Bacon, Ken Moody; "Toward open, secure, widely distributed services"; Magazine Communications of the ACM; Vol. 45; Issue 6; 2002.
- [5] Guilan Kong, Dong-Ling Xu, Jian-Bo Yang; "Clinical Decision Support Systems: A Review on Knowledge Representation and Inference under Uncertainties"; International Journal of Computational Intelligence Systems, Vol.1, No.2 (May 2008), 159-167.
- [6] Bodenreider, Olivier (2004), "The Unified Medical Language System (UMLS): integrating biomedical terminology", Nucleic Acids Research, 32, D267-D270.

- [7] Protégé official web site, Feb 2010. Available: <http://protege.stanford.edu/>
- [8] Terenziani P, Montani S, Bottrighi A et al. "The GLARE approach to clinical guidelines: main features". *Stud Health Technol Inform.* 2004;101:162-6.
- [9] Open Clinical: PROforma, Feb 2010. Available: http://www.openclinical.org/gmm_proforma.html
- [10] Open Clinical: ASBRU, Feb 2010. Available: http://www.openclinical.org/gmm_asbru.html.
- [11] w3c consortium. Available: <http://www.w3.org/XML/>
- [12] Ramakrishnan N., Hanauer D., Keller B.; "Mining Electronic Health Records"; *IEEE Computer Society*; Vol. 43; Issue: 10; PP. 77 – 81; 2010.
- [13] Eugenia G. Giannopoulou, "Data Mining in Medical and Biological Research," ISBN 978-953-7619-30-5.
- [14] Gamberger D., Prcela M., Jovic A., Smuc T., Parati G., Valentini M., Kawecka-Jaszcz K., Styczkiewicz K., Kononowicz A., Candelieri A., Conforti D., Guido R.; "Medical knowledge representation within Heartfaid platform." In *Proc. of Biostec 2008 Int. Joint Conference on Biomedical Engineering Systems and Technologies*, pp.205-217.
- [15] Aleksovska-Stojkowska L. Loskovska S.; "Clinical Decision Support Systems: Medical knowledge acquisition and representation methods"; *Electro/Information Technology (EIT) IEEE International Conference*; on pages: 1; 2010.
- [16] Reza Sherafat Kazemzadeh, Kamran Sartipi; "Interoperability of Data and Knowledge in Distributed Health Care Systems"; *Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP'05)*; Page(s): 230; 2005.
- [17] June Eichner, Maya Das, NORC at the University of Chicago; "Challenges and Barriers to Clinical Decision Support (CDS) Design and Implementation Experienced in the Agency for Healthcare Research and Quality CDS Demonstrations"; *AHRQ National Resource Center for Health Information Technology*; 2010
- [18] Liang Xiao, Gráinne Cousins, Lucy Hederman, Tom Fahey, Borislav Dimitrov; "The Design of an EHR for Clinical Decision Support"; *3rd International Conference on Biomedical Engineering and Informatics(BMEI 2010)*.
- [19] Jeong Ah Kim, InSook Cho, Yoon Kim; "CDSS (Clinical Decision Support System) Architecture in Korea"; *International Conference on Convergence and Hybrid Information Technology (ICHIT)*; Page(s): 700 – 703. 2008.

Simulation of Genetic Regulatory Networks

Rafat Parveen

Address

rafatparveen@yahoo.co.in

Asstt. Professor, Department of Computer Science,
Jamia Millia Islamia ,New Delhi, India.

Abstract: Dizzy is a chemical kinetics simulation software framework. On up gradating this package to simulate the dynamics of complex gene regulatory networks. Using Tauleap simplex and Tauleap complex algorithms, implemented in Java. Procedure have been improved for determining the maximum leap size which accelerates the speed of simulation. This paper focuses mainly on simulating Genetic Regulatory Networks using stochastic methods of simulation and introducing τ to accelerate the speed of simulation.

Keywords: Gene Regulatory Network, Endomesoderm, Sea Urchin, Stochastic Simulation , BioTapestry.

1. Introduction: Simulation is a powerful approach for understanding the complexity of biological systems. Recently, several successful attempts have been made for simulating complex biological processes like gene regulatory networks, metabolic pathways and cell signaling pathways[1][2]. The network models have not only generated experimentally verifiable hypothesis but have also provided valuable-insights into the behavior of complex biological systems. Many recent studies have confirmed the phenotypic variability of organisms to an inherent stochasticity that operates at a basal level of gene expression. Due to this reason, development of novel mathematical representations and efficient algorithms are critical for successful simulation of biological systems. Genetic Regulatory Networks (GRNs) control cellular state form and functions. They are responsible for executing embryonic developmental programs and, changing cellular state and metabolic processes based on environmental conditions. A specific example is the early cell specification process within the sea urchins embryo. GRNs typically involve feedback interactions among multiple genes [2][3]. The situation is frequently more complex in adult organisms, where feedback loops intertwine genetic networks closely.

Signaling and metabolic events change the state of a GRN, which in turn modifies the structure of the upstream [3]. BioTapestry is a software tool for modeling the genetic regulatory networks. The application of Bio-tapestry tool to enable computerized modeling of GRNs, we can model a network consisting of only up to 50 genes. I have upgraded the tool by using Kinetic Logic Model Framework and a number of other algorithms such that a GRN model of more than 186 genes can now be obtained. The output of BioTapestry is used as an

input to Dizzy package in order to simulate the modeled GRNs. On extending the Dizzy software tool by using stochastic Tauleap complex method in order to simulate GRNs. This paper is divided into different sections. Section 2 comprises a brief overview of the Dizzy software system, whereas in section 3 explains the different simulation algorithms and in section 4 the Simulation Methodology is described. In section 5, a new Tau selection procedure is proposed, in sections 6 sea urchins gene expression data and finally in section 7 we have discuss the results and conclusions.

