SESSION

GENE EXPRESSION AND REPRESENTATION, MICROARRAY, SEQUENCING, ALIGNMENT, AND RELATED BIOINFORMATICS ISSUES

Chair(s)

TBA

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Investigating the Stability of Housekeeping Genes Using Microarray Datasets

Fariba Nosrati, Ayse Bener
Data Science Lab, Department of Mechanical & Industrial Engineering, Ryerson University, Toronto, Canada
Asli Uyar
Department of Computer Engineering, Okan University, Istanbul, Turkey

Abstract*— *Results of microarray experiments would be considered reliable if the non-biological variations, which are introduced into the data during the multiple stages, are as low as possible. The elimination of technical variations is mainly performed during data normalization process. In the popular Housekeeping Gene normalization method genes designated as "housekeeping genes" have been used as internal reference genes under the assumption that their expression is stable and independent of experimental conditions. However, verification of this assumption is rarely performed. This study attempts to identify the suitable housekeeping genes for normalization purpose in the reproductive field for mice. The stability of housekeeping genes was investigated and it was confirmed that housekeeping genes will have different expression stability across different experimental conditions. Our results show that the most stable genes in this field are: GAPDH, TBP, ACTB and PPIA.

***Keywords*— Microarray, Housekeeping Genes, Stability, RMA, Normalization**

1. INTRODUCTION

Microarray technology is used for genome-wide transcription profiling [1]. Since microarray technology works on the whole genome expression simultaneously, it gives better spectra of interactions between thousands of genes in comparison with other gene expression analysis methods [1]. One of the main results in microarray analysis is finding the list of genes that are differentially expressed under different conditions [2].

In differential expression analysis the raw microarray data is preprocessed and then analyzed. Since microarray analysis is a multi-stage procedure, one of the main challenges is that variations could be introduced into the data in different stages. This may lead to less analytical precision because of the undesired variations. The preprocessing stage aims to eliminate these variations by removing the background noise and normalization [3]. Robust Multichip Average (RMA) is a common preprocessing method that includes background adjustment, quantile normalization, and probe set summarization stages [4].

Another popular normalization method is normalization against internal controls or Housekeeping Genes (HKG) [4]. HKG are the genes in basic cellular functions regardless of tissue or organism. They are assumed to be expressed uniformly in all cells at different conditions [5-6]. However, different studies have shown that popular well known HKG may show variations in their expression at different conditions [7-10]. So, when using HKG normalization method, the internal reference with higher expression stability across different experimental conditions should be identified.

This study attempts to identify the suitable HKGs for normalization purpose in the reproductive field for mice. Thus, the stability of HKGs was investigated. The structure of this paper is as follows: section II describes the background and motivation of the research. The methodology is introduced in section III. In section IV the results are presented and discussed. The threats to validity of the results are mentioned in section V and finally section VI concludes and discusses future directions.

2. BACKGROUND AND MOTIVATION

Microarray studies had a great impact on gene expression analysis and led to evolution from single-gene to whole-genome investigation [11]. Thus, high-throughput microarray procedures are one of the popular transcriptomic investigations nowadays [11].

2.1. Microarray Experiments

To perform a DNA microarray procedure, single-stranded polynucleotide probes which are prepared and fixed on a solid two-dimensional array, are prepared. A microscopic spot on the array corresponding to a specific mRNA is attached to each probe. Because of having a large number of probes on a single chip, the expression analysis of thousands of known genes can be performed simultaneously. The procedure continues with multiple steps including: labeling, hybridization, scanning, and data analysis [12]. In this section, these steps are explained with regards to Affymetrix, which is one of the most common microarray platforms [11]:

- The isolation of RMA and reverse transcription of mRNA is performed.
- Amplification and labeling of cRNA will be done.
- The fragmented cRNAs are hybridized on the array.
- After washing the array, the chip is illuminated by laser light. The more complemented to those labeled cRNAs, the more fluorescence will be emitted.
- The fluorescence intensity of each probe is captured by scanners into an image and these intensities would be an estimate of gene expression quantity.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

- From this stage on, the raw microarray data from the image files is available for preprocessing and differential expression analysis.

2.2. Need for Normalization

The entire microarray experiment is prone to errors and non-biological variations due to its multi-stage procedure. In order to have more accurate and reliable expression analysis results these unwanted variations should be reduced as much as possible [13]. One of the major steps performed in preprocessing of microarray data analysis is normalization of raw data that specifically aims in reduction of the variations. The main goal of any normalization method is to eliminate non-biological or technical variations which have been introduced into data from different sources. Plenty of methods and algorithms have been proposed during past decade, each of them has its own strengths and limitations [4]. It is important to choose a method that optimizes the performance in conducting any kind of microarray analysis [2].

2.3. RMA Normalization

One of the most popular preprocessing methods is Robust Multi-Array Analysis (RMA) algorithm [4]. This method consists of three steps: (1) Background adjustment, (2) Quantile normalization and (3) Summarization. It uses the convolution of signal and noise distributions in order to adjust the background fluorescence. RMA algorithm is based on Quantile normalization to remove the technical variations. Quantile method aims to make an identical distribution of probe intensities of a set of arrays [4]. This algorithm first adjusts the arrays for background noise on a raw intensity scale, and then the perfect match probes are corrected for background. After that, the intensity values are transformed to log 2 values and normalized by quantile algorithm and finally summarized [4].

The biological assumption is that a treatment would result in decrease or increase of the expression level in only a limited number of genes and the other genes' expression would remain stable [14]. In addition, it is supposed to have equal amounts of RNA on each array; therefore the sum of all expressions among the samples of one condition should be the same. As the RNA amount is not completely under control and these assumptions might not be met, in measuring the actual expression values, use of a control would be helpful [14].

2.4. Housekeeping Gene Normalization

Another popular normalization method is normalization using internal controls or HKGs. In this method, gene expression levels are normalized against a stably expressed reference [15]. HKGs are the genes in basic cellular functions with hypothetically stable and consistent expression values in all cells regardless of tissue or organism [6]. These genes are assumed to be expressed uniformly in all cells at any differentiation stage, tissue type, developmental stage or any external signal [5-7]. Therefore these genes could be perfect candidates as internal controls

[10, 16-19].

In normalization against HKGs using multiple HKGs proposed to be more accurate normalization algorithm [10]. The overall algorithm of normalization against HKG is [10]:

- Select the HKGs.
- Calculate the normalization factor per array by using the geometric mean of chosen genes' expressions.
- Divide all the expression values of the array by the normalization factor.

A major challenge facing these studies is the design of experimental controls that will permit comparison of quantitative expression profiles [20]. In many studies the popular s (such as ACTB, GAPDH, etc.) are selected under the assumption that their expression is stable and independent of experimental conditions However, verification of this assumption is rarely performed [7]. Furthermore, different

Table 1- EBI-EMBL datasets used in the current investigation

Experiment	Description	Type	Organism
E-GEOD-46875	Association of maternal mRNA with the spindle in mouse oocytes	other, transcription profiling by array	Mus musculus
E-GEOD-45668	The presence of the Y-chromosome, not the absence of the second X-chromosome, alters the mRNA levels stored in the fully grown XY mouse oocyte	transcription profiling by array	Mus musculus
E-GEOD-35106	Polysome-bound mRNA during oocyte maturation	transcription profiling by array	Mus musculus
E-GEOD-17985	Transcription profiling by array of mouse Dicer-deficient ooctyes	transcription profiling by array	Mus musculus
E-GEOD-5668	Transcription profiling of immature germinal vesicle stage oocytes with mature metaphase II oocytes to identify and characterize the changed and stable transcripts during mouse oocyte maturation.	transcription profiling by array	Mus musculus
E-MEXP-1146	Transcription profiling of mouse in vivo matured MII oocytes and fully in vivo grown germinal vesicle oocytes to identify gene transcripts linked to epigenetic reprogramming	transcription profiling by array	Mus musculus

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

studies have shown that popular well known HKG may show variations in their expression at different conditions [7-10, 20-21].

Thus, as the first step of this process, the most stable genes with minimum variation need to be identified and selected. Large-scale expression data profiling is used for this purpose to examine the expression values of genes and identify the stable genes across different conditions [22]. Different statistical algorithms and software bundles such as GeNorm [10], NormFinder [8] and BestKeeper [23] are tools to identifyHKGs.

In GeNorm method, the genes with initial expression stability (M) values within the recommended range ($M < 1.5$) are considered as normalizer [10]. In GeNorm algorithm a pair of HKG is chosen based on the M-value of multiple candidate genes. M-value is an estimation of pairwise variation for each gene. [10].

NormFinder is an algorithm for identifying the optimal normalization gene among a set of candidates [8]. It ranks the set of candidate normalization genes according to their expression stability in a given sample set. The algorithm is rooted in a mathematical model of gene expression and uses a statistical framework to estimate not only the overall expression variation of the candidate normalization genes, but also the variation between sample subgroups of the sample set [8]. Also, NormFinder provides a stability value for each gene, which is a direct measure for the estimated expression variation enabling the user to evaluate the systematic error introduced when using the gene for normalization [8].

BestKeeper determines the best suited standards, out of ten candidates, and combines them into an index. The index can be compared with further ten target genes to decide, whether they are differentially expressed under an applied treatment. When the BestKeeper index was constructed, it can be applied as an expression standard in the same way like any single HKG [23].

As it was mentioned, in using HKG normalization method, identifying the internal reference with higher expression stability across different experimental conditions, may be a challenge [7]. In an attempt to identify genes that are expressed at constant levels in the field of reproductive biology, in this study the stability of HKGs in oocyte samples were investigated by using NormFinder algorithm [8].

3. METHODOLOGY

Using the public data repository from European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) website [24], the study was narrowed to oocyte samples analyzed by transcriptional profiling using arrays in mouse. As shown in Table 1, there were 6 data sets available with raw data for further investigation [25-30]. Figure 1 illustrates an overview of the research methodology.

After choosing the datasets, the arrays of 6 available datasets were normalized by RMA normalization. This method is performed with Quantile algorithm that aims to make an identical distribution of probe intensities of a set of arrays. RMA normalization was performed by MATLAB software using standard bioinformatics toolbox and the expression analysis of the genes after preprocessing was also performed by MATLAB software (R2013b).

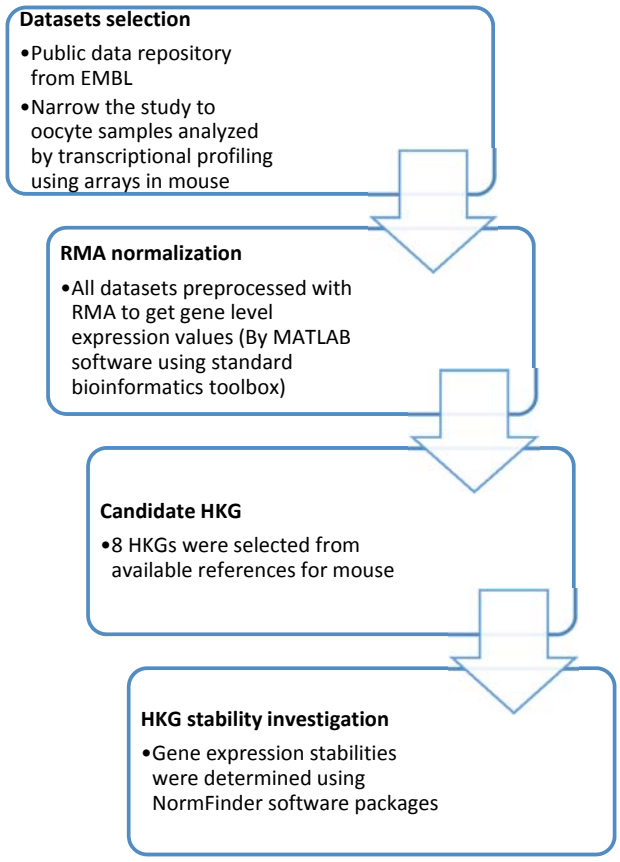


Figure 1- Overview of research methodology

The quantile normalization for a set of data vectors is performed as follows [4]:

- Having n arrays with length p, making matrix X with dimensions $p \times n$, in which each column is an array;
- Sort X to have X_{sort}
- Calculate the average across rows of X_{sort} and substitute each element of row with the average to have X'_{sort}
- Rearrange each column of X'_{sort} to have the same ordering as the original X and the result will be $X_{normalized}$.

Then, 8 candidate HKGs were selected based on the studies in the literature for mouse [31] and their inter-group variances, intra-group variances and their stability were studied. In this research, the stability of the following genes was investigated: GAPDH, ACTB, TUBB5, PPIA, B2M, PGK1, TBP and GUSB.

Using NormFinder methodology and software package [8], the intragroup variation, intergroup variation and the stability value for each candidate gene were calculated. There are some requirements for using NormFinder algorithm:

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

1) The validity of the approach is related to the number of samples and candidates analyzed. The more samples and candidates be used, the estimates would be better. The sample set should minimally contain 8 samples/group [8]. It should be mentioned that two of the investigated datasets (E-GEOD-45668 and E-GEOD-5668) had 6 samples/groups.
2) The number of candidates should be at least 3 for technical reasons, but 5–10 are recommended [8]. Since in this study 8 candidate genes were investigated this requirement was fulfilled.
3) It is also required that the candidates are chosen from a set of genes with no prior expectation of expression difference between groups. This requirement is used to say that the average expression level is approximately the same in the different groups. Therefore, instead of assuming the individual candidate genes to show no systematic intergroup variation, it is assumed that the average of the candidate genes to show no systematic variation [8]. The candidate genes in this study were selected from previously published references as HKGs and this fulfills the requirement.

4. RESULTS AND DISCUSSION

After estimation of both the intra-group and inter-group variations, they are combined into a stability value by NormFinder software package, which adds the two sources of variation and thus denotes a practical measure of the systematic error that will be introduced by using the gene for as a normalizer. NormFinder ranks the candidate genes according to their expression stability. Table 2 summarizes the results for the most stable genes and the best combination of genes in each of the experiments.

As it can be seen in table 2, the most stable gene in each experiment may be different and thus there is no one stable HKG for all conditions. Even when the study is completely narrowed to reproductive biology, in mouse and for oocyte samples, the results are not the same for different conditions and different experiments. Overall, the most stable genes in oocyte samples of mouse were identified as follows: GAPDH, TBP, ACTB and PPIA.

Plotting the inter-group variances will show the effect of using any of the candidates for normalization. Our aim is to look for a candidate gene with an inter-group variance as close to zero as possible. To obtain confidence intervals on the inter-group variances, the average of the intra-group variances will be plotted as error bars to inter-group variances [8]. We still aim that having a candidate gene with an inter-group variation as close to zero as possible, and at the same time having as small errors bars as possible. Therefore a candidate is found this way, and it will be the top-ranked candidate picked by NormFinder. The plots for all the 6 datasets are presented in Figure 2.

As it is illustrated in Figure 2, there are cases that the genes show small intergroup variations. Since their intra group variations are high the overall stability value would be high and thus it is not suitable to be used as internal control for normalization. This can be seen for GUSB in E-GEOD-46875 experiment or for GUSB in E-GEOD-5668 dataset.

It is also obvious based on Figure 2 that both inter-group and intra-group variations depend on the experiment and its conditions and vary in different situations. This confirms that one cannot choose the internal control genes based on popular previously published HKGs and the need for analysis of gene expression stability and choosing the most stable genes for a specific condition and then use them as normalizer.

Table 2- Result of NormFinder for the most stable gene (with the lowest stability value) and the best combination in each experiment

Experiment	Most Stable Gene	Best Combination
E-GEOD-46875	PPIA	PPIA and TUBB5
E-GEOD-45668	TBP	GAPDH and TBP
E-GEOD-35106	GAPDH	GAPDH and PGK1
E-GEOD-17985	TBP	GAPDH and TBP
E-GEOD-5668	GAPDH	ACTB and TBP
E-MEXP-1146	ACTB	B2M and TBP

5. THREATS TO VALIDITY

Threats to validity of this research include external, internal, construct, and statistical threats:

- *External validity:* This study has been performed on 6 publicly available datasets. These data sets may not be a good representative of the population and this may be a threat to external validity of the investigation. Also the study was narrowed to reproductive biology and this could be an external threat to validity of the results. Investigating other systems and genomes may lead to different results. Another threat could be the choice of organism; in this study the datasets were chosen from mouse genome and analysis on other organisms may have different results. Another threat to external validity of this study may be due to the type of microarray experiments which was RNA assay. Other microarray types may show different outcomes.
- *Internal validity*: There may be inherent uncertainties for our experiment such as unknown biological factors that would affect the cause and effect relationships. Also, it could not be assured that other extraneous factors were under control because the data were imported from online datasets.
- *Construct validity:* The procedures in this study were well defined and thus it is fulfilled.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

- *Statistical validity:* One threat could be using RMA normalization for the preprocessing stage and changing preprocessing method may lead to different results. Also in this Study NormFinder methodology was used to evaluate the stability of the genes and using other methodologies may show different outcomes.