2. Overview of the Dizzy Software System

In this section, we give an overview of the major features of Dizzy, a software framework for simulating the dynamics of complex Genetic Regulatory Network systems. Dizzy provides a collection of simulators for solving the dynamics of a model. Features of Dizzy simulator are:

a. Modular simulation framework: Dizzy employs a modular design in which each simulator is a software unit that conforms to a simple, well-defined interface specification. This architecture facilitates an iterative model development cycle in which the model is analyzed using various simulation algorithms [4].

b. Templates reusable and hierarchical model elements: Dizzy's model definition language permits the definition of reusable, parameterized model elements called templates. This enables the construction of a prepackaged library of templates that can simplify the task constructing a complex model.

c. Multi-step and delayed reaction processes: Dizzy enables the simulation of complex multi-step processes such as elongation and translocation during transcription or translation, through two methods. One may define it as a multi-step reaction process, a reaction process with an intrinsic, phenomenological time delay [5][6].

d. Estimation of steady-state stochastic noise: Dizzy provides a feature for estimating or calculating the steady-state stochastic fluctuations of the species in a biochemical model, requiring only the solution of the deterministic dynamics [7][8].

e. Integrated, graphical, and portable software framework: Dizzy has several important software features including integration with external software tools, a graphical user interface (GUI) and a high level of portability. Many software tools are available for solving the deterministic and stochastic dynamics of complex biochemical networks but not for GRNs. A detailed overview of the most common algorithms for simulating GRNs is presented in section 3. We compare some of the most widely used simulation software tools against a specific list of simulation algorithms and features described above. To the best of our knowledge, Dizzy is the only software tool available that includes all the features enumerated above. In addition, it includes novel implementations of the number of simulation algorithms[9].

3. Simulation Algorithms: A Number of algorithms can be used for simulating the GRNs. These can be divided into two broad categories: a. Deterministic and b. Stochastic Algorithms

a. Deterministic Algorithms:

If no noise or any stochastic variations are present in the process, then we may use Deterministic Algorithms to solve a group of non-linear differential equations. If the system includes both very fast and very slow dynamics, that is some reactions are much faster than others, the system is called stiff. Stiff systems are hard to simulate since the fast dynamics require for short step size and the slow dynamics increase the total simulation time interval. Using a small step size, the simulation of the whole process becomes very slow[10]. Consequently, some numerical algorithms are developed especially for the simulation of this kind of systems. The deterministic algorithms available in Dizzy are listed below.

i. Fifth order Runge-Kutta Method: This method is particularly useful for simulating models in which a derivative function is discontinuous & the step size is adaptively controlled, based on a fourth order error estimation formula. Both relative and absolute error tolerances may be independently specified, as well as the initial step size[11].

ii. Fifth Order RK Fixed: In this method, the differential equations are solved using a finite difference method, with a fixed step-size. The step size is specified by the user, as a fraction of the total time interval for the simulation.

iii. ODE to Java-dopr54-adaptive: In this algorithm control adaptive step-size is used. Implemented by Murray Patterson and Raymond Spiteri.

iv. ODE to Java-imex443-stiff: An implicit-explicit ODE solver with step doubling. Works well for models with a high degree of stiffness.

b. Stochastic Algorithms:

Gene regulation is an inherently stochastic process, which cannot be exactly simulated by deterministic algorithms. In addition, the stochastic algorithms are designed for continuous changes in the state[12]. Some genes in the network may be weakly expressed but the model must handle the exact numbers of genes. In these cases the stochastic simulation methods have to be used. For biological systems involving genes of small

populations, the stochastic simulation algorithm (SSA) derived by Gillespie is an essentially exact procedure for studying noise in gene networks systems[13][14]. However, the computational load of the SSA is often very high when it is applied to simulate large biological systems. Thus, it is imperative to design efficient numerical methods for simulating stochastic Gene Regulation Networks. There are two significant approaches for reducing the computational time of SSA describe in methodology section. In dizzy, the following stochastic algorithms are realized[15]

i. Gibson-Bruck : An algorithm used for simulating the large scale models but are less dynamic.

ii. Gillespie : This algorithm is useful for simple systems with less complexity

iii. Tauleap-Simple: An approximate accelerated stochastic simulator implemented using the Gillespie Tau-Leap algorithm. This implementation is intended for models in which the models are less complex.

iv. Tauleap-Complex: An approximate accelerated stochastic simulator implemented using the Gillespie Tauleap algorithm. This implementation is used for large complex models.

4. Simulation Methodology

In a Genetic Regulatory Networks system, the state vector $\mathbf{X}(t) = (X_1(t), \dots, X_N(t))$, where $X_i(t)$ is the number of gene of species S_i in the system at time t , evolves stochastically because of the inherent random interactions of genes. Random genes interactions give rise to random chemical transmutations in accordance with some specified set of reaction channels $\{P_1, \dots, P_M\}$. The dynamics of genes R_j are mathematically defined by a *propensity function* a_j together with a *state-change vector* $\mathbf{v}_j = (v_{1j}, \dots, v_{nj})$. $a_j(X)dt$ gives the probability that one R_j reaction will occur in state \mathbf{X} during the next infinitesimal time interval dt , and τ_{ij} gives the change in the S_i molecular population produced by one R_j reaction[16].