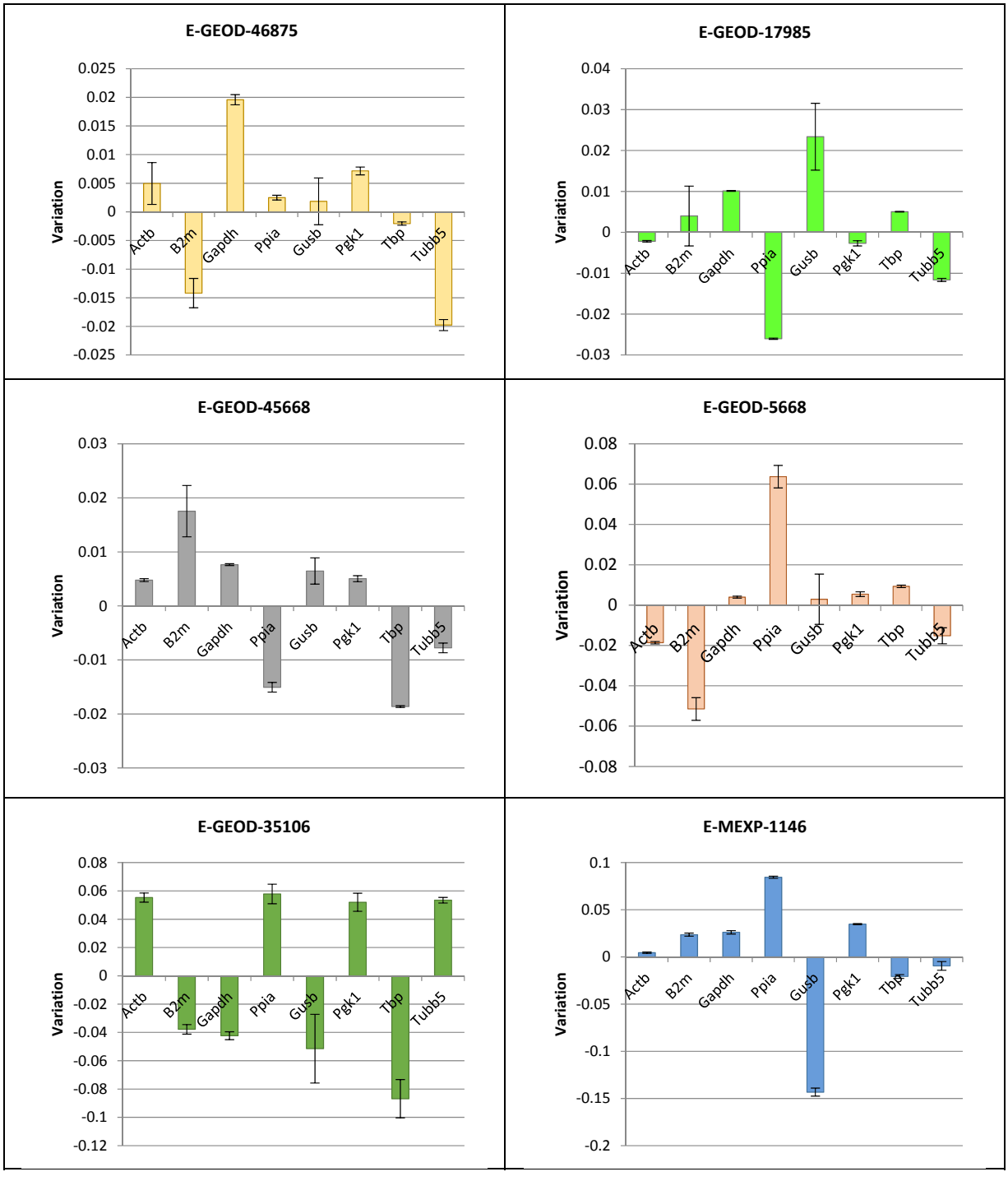


Figure 2- The results of plotting the inter-group variations and the average of intra-group variations as the error bars

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

6. CONCLUSION AND FUTURE WORK

In investigating the stability of HKGs using microarray datasets, our results showed that the most stable gene in each experiment may be different and thus there is no one stable HKG for all conditions. Even when the study is completely narrowed to reproductive biology, in mouse and for oocyte samples, the result is not the same for different conditions and different experiments.

The most stable genes in oocyte samples of mouse were identified as follows: GAPDH, TBP, ACTB and PPIA.

We have also seen that both inter-group and intra-group variations depend on the experiment and its conditions and vary in different situations. Therefore, choosing any normalization methods may result in wrong conclusions. We recommend that clinical researchers should analyze gene expression stability and then choose the most stable genes for a specific condition before they use them as normalizer.

Going forward, we would like to investigate different methodologies for evaluation of gene stability. We would like to study other organism and cells as well. Another research direction may also be investigating different preprocessing method and its effects on the results.

REFERENCES

[1] A Brazma, A Robinson, G Cameron, M Ashburner, "One-stop Shop for Microarray Data", Nature 403: 699-700, 2000.

[2] P Baldi, GW Hatfield "DNA Microarrays and Gene Expression : from Experiments to Data Analysis and Modeling", Cambridge University Press, United Kingdom, 2002

[3] JH Do, D Choi, "Normalization of Microarray Data: Single-Labeled and Dual-Labeled Arrays", Molecules and cells, 22(3), 254, 2006.

[4] BM Bolstad, "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias", TP Speed - Bioinformatics, 2003 - Oxford Univ Press

[5] JL Fritchman, JF Weidman, KV Small, M Sandusky , " The Minimal Gene Complement of Mycoplasma Genitalium", Science 270, 397–403, 1995

[6] EV Koonin, "How Many Genes Can Make a Cell: the Minimal-Gene-Set Concept". Annual Review of Genomics and Human Genetics 1, 99–116, 2000.

[7] NC Noriega, SG Kohama, "Microarray Analysis of Relative Gene Expression Stability for Selection of Internal Reference Genes in the Rhesus Macaque Brain-BMC Molecular Biology", 2010 - biomedcentral.com

[8] CL Andersen, JL Jensen, TF Ørntoft , "Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets Cancer research", Cancer Research 64, 5245–5250, 2004

[9] A Radonić, S Thulke, IM Mackay, O Landt… - Biochemical, W Siegert, and A Nitsche, "Guideline to Reference Gene Selection for Quantitative Real-Time PCR", Biochemical and Biophysical Research Communications 313, 856–862, 2004

[10] J Vandesompele, K De Preter, F Pattyn, B Poppe, NV Roy, N De Paepe and F Speleman, "Accurate Normalization of Real-Time Quantitative RT-PCR Data by Geometric Averaging of Multiple Internal Control Genes" , Genome Biology 2002, 3(7):research 0034.1–0034.11, genomebiology.com

[11] A Uyar, S Torrealday, E Seli, "Cumulus and Granulosa Cell Markers of Oocyte and Embryo Quality", Fertility and Sterility, 99(4), 979-997, 2013

[12] DM Mutch, A Berger, R Mansourian, A Ryt, "Microarray data analysis: a practical approach for selecting differentially expressed genes". Genome Biol 2001, 2:preprint0009.0001-0009.0029.

[13] T Park, SG Yi, SH Kang, SY Lee, S Kang, R Simon, "Evaluation of Normalization Methods for Microarray Data". BMC Bioinformatics, 4(1), 33, 2003.

[14] MA Hannah, H Redestig, A Leisse, L Willmitzer, "Global mRNA Changes in Microarray Experiments", Nature biotechnology 26:741–2, 2008.

[15] SA Bustin, T Nolan, "Pitfalls of Quantitative Real-Time Reverse Transcription Polymerase Chain Reaction". J Biomol Tech 15:155-166, 2004.

[16] O Thellin, W Zorzi, B Lakaye, B De Borman "Housekeeping Genes as Internal Standards: Use and Limits", Journal of Biotechnology, 75, 291–295, 1999.

[17] MD Robinson, A Oshlack, "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data". Genome Biology. 11, R25 5, 2010.

[18] K Dheda, JF Huggett, SA Bustin, MA Johnson, "Validation of Housekeeping Genes for Normalizing RNA Expression in Real-Time PCR". BioTechniques 37:112-119, 2004.

[19] C Rubie, K Kempf, J Hans, T Su, B Tilton, T Georgb, B Brittnera, B Ludwiga, M Schilling, "Housekeeping Gene Variability in Normal and Cancerous Colorectal, Pancreatic, Esophageal, Gastric and Hepatic Tissues". Molecular and Cellular Probes 19, 101–109, 2005.

[20] PD Lee, R Sladek, CMT Greenwood, TJ Hudson "Control Genes and Variability: Absence of Ubiquitous Reference Transcripts in Diverse Mammalian Expression Studies." Genome Research 12.2, 292-297, 2002.

[21] S Manafi, A Uyar, A Bener "Sampling Bias in Microarray Data Analysis: A Demonstration in the Field of Reproductive Biology". In Health Informatics and Bioinformatics (HIBIT), 2013 8th International Symposium,IEEE , 2013.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[22] J Zhu, F He, S Song, J Wang, J Yu, "How Many Human Genes Can Be Defined as Housekeeping with Current Expression Data?" BMC Genomics 9, 172, 2008.

[23] MW Pfaffl, A Tichopad, C Prgomet, TP Neuvians, "Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – Excel-based tool using pair-wise correlations", Biotechnology Letters 26: 509–515, 2004.

[24] European Bioinformatics Institute, www.ebi.ac.uk

[25] EJ Romasko, D Amarnath, U Midic, KE Latham, "Association of Maternal mRNA and Phosphorylated EIF4EBP1 Variants with the Spindle in Mouse Oocytes: Localized Translational Control Supporting Female Meiosis IN Mammals", Genetics 195, 349-358, 2013

[26] B Xu, Y Obata, F Cao, T Taketo, "The Presence of the Y-Chromosome, not the Absence of the Second X-Chromosome, Alters the MRNA Levels Stored in the Fully Grown XY Mouse Oocyte", PloS one, 2012

[27] J Chen, C Melton, N Suh, JS Oh, K Horner, "Genome Wide Analysis of Translation Reveals a Critical Role for Deleted In Azoospermia-Like (Dazl) at the Oocyte-to Zygote Transition." Genes Dev. 25: 755-766, 2011.

[28] J Ma, M Flemr, P Stein, P Berninger, R Malik, "MicroRNA Activity Is Suppressed in Mouse Oocytes", Current Biology 20, 265–270, 2010

[29] YQ Su, K Sugiura, Y Woo, K Wiggleswort, "Selective Degradation of Transcripts During Meiotic Maturation of Mouse Oocytes", Developmental Biology, 302, 104–117, 2007.

[30] RS Oliveri, M Kalisz, CK Schjerling, "Evaluation in Mammalian Oocytes of Gene Transcripts Linked to Epigenetic Reprogramming", Society for Reproduction and Fertility, 1741–7899 (online), 2007.

[31] "Mouse Endogenous Control Gene Panel", TATAA Biocenter, www.tataa.com

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

A Comparative Analysis of Computational Indel Calling Pipelines for Next Generation Sequencing Data

Jacob Porter, Jonathan Berkhahn, and Liqing Zhang
Computer Science, Virginia Tech, Blacksburg, VA, USA

Abstract - *Insertions and deletions (indels), are one of the most common class of mutations in the human genome. Correctly detecting and identifying indels is important in the study of human genetics and disease. We evaluated the precision and recall of combinations of read mapping and indel calling software on calling short and longer indels with variable read coverage on simulated data. We examined the popular read mappers BFAST, Bowtie2, BWA, and Shrimp and the indel callers Dindel, FreeBayes, and SNVer. Interestingly there were interactions between read mappers and indel callers. On simulated data, the BFAST-Dindel and Shrimp-SNVer pipelines showed superior performance in most cases. Real data from human chromosome 22 with indels determined from an alternative indel pipeline were used to validate the computational pipelines and to assess run-time. The Shrimp-SNVer pipeline was the most accurate, while pipelines with FreeBayes did poorly. We discuss reasons for pipeline accuracy.*

Keywords: indel calling bowtie2 bfast bwa snver

1 Introduction

Indels are the second most common class of mutation in the human genome [1]. They consist of an insertion or deletion of one or more DNA bases into a genome. This can have far ranging effects concerning gene expression and genetic disease [1]. Detecting and identifying indels is a multistep process that can introduce error at every step. Starting with a set of DNA sequence reads and a reference DNA genome, reads are first mapped to the reference genome with a mapping program and then the mapped results are inputted into indel calling software to identify indels.

A growing variety of software is available for read mapping and indel calling. New tools are constantly being developed with an eye towards better performance and increased accuracy. As the variety of available tools and the complexity of the technologies involved in indel calling increases, it becomes increasingly important to understand the relationships between mapping software and indel calling software. While previous work [2, 3] assessed the accuracy of indel calling by only varying mapping software or by only varying indel calling software, we studied the accuracy of mapper and indel calling software combinations. We evaluated the accuracy of pipelines consisting of four popular mapping programs and three indel calling programs on simulated data based on a portion of human chromosome one. For mapping, BFAST [7], Bowtie2 [4], BWA [5], and SHRIMP [6] were selected. For indel calling, we used Dindel [8], Freebayes [10], and SNVer [9]. We varied the coverage of the reads inputted to the mappers to study the effects of different levels of read coverage on the precision and recall of called indels. We evaluated the accuracy of these pipelines on indels from 1-30 bases long.

Furthermore, pipeline accuracy was assessed with real human data from chromosome 22. The indels were validated with an alternate method described in the paper [12].

The remainder of this paper is organized as follows. Section 2 discusses the read mapping software and the indel calling software we selected. It discusses methods we used to generate our simulated and real data sets, and the statistics that we used to evaluate the accuracy of our pipelines. Section 3 discusses the accuracy results of the pipelines and runtime on real data. Section 4 concludes.

2 Methods

2.1 Software Workflow

We selected mapping software that was both widely used and that covered a variety of different algorithms. Bowtie2 [4] and Burrows-Wheeler Aligner (BWA, [5]) were both popular tools that use the Burrows-Wheeler transform to map reads. SHRiMP [6] and BFAST [7] are both hash-based mapping tools. Shrimp creates a hash table index of the read sequences, but BFAST creates a hash table index of the reference sequences. For indel calling, we selected two programs that use Bayesian statistics, Dindel [8] and Freebayes [10]. SNVer [9] is based on a frequentist binomial-binomial model developed by the SNVer authors.

All of our experiments were run on SystemG nodes. SystemG is a research cluster at Virginia Tech. Each node had two quad-core 2.8 GHz Intel Xeon processors and 8 gigabytes of RAM. The mappers were run with four threads when possible, but the indel callers were single-threaded only.

Each tool was run with default settings since that is how the tools will most likely be used. The workflow consisted of creating SAM files from each read-mapper and then

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

transforming the SAM files into BAM files with samtools. Finally, each indel-caller produced a VCF file from the BAM files. The following are the version numbers for the software: bfast-0.7.0a, bowtie2-2.1.0, bwa-0.7.1, SHRiMP-2.2.3, dindel-1.01, freebayes-0.9.9, and SNVer-0.4.1. The real data workflow was similar to the simulated data workflow. The differences are that Shrimp was run with --no-qv-check and --qv-offset 33 because the real data were Sanger traces rather than Illumina reads, and Dindel was run with --numWindowsPerFile 1000000. Dindel was not used much on real data because at one day of running, it was still not finished. The workflow and arguments used were the following.

1. Read Mapping:

BFAST:
```
bfast fasta2brg -f reference.fasta
bfast index -f reference.fasta -m
1111111111111111111111 -w 14 -n 4
bfast match -f reference.fasta -r reads.fastq -n 4 -t 1>
bfast.matches.bmf 2> bfast.matches.out
bfast localalign -n 4 -t -f reference.fasta -m
bfast.matches.bmf 1> bfast.aligned.baf 2> bfast.aligned.out
bfast postprocess -f reference.fasta -i bfast.aligned.baf -o 3
-a 3 > align.sam
```

Bowtie 2:
```
bowtie2-build reference.fasta reference
bowtie2 -x ref -U reads.fastq -S align.sam
```

BWA:
```
bwa index reference.fasta
bwa aln reference.fasta reads.fastq > align.sai
bwa samse reference.fasta align.sai reads.fastq > align.sam
```

SHRiMP:
```
gmapper -N 4 reads.fastq reference.fasta > align.sam
```

2. BAM Conversion:

Samtools:
```
samtools faidx reference.fasta
samtools view -b -S align.sam > align.bam
samtools sort align.bam align.sorted
samtools index align.sorted.bam
```

3. Indel Calling

Dindel:
```
dindel --ref reference.fasta --outputFile dindel_output --
bamFile align.sorted.bam --analysis getCIGARindels
makeWindows.py --inputVarFile dindel_output.variants.txt
--windowFilePrefix realign_windows --
numWindowsPerFile 1000
dindel –analysis indels --bamFile align.sorted.bam –
doDiploid --ref reference.fasta --varFile
realign_windows.1.txt --libFile dindel_output.libraries.txt --
outputFile stage3_output
echo "stage3_output.glf.txt" > list.txt
mergeOutputDiploid.py -i list.txt -o indels.vcf -r
reference.fasta
```

Freebayes:
```
freebayes --no-snps --no-mnps --no-complex -b
align.sorted.bam -f reference.fasta -v indels.vcf
```

SNVer:
```
java -jar SNVerIndividual.jar -i align.sorted.bam -o
indels.vcf -r reference.fasta
```

2.2 Simulated Data

The simulated data was generated from 10 megabases of chromosome one from a publicly available human genome available from the National Center for Biotechnology Information. Artificial mutations were introduced using inGAP, a software tool for the manipulation of genetic data [11]. SNPs were inserted at a divergence rate of 0.1%, and indels were inserted at a divergence rate of 0.02%. These values were chosen since they were realistic [1]. Indel lengths were uniformly distributed from one to thirty bases. Ten replicates of simulated reads were produced to generate average and error statistics for the tests. Reads of uniform 50 base pair length were generated from the mutation sequences in the fastq file format using inGAP. Reads of length 50 were chosen because indel identification is more complex for shorter single-end reads since they "lack insert length variance" [2], so short single end reads represented a good test of indel pipeline sensitivity. In order to study the effects of varying coverage on the accuracy of the pipelines, reads were generated for 10x, 50x, and 100x coverage for each of the mutation sequences.