For simulating the stochastic evolution of $\mathbf{X}(t)$, there exist several exact procedures that actualize every molecular reaction event[17][18]. But efforts to model the complex biological networks inside living cells, where small number of genes can set the stage for major stochastic effects[19], have revealed the need for faster, possibly less meticulous stochastic simulation strategies. The newly proposed "leaping" methodology attempts to sacrifice accuracy for greater speed, and to do so in a way that segues as the system size becomes infinite to standard solution methods for the conventional deterministic reaction rate equation. The " τ -leap method," for instance, tries to leap down the history axis of the system by some chosen time τ that encompasses many reaction events. But theoretical considerations demand that the size of τ be constrained by a *Leap Condition*, which says that the state change in any leap should be small enough that no propensity function will experience a macroscopically significant change in its value. The mathematical rationale for the τ -leap method [19] is the fact that, to the extent that the Leap Condition is satisfied, then given $\mathbf{X}(t) = \mathbf{x}$, the number of times $K_j(\tau, \mathbf{x})$ that genes R_j will be expressed in $(t, t + \tau)$ can be approximated by a *Poisson* random variable:

$$K_j(\tau; \mathbf{x}) \approx \rho_j(a_j(\mathbf{x}), \tau) \tag{1}$$

This is so because the generic Poisson random variable $\rho(a, \tau)$ can be defined as the number of events that will occur in a time τ , given that the probability for an event to occur in the next infinitesimal time dt is adt , where a can be any non-negative constant.

This last requirement is approximately ensured by the Leap Condition, and the consequent approximation (1) allows us to estimate the state change in the leap,

$$X(t+\tau) - x \equiv \Lambda(\tau; \mathbf{x}) = \sum_{j=1}^M K_j(\tau; \mathbf{x}) v_j \tag{2}$$

by simple Poisson sampling[8]. But for this approach to be practicable, we need a reliable, expeditious, and preferably automatic way of determining the largest value of τ , that is compatible with the Leap Condition. A plausible mathematical framing of the leap condition would be require the leap time τ to be such that

$$|a_j(\mathbf{x} + \Lambda(\tau; \mathbf{x})) - a_j(\mathbf{x})| \leq \varepsilon a_0(\mathbf{x}), \quad \forall_{j=1, \dots, M} \tag{3}$$

where τ is a pre-specified error control parameter. But of course, smaller values of τ also imply shorter leaps, and therefore longer simulation times. How can we find the largest value of ε that is consistent with (3) for a specified value of τ ? This would be a reasonably straightforward problem were it not for the fact that the left-hand side of (3) is a random variable (since $\Lambda(\tau; \mathbf{x})$ is a random variable). In any case, we would like to make our determination of τ without performing repeated "trial" leaps, checking after each one to see if condition (3) is satisfied and adjusting τ accordingly, such a post-leap procedure not only would consume much time and many random numbers, but it might also discriminate against statistically rare but nonetheless legitimate large changes in the system's state. In this section we present a new τ -selection procedure that is more robust.

6. Sea Urchins Gene Expression Data

# time	Gene 01	Gene 02	Gene 03	Gene 04	Gene 05	Gene 06	Gene 07	Gene 08	Gene 09	Gene 10	Gene 11	Gene 12	Gene 13	Gene 14	Gene 15
0.000	100.350	99.933	100.070	0.320	0.001	0.001	0.001	0.001	0.604	0.604	0.001	0.001	0.001	0.001	0.604
10.101	108.310	98.566	101.680	7.932	0.337	0.337	0.337	0.619	14.447	14.447	0.619	0.619	0.619	0.619	14.447
20.202	114.820	97.661	103.050	14.929	1.200	1.200	1.200	2.153	26.291	26.291	2.153	2.153	2.153	2.153	26.291
30.303	118.640	97.243	103.910	19.476	2.052	2.052	2.052	3.621	33.530	33.530	3.621	3.621	3.621	3.621	33.530
40.404	122.580	96.920	104.850	24.665	3.311	3.311	3.311	5.729	41.342	41.342	5.729	5.729	5.729	5.729	41.342
50.505	126.180	96.750	105.810	30.029	4.938	4.938	4.938	8.372	48.905	48.905	8.372	8.372	8.372	8.372	48.905
60.606	129.180	96.732	106.710	35.186	6.823	6.823	6.823	11.334	55.673	55.673	11.334	11.334	11.334	11.334	55.673
70.707	131.620	96.839	107.570	40.143	8.938	8.938	8.938	14.548	61.700	61.700	14.548	14.548	14.548	14.548	61.700
80.808	133.550	97.047	108.390	44.907	11.257	11.257	11.257	17.955	67.041	67.041	17.955	17.955	17.955	17.955	67.041
90.909	135.010	97.334	109.200	49.486	13.756	13.756	13.756	21.503	71.745	71.745	21.503	21.503	21.503	21.503	71.745
101.010	136.040	97.681	109.980	53.888	16.415	16.415	16.415	25.146	75.858	75.858	25.146	25.146	25.146	25.146	75.858
111.110	136.670	98.071	110.740	58.120	19.211	19.211	19.211	28.844	79.425	79.425	28.844	28.844	28.844	28.844	79.425
121.210	136.940	98.488	111.490	62.190	22.129	22.129	22.129	32.561	82.487	82.487	32.561	32.561	32.561	32.561	82.487
131.310	136.880	98.919	112.230	66.105	25.150	25.150	25.150	36.266	85.082	85.082	36.266	36.266	36.266	36.266	85.082
141.410	136.520	99.353	112.960	69.872	28.258	28.258	28.258	39.933	87.247	87.247	39.933	39.933	39.933	39.933	87.247
151.520	135.880	99.779	113.680	73.496	31.441	31.441	31.441	43.538	89.015	89.015	43.538	43.538	43.538	43.538	89.015
161.620	135.000	100.190	114.390	76.986	34.684	34.684	34.684	47.062	90.420	90.420	47.062	47.062	47.062	47.062	90.420
171.720	133.900	100.570	115.090	80.346	37.977	37.977	37.977	50.488	91.490	91.490	50.488	50.488	50.488	50.488	91.490
181.820	132.590	100.920	115.780	83.582	41.307	41.307	41.307	53.800	92.255	92.255	53.800	53.800	53.800	53.800	92.255

5. The New Tau-Selection Procedure

The new τ -selection procedure requires us to determine in advance first the M^2 functions

$$f_{jj'}(\mathbf{x}) = \sum_{i=1}^N \frac{\partial a_j(x)}{\partial x_i} \quad j, j' = 1, \dots, M \tag{4}$$

and then the $2M$ functions

$$\mu_j(\mathbf{x}) = \sum_{J'=1}^M f_{JJ'}(x) a_{J'}(x) \quad (j=1, \dots, M) \tag{5a}$$

$$\sigma_j^2(\mathbf{x}) = \Delta \sum_{J'=1}^M f_{JJ'}(x) a_{J'}(x) \quad (j=1, \dots, M) \tag{5b}$$

and then the $2M$ functions

$$\mu_j(\mathbf{x}) = \sum_{J'=1}^M f_{JJ'}(x) a_{J'}(x)$$

This obviously represents some computational overhead, but the task is not quite as daunting as it might at first appear, the functional dependence of a_j on each x_i is typically be very simple often constant, sometimes linear, but rarely more than quadratic. Furthermore, for large systems the matrix v_{ij} will typically be sparse. In any case, with the functions (4) and (5) determined, then given a current state $\mathbf{X}(t) = \mathbf{x}$, the largest τ that is compatible with the Leap Condition (3) is taken to be

$$\tau = \min_{J \in [1, M]} \left\{ \frac{a_0(x)}{\mu_j(x)}, \frac{2a_0(x)}{\sigma_j^2(x)} \right\} \tag{6}$$

Acceptance of this τ -value is, however, subject to the provision that if it is less than a few multiples of $1/a_0(x)$, which is the mean time step for the exact stochastic simulation algorithm.

7. Results and Conclusions

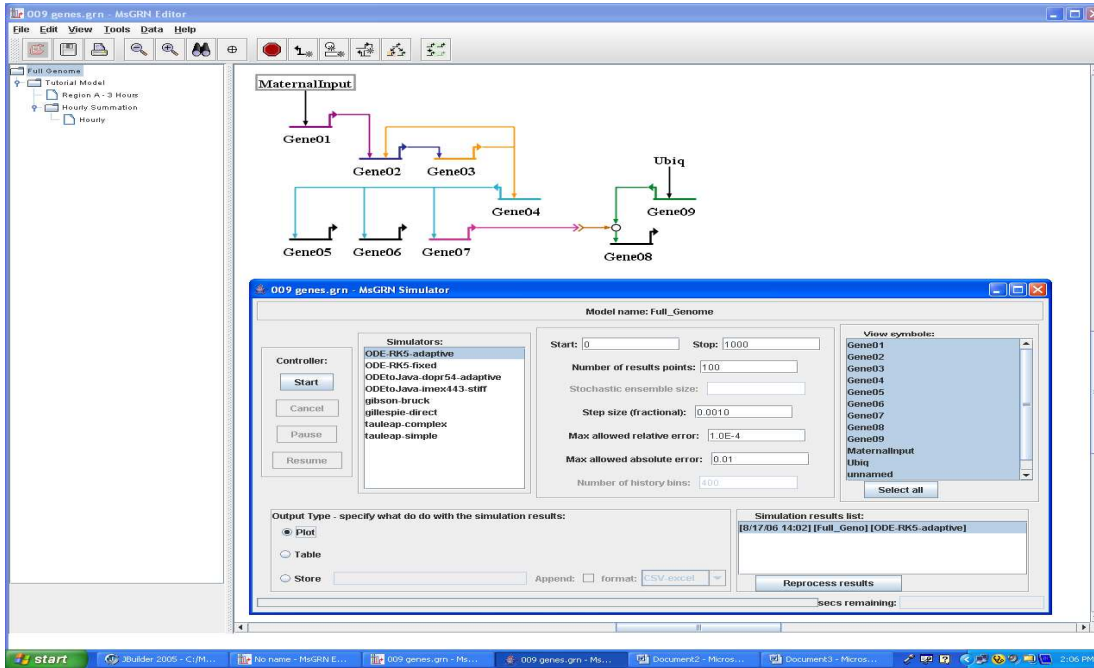


Fig.1 A screen capture of the Dizzy program showing a simulation of a model of genetic regulatory network consisting of different number of genes in sea urchin's embryo.