2.3 Real Data

Applied Biosystems (Sanger) paired-end traces from the set Chr_22_7340 were identified and downloaded from the NCBI trace archive. These traces were used in a Devine lab study that searched for indels in human chromosome 22 [12]. The paper identified 6487 indels for the Chr_22_7340 traces.

We cleansed the traces of contamination using NCBI VecScreen where traces with vector contamination in the middle were discarded, and traces with vector contamination on the ends had the contamination trimmed off. After this, there were 217,924 traces with sizes as much as 2000bp. Since short read mappers perform poorly with very long sequences, 10 million 100bp portions of the traces were sampled with replacement in order to simulate short reads. The 10 million simulated single-end sequences were run through the pipelines with timing tracked with the Linux "date" command.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2.4 Indel Detection

A confusion matrix for each pipeline on each data set was created that recorded true positives, false positives and false negatives. Indels were recorded as true positives if the predicted indel's position was plus or minus 5 nucleotides of the actual indel's position and the predicted length was within 5 percent of the actual length (with all lengths set to be one if 5 percent of the actual length was less than 1). The indel had to be correctly classified as an insertion or a deletion to be marked a true positive; otherwise, it was classified as a false positive. The sequence identity of predicted and actual indels was not checked since differences in sequence identity were rare. A false positive was a predicted indel that didn't meet the preceding criteria, and a false negative was an actual indel that wasn't identified by the indel classifier. Precision, recall, and F1-score were calculated for all pipelines to assess accuracy. Python 2.6 and Bash scripts were created to do the statistical analysis and workflow.

3 Results and Discussion

3.1 Analysis of F1-Score and Coverage on Simulated Data

In our results on simulated data with indels of size 1-30 bases there were clearly pipelines that performed better than other pipelines as measured by the F1-score (Figure 1). For each pipeline, Figure 1 shows average F1-score and the minimum and maximum F1-score of the 10 replicates. Most indels called had fewer than 10 bases.

Figure 1 shows that pipelines with 10X coverage have the best F1-scores, and that 50X and 100X coverage perform less well. This was explained by a tradeoff between precision and recall caused by both increasing false positives and increasing true positives. As coverage increased, recall increased since indel callers return more predicted indels and thus more genuine indels. However, there were more false positives as coverage increased, so precision went down as coverage increased. The general downward trend of the F1-score was because precision decreased more than recall increased with increasing coverage. This suggests that there is some coverage amount that maximizes F1-score for the data, and increasing coverage isn't always desirable. This result was consistent with other work that showed statistically significant precision and recall trends with increasing coverage [2].

The three top performing pipelines were BFAST-Dindel (avg F1-score 0.66), SHRIMP-SNVer (0.53), and BFAST-SNVer (0.51) in the 10X coverage for 1-30 indels. The BFAST-Dindel pipeline had the best average F1-score for all coverage amounts (Figure 1). SHRIMP pipelines were interesting since the F1-score varied considerably. The Shrimp-SNVer pipeline was among the top performing, but Shrimp-Freebayes and Shrimp-Dindel performed poorly. By default, Shrimp mapped some reads to multiple positions.

Read mappers use a seed and extend strategy, and BFAST's seeding strategy used a sliding window at every base. Bowtie2 used multiple 20bp seeds with an offset determined by the read length. Perhaps BFAST's seed strategy allowed it to be more

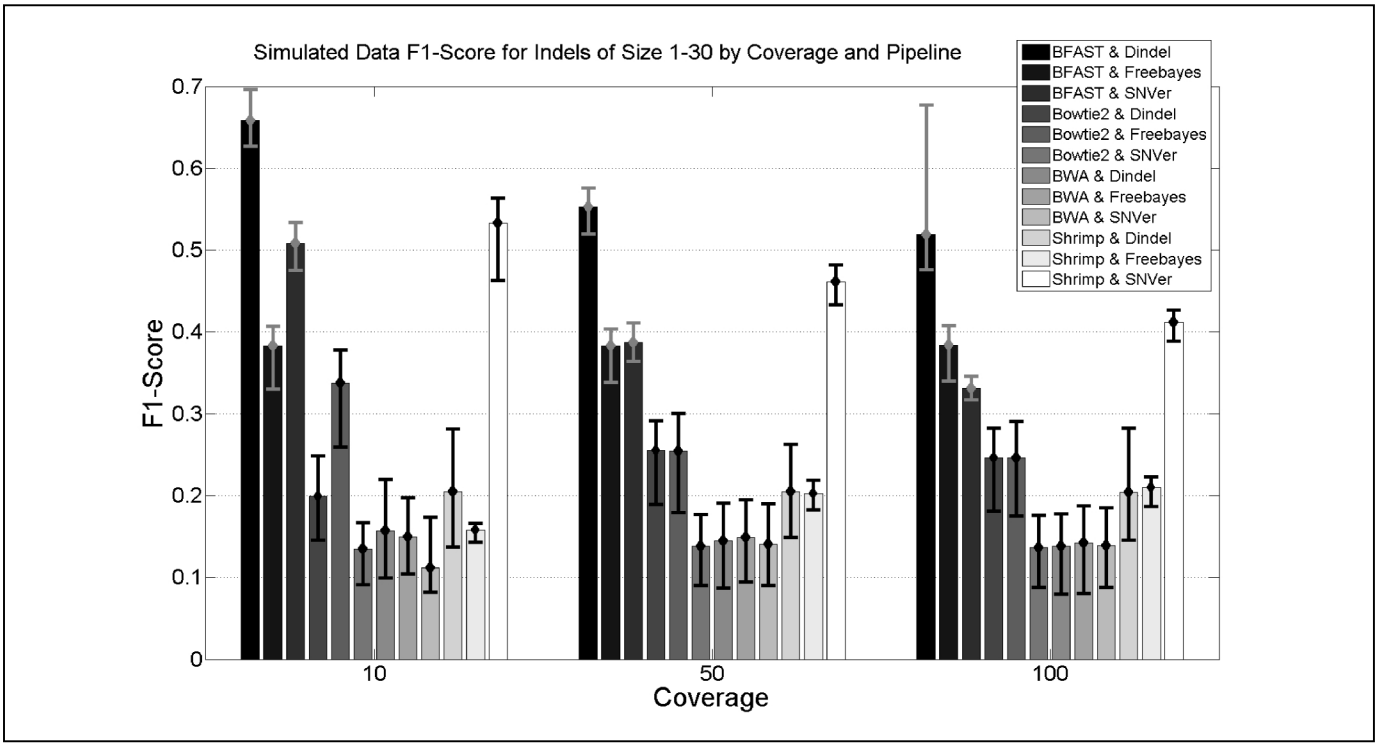


Figure 1 : The F1-score of indel calling pipelines on simulated data with reads containing indels of 1-30 bases. The results are divided into sets of 10X, 50X, and 100X coverage. The F1-score is shown with average, low and high scores.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

accurate. All the mappers' extension phases are similar since they use local or global alignment algorithms [4, 5, 6, 7].

The F1-score difference for 10X coverage between the best pipeline (BFAST-Dindel) and the worst pipeline (BWA-SNVer) was about 0.546. BWA generally performed poorly and this could be because it only supported gaps less than 10 bases in its alignments [5]. Even though BWA and Bowtie2 used a similar seeding strategy with the Burrows-Wheeler transform, Bowtie2 pipelines usually had better F1-scores. Bowtie2's split seed approach handles some variation in the seed [4].

Interestingly, there isn't one clear indel caller that did the best overall. Bowtie2-SNVer was among the lowest performing while SHRIMP-SNVer was among the best performing. Dindel did well with BFAST but not very well with SHRIMP.

3.2 Precision and Recall on Simulated Data With Smaller and Longer Indels

Figures 2 and 3 show a comparison of the effects of longer indels on precision and recall at 10X coverage. Pipelines performed worse for data with indels of 1-30 bases (Figure 3) than for indels with 1-10 bases (Figure 2). Figure 3's precision-recall tuples are generally shifted left when compared to Figure 2. The Bowtie2-Freebayes pipeline did noticeably better with 1-10 indel lengths.