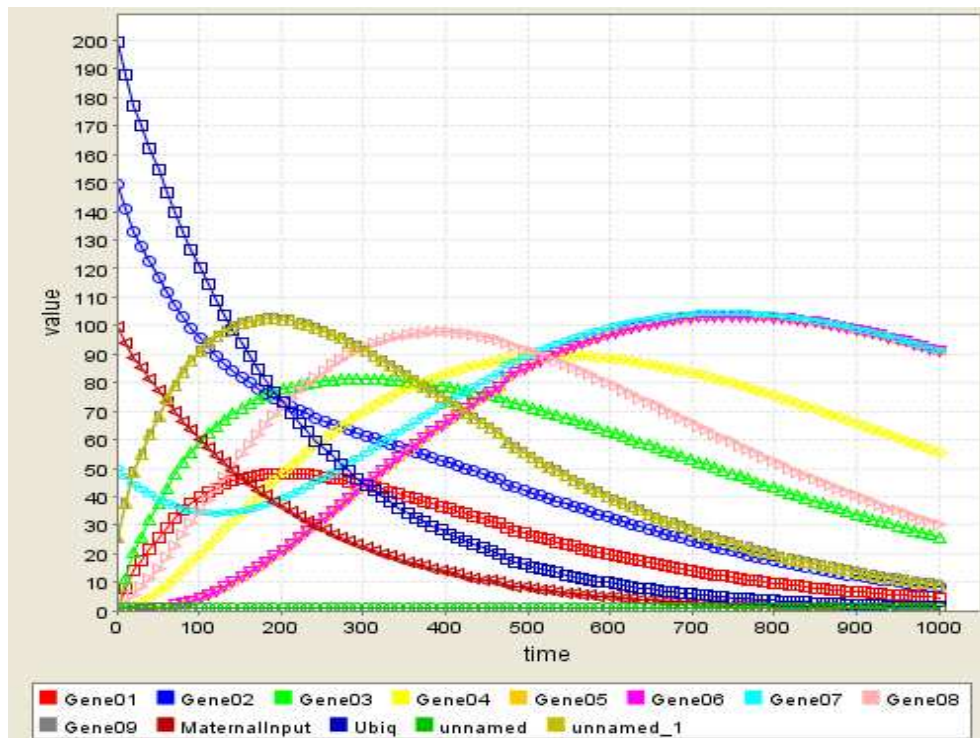


Fig.2 Simulation of GRN consisting of 9 genes of sea urchins embryo

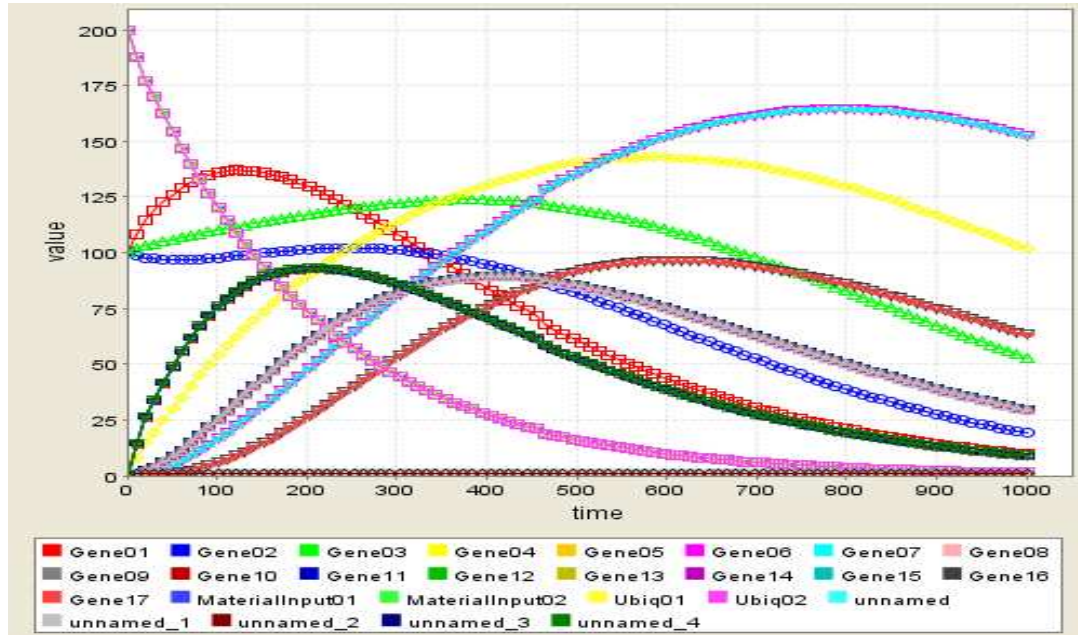


Figure 3. Simulation of GRN consisting of 17 genes of sea urchins embryo

Algorithms	Computational cost	Modeling Knowledge	Speed	Accuracy
Tauleap Simplex	High	Medium	Fast	High
Tauleap Complex	Low	High	Very Fast	Medium

Table1: Comparison of various stochastic simulating algorithms

In this paper we have presented a comprehensive software tool for conducting stochastic simulations of the dynamics of complex gene regulatory networks. The tool is particularly well suited for simulating the dynamics of integrated large-scale genetic, metabolic and signaling networks. In this paper we have implemented & tested various forms of stochastic algorithms and their application to simulation of biological systems. Each algorithm imposes a certain constraint on the computational power, knowledge of the system and input of the numerical parameters. In addition, the algorithms provide different abstractions of the system and produce solutions with very accuracy. Tauleap methods, Tauleap Simplex and Tauleap Complex algorithms are among the fastest simulation algorithms. However, due to various numerical treatments to the algorithms, both Tauleap methods require substantial modeling knowledge to ensure the accuracy of the solutions. Besides that, both the algorithms are efficient algorithms, which increase the

speed of simulation without sacrificing the accuracy of solutions.

As the number of genes in a network increases the network becomes more complex because the connecting lines start criss crossing leads to more complex network. By simulating these genetic regulatory networks we can choose the less complex network that corresponds to the less complex biological system and our aim in systems biology is to find the less complex system so that we can use that, in preventive medicine because gene target prediction becomes easier. If we compare the results obtained in fig.2 and figure 3 which shows that we can easily choose the less complex network because as the number of genes increases the lines in graphs starts overlapping consequently our graphs becomes more complex, that indicates more complex network. In this way we have simulated the GRNs.