The precision-recall plots show which pipelines are conservative, which generous, and which are balanced. The pipelines involving BWA and Bowtie2 were the most conservative with high precision but low recall. Pipelines involving Freebayes were the most generous with low precision but higher recall. BFAST-Dindel and Shrimp-SNVer had the most balanced precision and recall results with (0.648,0.771) and (0.640,0.799) respectively for indels of length 1-10. The BFAST-Dindel pipeline performed better than Shrimp-SNVer for indels of length 1-30.

Bowtie2 pipelines generally appeared mediocre in our tests. BFAST pipelines had generally good performance with different indel callers while SHRIMP pipeline performance varied considerably with indel calling software. BWA pipelines performed poorly.

3.3 Accuracy of Real Data

The only accurately called indels on the real data were smaller than 5bp. Figure 4 shows the F1-scores of the real data pipelines. SNVer pipelines had the best F1-score. Similar to the simulated data, pipelines with Freebayes were too generous with high recall but low precision. Average precision and recall for Freebayes was 0.000440955 and 0.010906428, but with SNVer it was 0.00117591 and 0.001079081. Thus, SNVer was more conservative in indel calling. The Shrimp-SNVer and Bowtie2-SNVer pipelines did the best while BWA-Freebayes was the worst. The choice of read mapper made little difference, and this could be because only small indels were called. True positives were few relative to indels called (Table 1). For the BWA mappings, Dindel completed in 6.6 hours with similar precision (0.00062) and recall (0.0026) to the BFAST-SNVer pipeline (0.00079, 0.0012).

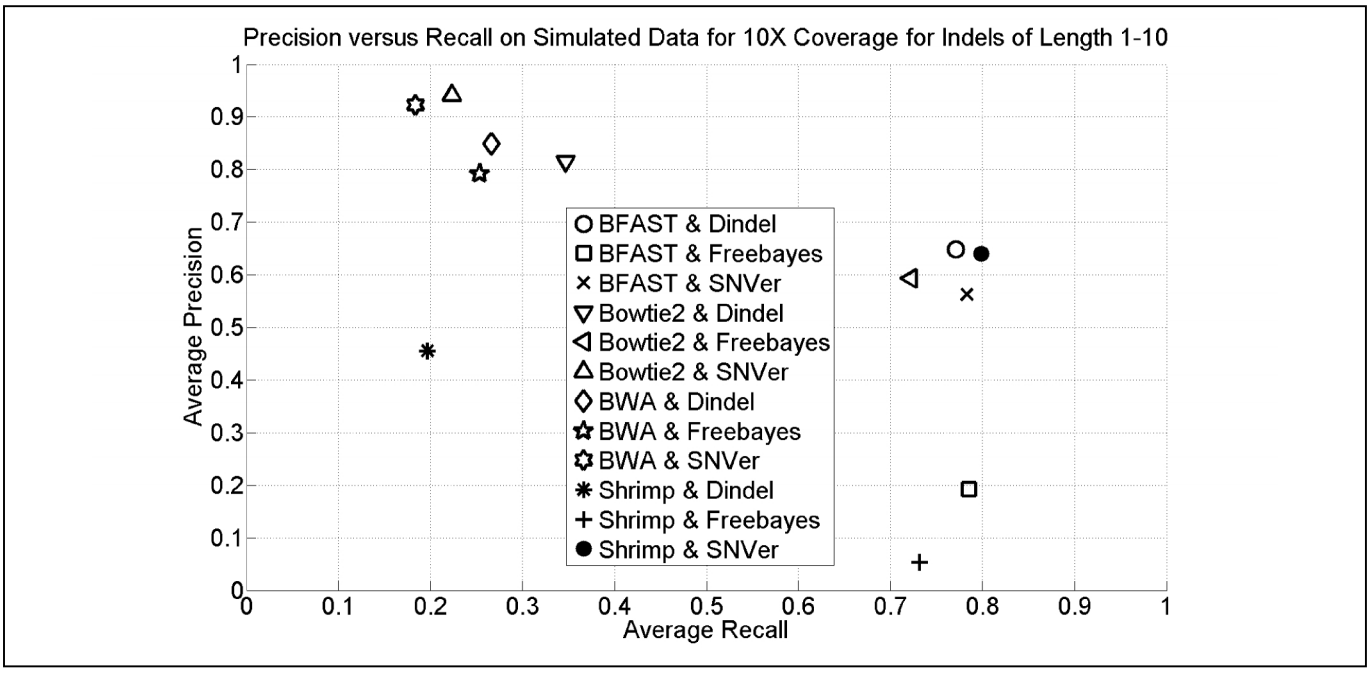


Figure 2 : Average precision and average recall for 10 simulated data replicates for the indel calling pipelines. Precision and recall was calculated for indels with only 1-10 bases at 10X coverage. The reads contained indels as large as 30 bases.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

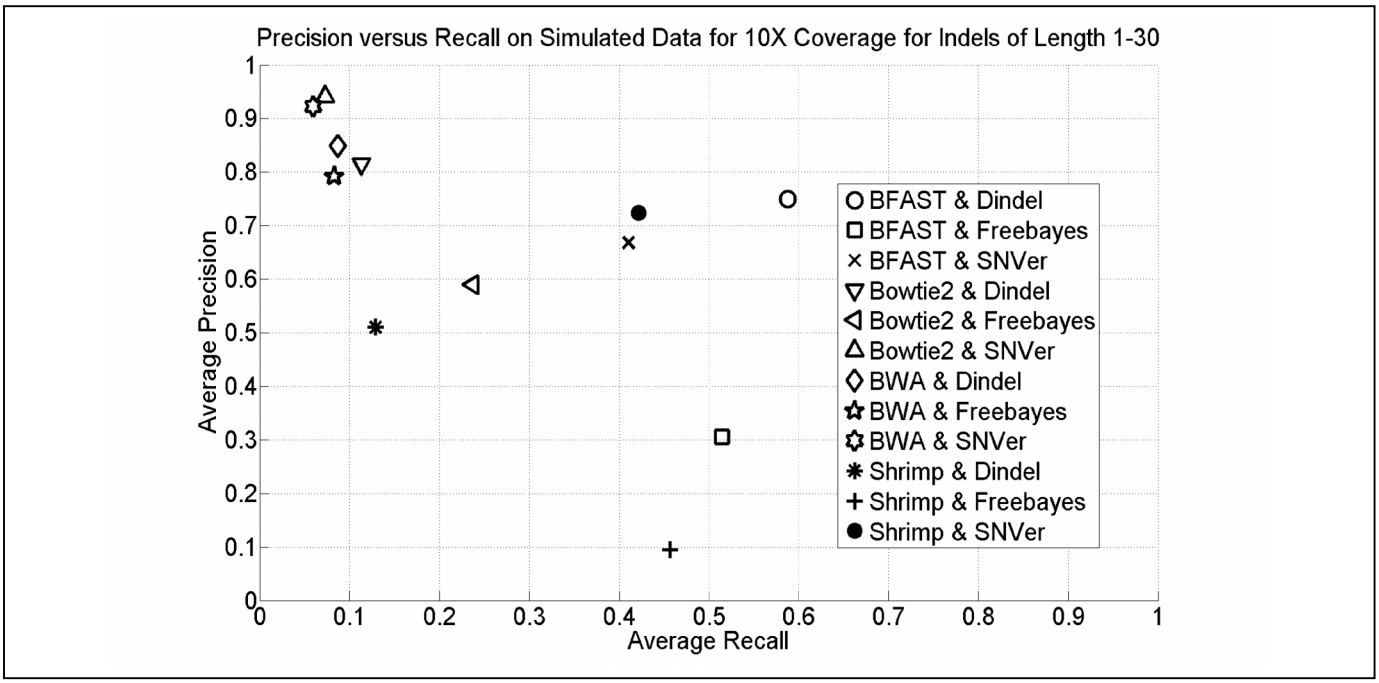


Figure 3 : Average precision and average recall for 10 simulated data replicates for the indel calling pipelines. Precision and recall was calculated for all indels. Indels had 1-30 bases at 10X coverage.