References:

- [1]. E. H. Davidson et al., "A genomic regulatory network for development," *Science*, vol. 295, pp. 1669-1678, 2002.
- [2]. S. R. Biggar and G. R. Crabtree, "Cell signaling can direct either binary or graded transcriptional responses," *EMBO J.*, vol. 20, no. 12, pp. 3167-3176, 2001.
- [3]. P. de Atauri, D. Orrell, S. Ramsey, and H. Bolouri, "Evolution of 'design' principles in biochemical networks," *IEE Systems Biology*, vol. 1, pp. 2840, 2004.
- [4]. H. Bolouri and E. H. Davidson, "Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 16, pp. 9371-9376, 2003.
- [5]. P. Guptasarma, "Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of *E. coli*?" *BioEssays*, vol. 17, pp. 987-997, 1995.
- [6]. M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, pp. 1183-1186, 2002.
- [7]. E. M. Ozbudak et al., "Regulation of noise in the expression of a single gene," *Nature Gen.*, vol. 31, pp. 6973, 2002.
- [8]. W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins, "Noise in eukaryotic gene expression," *Nature*, vol. 422, pp. 633-637, 2003.
- [9]. D. T. Gillespie, "Approximate accelerated stochastic simulation of chemically reacting systems," *J. Chem. Phys.*, vol. 115, no. 4, pp. 1716-1733, 2001.
- [10]. E. L. Haseltine and J. B. Rawlings, "Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics," *J. Chem. Phys.*, vol. 117, pp. 6959-6969, 2002.
- [11]. D. T. Gillespie and L. R. Petzold, "Improved leap-size selection for accelerated stochastic simulation," *J. Chem. Phys.*, vol. 119, no. 16, pp. 8229-8234, 2003.
- [12]. J. Puchalka and A. M. Kierzek, "Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks," *Biophys. J.*, vol. 86, pp. 1357-1372, 2004.
- [13]. T. R. Kiehl, R. M. Mattheyses, and M. K. Simmons, "Hybrid simulation of cellular behavior," *Bioinf.*, vol. 20, no. 3, pp. 316-322, 2004.
- [14]. C. V. Rao and A. P. Arkin, "Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm," *J. Chem. Phys.*, vol. 118, no. 11, pp. 4999-5010, 2003.
- [15]. K. Vasudeva and U. S. Bhalla, "Adaptive stochastic-deterministic chemical kinetic simulation," *bioeng.washington.edu*, 2003.
- [16]. D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J. Comp. Phys.*, vol. 22, pp. 403-434, 1976.
- [17]. D. A. McQuarrie, "Stochastic approach to chemical kinetics," *J. Appl. Prob.*, vol. 4, p. 413, 1967.
- [18]. M. A. Gibson and J. Bruck, "Efficient exact stochastic simulation of chemical systems with many species and many channels," *J. Phys. Chem. A*, vol. 104, pp. 1876-1889, 2000.

TTCS: Three-Dimensional and Two-Dimensional Compound Structure Search Online Tools

Hongzhi Li, Haiqing Li, Edward Lee and Yate-Ching Yuan

Bioinformatics Core Facility, Department of Molecular Medicine, Beckman Research Institute,
City of Hope Medical Center, Duarte, CA 91010, USA

Abstract - TTCS is a web based online tool to perform Two- and Three-Dimensional Compound Similarity structure search. It performs compound similarity search from several in-house compound libraries integrated with millions of public available compound libraries such as NCI DTP database, NCI FDA-approved Drug database, and DrugBank, etc. TTCS offers user friendly web interface for 2D and 3D structure query with features such as tracing user's search history, job queuing, and visualization. The website is specially designed for researches' need in bioactive chemical discovery and lead optimization for drug discovery.

Keywords: similarity search, drug discovery, ligand-based drug discovery, three-dimensional similarity, cheminformatics

1 Introduction

Computer-assisted drug design can reduce the number of potential drug candidates from millions of compounds to tens for experimental synthesis and validation. There are mainly two types of approaches, structure-based and ligand-based drug design. Compound structure similarity search is widely used cheminformatics tool to identify clusters of structural similarity compounds for lead optimization in ligand-based drug discovery. TTCS offers the compound similarity searching measured by calculating the Tanimoto coefficient of the lead compound two-dimensional (2D) fingerprints[1], as well as comparing 3D pharmacophore fingerprints[2] and comparing with integrated large bioactive compound libraries with purchase information.

Structure similarity search by using 2D molecular fingerprint is one of the common methods utilized for comparing large compound libraries in *silico*. TTCS integrated a suite of compound structure similarity search tools which are effectively parallelized running on a scalable ScaleMP computation server to search millions compound libraries with reasonable accuracy. The similarity can be determined by calculating the Tanimoto coefficient of the 2D fingerprints and 3D pharmacophore patterns. For example, DiverseSolution BCUT descriptors from Tripos capture the properties of the molecule in chemical space, including

topological or distance information, such as hydrogen bonds between atoms in a molecule. Many BCUT descriptors are calculated for a compound library/database, but only the most important 3-6 descriptors that are most useful for the diverse chemical space of the library. The pre-selected chemical space determined as the nearest neighbor representing a clusters of structural similarity compounds for lead optimization in ligand-based drug discovery.