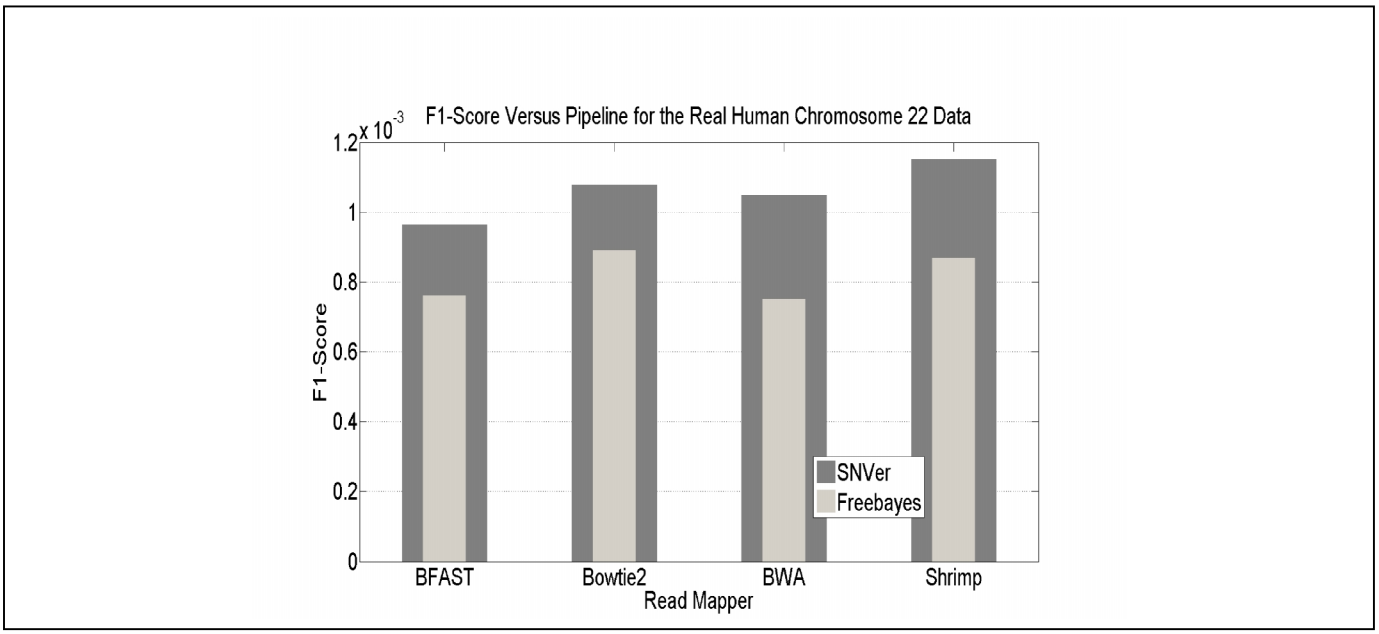


Figure 4 : F1-Score for the indel calling pipelines on 10 million real human chromosome 22 traces.

3.4 Run-Time Performance on Real Data

Table 1 summarizes run-time performance for the pipelines for the real human chr22 data. Bowtie2 and BWA, had the fastest runtimes at 27 and 23 minutes respectively. BFAST and Shrimp were the slowest mappers, and both used a sliding window hashing seed strategy. BFAST was about 10 minutes slower, but Shrimp was 6.2 times slower than BWA, making Shrimp pipelines the slowest. The Shrimp read mapping percent is more than 100 percent since it mapped some reads to multiple positions by default.

The indel callers did not have multithreading, so they were slow. Freebayes was always faster than SNVer, and SNVer took 145 minutes with Shrimp's input making the Shrimp-SNVer pipeline the slowest. Dindel took over a day (except

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 1 : Mapper, caller, pipeline run-time, percent mapped, and indels called on 10 million real human 100bp reads

Mapper	Caller	Minutes	Total Minutes	% Reads Mapped	Total Indels Called	True Indels
BFAST		38		0.5437498		
	Samtools	5				
	SNVer	18	61		20486	8
	Freebayes	37	80		358574	139
Bowtie2		27		0.4850195		
	Samtools	4				
	SNVer	19	50		9812	6
	Freebayes	16	47		114679	54
BWA		23		0.412648		
	Samtools	4				
	SNVer	18	45		6388	5
	Freebayes	9	36		30789	14
Shrimp		143		1.7618786		
	Samtools	4				
	SNVer	145	292		9145	9
	Freebayes	59	206		168684	76

with BWA input), making it less tolerable for big indel calling projects; however, Dindel has the ability to split its work into multiple files for manual multiprocessing.

4 Conclusions

To our knowledge, this work is the first to look at the accuracy of the combination of mapping software and indel calling software with larger (>10 nucleotides) indels. F1-score, a measure of accuracy, fell with increased coverage, belying expectations. Indel calling accuracy depended on the combination of mapping software and indel calling software. Some of the top performing pipelines were BFAST-Dindel, SHRIMP-SNVer, and BFAST-SNVer on simulated data. The best pipeline had an F1-score 0.6 higher than the worst pipeline on simulated reads. On real data, SNVer pipelines were more accurate than FreeBayes pipelines in all cases. SNVer and Shrimp can have slow runtimes, but Dindel was by far the slowest. Future work could include exploring the parameter space of the tools to observe the effects of argument selection on sensitivity.

5 References

[1] Mullaney JM, Mills RE, Pittard WS, *et al.* Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* 2010;19:R 131-b

[2] Neumann JA, Isakov O, Shomron N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform* 2013 Jan;14(1):46-55

[3] Pabinger S. *et al.* survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 2013 Jan 21.

[4] Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

[5] Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 2009, 25:1754-60.

[6] David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: sensitive yet practical Short Read Mapping. *Bioinformatics* 2011, 27:1011-102.

[7] Homer N, Merriman B, Nelson SF. BFAST: An alignment tool for large scale genome resequencing. *PLoS ONE* (2009). 4(11): e7767.

[8] CA Albers, G Lunter, Daniel G MacArthur, Gilean McVean, Willem H Ouwehand, Richard Durbin. Dindel: Accurate indel calls from short-read data. *Genome Research*: 2010

[9] Erik Garrison, Gabor Marth. Haplotype-based variant detection from short-read sequencing. ArXiv:1207.3907[q-bio.GN]

[10] Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 2011 Oct; 39(19):e132

[11] Qi J, Zhao F, Buboltz A, *et al.* InGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics* 2010;26:127-9.

[12] Ryan E. Mill, Stephen Pittard, et al. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*. 2011. 21: 830-839

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Relevance of Information in Cell Signaling Pathways using Default Logic

A. Doncescu[1], **P. Siegel**[2], **and T. Le**[1]
[1]LAAS-CNRS, University of Toulouse, Toulouse, France
[2]Aix Marseille Université, CNRS, LIF UMR 7279, 13288, Marseille, France

Abstract— *Cell Signaling Pathway Simulation is a very useful tool in the drug discovery process. These simulation programs can be divided into dynamic simulation and Knowledge-Based Discovery. In the first case the simulation is based on differential equations and could be considered in "real-time", meanwhile in the case of Knowledge Based Discovery Programs KBDP the consistency of the model is checked. The most efficient KBDP approach is based on first order logic (FOL). In this paper, algorithms based on Default Logic are proposed to check-out the consistency of the simplest representation of DNA double strand breaks. DNA double-strand breaks are among the most severe genomic lesions. This representation is concise and adequat for keeping the flow of information represented by gene expression, receptor and protein structure through the apoptosis and cell cycle.*

Keywords: Double Strand Breaks, DNA Damage, Default Logic, Extensions, Abduction, Consistent Pathway

1. Introduction

Today the conception of artificial systems attempts to imitate the natural systems by developing new concepts of reasoning able to handle a high level of heterogeneity and uncertainty. These complex systems have a dynamic evolution in terms of structure and organization. In order to model and control these systems there is a need to observe and reconstruct their behavior by a relevant model which should make sense of large amounts of heterogeneous data gathered on various scales. System Biology is a research field, which needs an appropriate evaluation of their know-how corroborated with the available experimental data in order to represent knowledge and discover new knowledge. Therefore, System Biology could be view as a complex network constituted of protein-protein interactions, small-molecule metabolism and gene regulation.

From the standpoint of Artificial Intelligence, cells are sources of information that include a large amount of intra and extra cellular signals. Disease and cancer in particular can be seen as a pathological alteration in the signaling networks of the cell. The study of signaling events appears to be the key of biological, pharmacological and medical research. For a decade signaling networks have been studied using analytical methods based on the recognition of proteins by specific antibodies. Parallel DNA chips (microarrays) are widely used to study the co-expression of candidate genes to explain the etymology of certain diseases, including cancer. The resulting data allows the modeling of gene interactions. The biological experts look for evidence of interactions between metabolites and genes. Therefore the representation by graphs is the best way to understand biological systems. This representation includes mathematical properties as connectivity; presence of positive and negative loops which is related to a main property of genetic regulatory networks. Biochemical reactions are very often a series of time steps instead of one elementary action. Therefore, one direction research in system biology is to capture or to describe the series of steps called pathways by metabolic engineering. All reactions that allow the transformation of one initial molecule to a final one constitute metabolic pathways. Each compound that participates in different metabolic pathways is grouped under the term metabolite.