Several public available compound databases, such as NCBI PubChem[3], UCSF ZINC[4] or NCI DTP integrates 2D fingerprint similarity search engine on their online tools. However, none of them provides both 2D and 3D pharmacophore fingerprints similarity search tools due to computational limitation. In this study, we present an online comprehensive similarity searching tools Two- and Three-Dimensional Compound Similarity structure search (TTCS) to provide users with more searching functionality to efficiently narrow down clusters of potential leads from several in-house commercial compound libraries integrated with large public available compound libraries such as NCI DTP database, NCI FDA-approved Drug database, and DrugBank database for further validation in ligand based drug discovery.

2 Methods and Design

2.1 Methodology

The 2D fingerprint similarity of two compounds is determined by the Tanimoto coefficient:

$$T_c = \frac{N_{ab}}{N_a + N_b - N_{ab}}$$

where N_a is the number of the bits in the 2D fingerprint of the first molecule, N_b is that of the second molecule, and N_{ab} is the number of common bits for both molecules. The chemistry space for 3D pharmacophore fingerprint searching is selected from the best 4 BCUT descriptors that provide the best separation of the compound library. The low-dimensional BCUT chemistry space is divided into cells and the compounds in the same cell of the target compound are selected as the similar compounds.

The screenshot displays the TACS web application interface. On the left, a 'Jobs' sidebar shows a 'Ligand Pipeline' with folders for 'In Queue (0)', 'Processing (0)', and 'Completed (5)'. Under 'Completed (5)', there are five entries: 'STAT3', 'test', 'ttt', 'test', and 'test'. The main area shows a 'Result List' for 'Job ID: 273'. The list has a table with columns 'Result ID', 'GScore', 'GRank', and 'EXTREG'. The table contains five rows, with the third row highlighted. Below the table, 'Result ID: 3' is selected, and its chemical structure is displayed. The structure is a complex molecule consisting of a pyridine ring, a benzyl group, and a substituted coumarin core.

Result ID	GScore	GRank	EXTREG
1			
2			CMP-090008
3			CMP-130913
4			CMP-083020
5			CMP-106537

2.2 Web Design

The TACS website is specially designed with a user friendly web interface for researchers to perform effective 2D and 3D structure similarity query to explore large chemical libraries. The parallelization of the 2D and 3D similarity algorithms runs as back-end on a scalable ScaleMP HP LINUX servers. As shown in the figure, the web application can trace user's search history, summary status for the batch job and display the output compounds in image. It also provide results downloading in different format, email results to the user, and many other features.

3 Conclusions

TACS, a user-friendly online compound similarity tool that combines the 2D fingerprint searching and 3D pharmacophore fingerprint searching for lead optimization in ligand based drug discovery. It will be available to public soon for users' feedbacks.

4 References

- [1] Peter Willett. "Similarity-based virtual screening using 2D fingerprints", *Drug Discovery Today*, Volume 11, Issues 23-24, 1046-1053, December 2006
- [2] Brett R. Beno and Jonathan S. Mason. "The design of combinatorial libraries using properties and 3D pharmacophore fingerprints", *Drug Discovery Today*, Volume 6, Issue 5, 251-258, March 2001
- [3] Bolton E, Wang Y, Thiessen PA, Bryant SH. "PubChem: Integrated Platform of Small Molecules and Biological Activities". Chapter 12 in *Annual Reports in Computational Chemistry*, Volume 4, American Chemical Society, Washington, DC, April 2008
- [4] Irwin JJ and Shoichet BK. "ZINC—a free database of commercially available compounds for virtual screening", *J Chem Inf Model*, Volume 45, Issues 1, 177-182, Jan-Feb, 2005

Cyberinfrastructure: A Case Study of IT Infrastructure for Next Generation Bioinformatics and Computational Biology

Haiqing Li and Yate-Ching Yuan

Bioinformatics Core, Department of Molecular Medicine, Beckman Research Institute,
City of Hope Medical Center, Duarte, CA 91010, USA

Abstract - *This presentation will share our experiences in establishing cost-effective translational bioinformatics platforms using an integrated cyberinfrastructure to support high-throughput data analysis, management, and integration in order to streamline analysis pipelines for predictive, preventive, personalized and participatory medicine. In this case study, we present the architecture of cyberinfrastructure and the challenges we face during design, deployment and management of cyberinfrastructure. We also show how new computational technologies, such as GPGPU and Cloud Computing, can help to speed up the bioinformatics analysis and data management. At the end, we discuss the future directions of IT infrastructure to support bioinformatics and computational biology.*

Keywords: Cyberinfrastructure, Bioinformatics, Computational Biology, Infrastructure, charge back, GPGPU, Cloud Computing

1 Introduction

IT infrastructure is the essential foundation for the bioinformatics and computational biology. The different types of biological computing have different computer utilization requirements[1]. The next generation bioinformatics and computational biology needs scalable and flexible IT resources to support the analysis of the next generation high throughput data, such as next generation sequencing, mass spectroscopy, HTS, high content screening technologies, etc. In spite of the broad spectrum of growing fields of OMICs technologies, the traditional IT data center has not yet evolved to provide adequate scalable cyberinfrastructure. In this study, we will demonstrate our efforts to integrate new parallelization approaches of using shared memory ScaleMP and GPGPU servers to leverage the growing needs of IS&T support.

2 Cyberinfrastructure

Cyberinfrastructure is a project to design and deploy an IT infrastructure for next generation bioinformatics and computational biology in a national comprehensive cancer

research center. The motivation of Cyberinfrastructure is to establish the IT infrastructure to support the integration bioinformatics of genomics, proteomics, cheminformatics, imaging, animal study, to enable our support to translational research.

2.1 Architecture

The architecture of Cyberinfrastructure (figure 1) includes two layers. The first layer is the system management and usage monitoring. The second layer is the IT infrastructure, which includes three components: the internal IT, scientific grid, and external cloud computing[2]. The internal IT is the core of cyberinfrastructure to support the on-demand high performance computing with several thousands of processors connected with Petabyte tiered of disk storage connected with infinity band network to processing TBs raw data generated from scientific high throughput instruments. We adopt several new technologies within the internal IT, which include the private cloud for application virtualization, high performance computing system using GPGPU and cluster technology, shared TBs memory computation, the tiered cluster storage system, high performance network system using Infiniband, and integrated lab information management system (LIMS). The second component of cyberinfrastructure is the scientific grid resources, which plays an important role in computational biology[2]. Cyberinfrastructure provides the interface to connect the scientific grid resources. The last component is external cloud computing, which steadily grows in demand for collaborative research by integrating large open source database such as EBI, NCBI, and UCSC Genome browser, etc. The challenges are to reconcile various of high throughput data analysis workflow in the area of genomics, proteomics, imaging, cheminformatics, small animal studies, and translational research and provide the computational and storage on demand Cloud computing for collaborative research. We have evaluated several Pay-As-You-Go external cloud services such as Amazon and Penguin-On-Demand for our current needs. We hope to establish a global platform that we can dynamically and cost effectively to support multidisciplinary translation research.

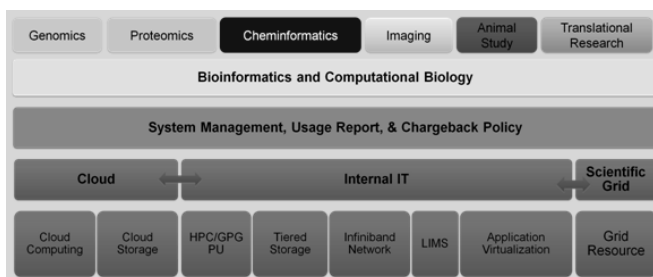


Figure 1. Architecture of Cyberinfrastructure

Figure 2 shows some highlights of the current deployment of Cyberinfrastructure at City of Hope. We adopt several new IS&T (information system and technology) technologies, which include high demand of CPU and memory intensity SMP system, virtual SMP system, high performance computing system using GPGPU and beowulf cluster technology, the tiered Isilon storage system, application virtualization using CITRIX and VMware, and integrated lab information management system (LIMS). It provides a powerful bioinformatics platform for the cancer research at City of Hope. Figure 3 shows the benchmark of the GPGPU cluster system with different GPU/CPU ratio configurations and how it speeds up Molecular Dynamics simulation for just few days running time that took months using super computational center resource.

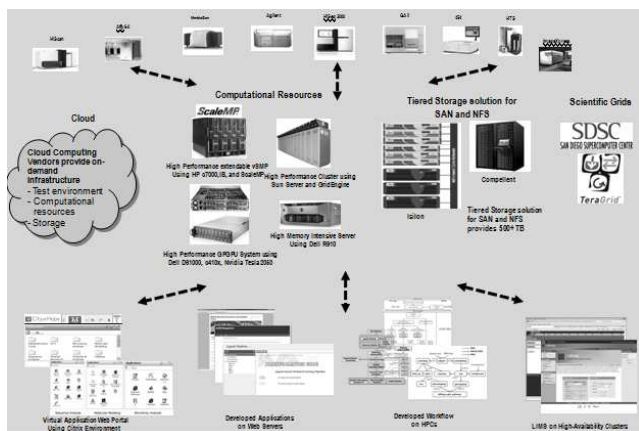


Figure 2. Current development of Cyberinfrastructure

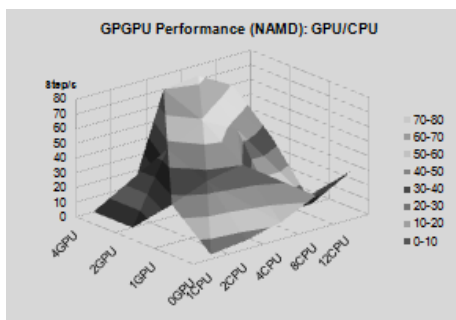


Figure 3. Benchmark of GPGPU

2.2 User access and Charge Back

As a core service, Cyberinfrastructure is open to all bioinformatics core subscribers. Subscribers can access cores Cyberinfrastructure resources, which include high performance servers, large scale tiered data storage, CITRIX web portal, and high performance workstations on campus. Tiered subscriber fee schema was established and fits all different type PIs and their research projects. Usage Metrics reports help PIs to cost effectively strategize their resources, and help administration team to strategic planning IS&T infrastructure needs.

3 Conclusions and Discussions

Strategic Planning for IS&T infrastructure is critical to support modern high throughput technologies such as next generation sequencing, high throughput screening, imaging, and high content screening. This presentation share our experiences in establishing cost-effective translational bioinformatics platforms using an integrated cyberinfrastructure to support high-throughput data analysis, management, and integration in order to streamline analysis pipelines for predictive, preventive, personalized and participatory medicine.

4 References

- [1] Eric Jakobsson. "Specifications for the Next-Generation Computational Biology Infrastructure," *CTWatch Quarterly*, Volume 2, Number 3, August 2006. <http://www.ctwatch.org/quarterly/articles/2006/08/specifications-for-the-next-generation-computational-biology-infrastructure/>
- [2] Brian Hayes, "Cloud computing". *Commun. ACM* 51, 7 (July 2008), 9-11
- [3] El-Ghazali Talbi , Albert Y. Zomaya, "Grid Computing for Bioinformatics and Computational Biology", Wiley, 2007