The study of gene networks poses problems well identified and studied in Artificial Intelligence over the last thirty years. Indeed, the description of network is not complete: biological experiments provide a number of protein interactions but certainly not all of them. On the other hand the conditions and sometimes the difficulties of the experiments involves these data are not always accurate. Some data may be very wrong and must be corrected or revised in the future. Finally the information coming from different sources and experiences can be contradictory. It is the goal of different logics, and particularly non-monotonic logics, to handle this kinds of situation. Afterwards this interaction maps should be validated by biological experiments. Of course, these experiments are time consuming and expensive, but less than an exhaustive experiment.

In this paper we focus on three main problems: handling the conflicts which can occur in the gene representation, completing in-silico the gene network and the practical handling complexity of the algorithm allowing the inferences for knowledge discovery on these networks. Our approach is based on default logic allowing to handle the incomplete information, and abductive reasoning to complete the missing information from the gene network. The last part is dedicated to a new language of representation, which seems to be the key to algorithm complexity handling.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2. Declarative Representation of Signalling Pathway

Figure 1 shows a simplified pathway of interactions in a cell. Through different mechanisms, not shown here, the ultraviolet UV drives the cell to cancer. This is represented by an arrow : $UV \rightarrow cancer$. On the other hand the UV activate the protein P53 ($UV \rightarrow P53$). This protein activates a protein A ($P53 \rightarrow A$)) and A blocks cancer ($A \dashv cancer$). But, in some conditions, Mdm2 binds protein P53 and the obtained complex, activates B and B blocks A. This example is of course very elementary, but it helps to ask questions related to the representation of networks of genes side view of artificial intelligence and algorithms. In practice, this may include pathways with thousands of genes, which will pose problems of computational complexity. In this article, to test the representation and the algorithms, we use the example introduced in [1] and [18] (Fig. 1) and the map given by Pommier [21] (Fig. 2).

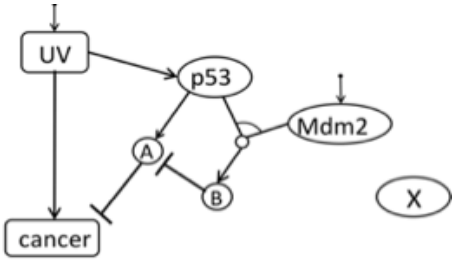


Fig. 1: The Simplified Model of Double Strand Break.

2.1 Double strand break of DNA

The cell's response to a double strand break of DNA (DBS) has been studied for some years, but the ATM-dependant signaling pathway has only been clarified since the discovering of H2AX [2], the phosphorylated form of histone H2AX. All the protein interactions of this pathway have been reported [21], including the signalization of the double-strand break (involving important proteins such as: H2AX, MDC1, BRCA1 and the MRN complex) but also for the checkpoint mechanisms (involving p53, the Cdc25's and Chk2).

In a general way, the cell can receive information by protein interactions that will transducer signals. First, the information is discovered by sensor proteins, which will recruit some other mediator proteins whose function will be to help all the interactions between the sensors and the transducers. These transducers are proteins that will amplify the signal by biochemical methods such as phosphorylation. In the end, the signal will be given to effectors that will engage important cell process. In this pathway, the DSB is recognized by the MRN complex, which in turn will recruit ATM in its inactive dimer form, and then ATM will phosphorylate itself and dissociate to become an active

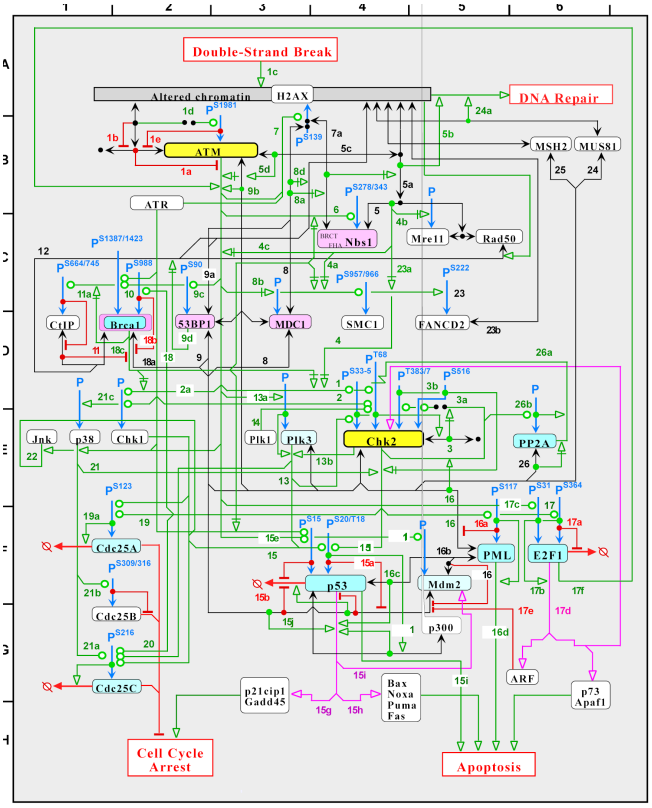
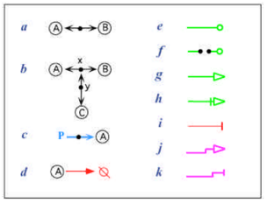


Fig. 2: DNA Double Strand Break Map.



Symbols :
(a) Proteins A and B could link, the knot represents the binding A:B ;
(b) Multimolecular Complex binding ;
(c) Covalent Modification of A ;
(d) Degradation of A ;
(e) Enzymatic stimulation in transcription ;
(f) Autophosphorylation
(g) General Stimulation;
(h) Necessity ;
(i) Inhibition ;
(j) Activation of transcription ;
(k) Inhibition of transcription

monomer. This active form of ATM will phosphorylate many mediators such as $\gamma-$H2AX, MDC1, BRCA1 or 53BP1. Then, the signal is transduced by important proteins such as Chk2, p53 (a very important protein, which can cause cancer if mutated) or the Cdc25's. The effectors can be different with the context: p21 and Gadd45 will induce the cycle arrest, whereas Box, Nas, Puma and Fas will induce the cell apoptosis.

3. Logic representation

Genes and proteins are considered the same object (the genes produce proteins). In this article, we often use a propositional representation. But, in practice the detailed study of interactions will be asked to represent increases or decreases protein concentration. It therefore falls outside the scope of propositional logic, but the basic problems are the same, especially for the issue of computational complexity.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Indeed the protein concentration is rarely precise, and often in practice, the biologists experiment shows a qualitative interpretation of increasing or decreasing of the concentration. To represent a change in concentration some predicates such as $increased$ or $decrease$ are used.

To describe interactions between proteins it is possible to use a language of classical logic (propositional or first order logic). We can say, for example $stimulation(UV)$ to say that the cell is subjected to ultraviolet or $GlassScreen \rightarrow \neg stimulation(UV)$ to say that a glass screen protects against ultraviolet. We are in a logical framework, so it is possible to represent almost everything in a natural way. The price to pay, if you use the entire first order language, is incompleteness and the combinatorial explosion of complexity algorithms. It is therefore essential to reduce the expressive power of language.

3.1 Causality and Classical Inference

Interactions between genes can be seen as a very simple form of causality. To express basic interactions, it is common to use two binary relations $trigger(A, B)$ and $block(A, B)$ [1], [10]. The first relation means, for example, a protein A triggers or activates the production of a protein B. The second relation is an inhibition. Conventionally, these relations are represented by $A \rightarrow B$ and $A \dashv B$. Theses kind of relations gives a basic form of causality.

Many works were written to represent causality. It is possible to use symbolic or numeric formalisms. You can use Bayesian approaches, probabilistic logics [24]. It is impossible here to go around all these works. We will simply try to describe and use a form of causality (the simplest possible) sufficient for the application to the cell.

The inferences of classical logic $A \rightarrow B$ or $A \vdash B$ are fully described formally, with all the "good" logic properties (tautology, not contradiction, transitivity, contraposition..). But the causality cannot be seen as a classical logic relation. A basic example is "If it rains the grass gets wet". The formula $Rain \rightarrow lawn-wet$ meant that if it rains the grass is always wet. But this formula is too strong. Indeed, there may be exceptions to this rule (the lawn is under a shed...).

In a first approach, the first properties that we want to give can be expressed naturally:

(1) If A triggers B and A is true, then B is true.

(2) If A blocks B and A is true, then B is false.

Depending on the context, true can mean the known, certain, believed, proved... The first idea is to express these laws in classical logic by axioms:

$trigger(A, B) \wedge A \rightarrow B$
$block(A, B) \wedge A \rightarrow \neg B$

They can also be weakly expressed by inference rules, close to Modus Ponens :

$trigger(A, B), A \vdash B$
$block(A, B), A \vdash \neg B$

But these two formulations are problematic when there is conflict. If for example we have a set of four formulas $F = \{A, B, trigger(A, C), blocks(B, C)\}$, we will in the two approaches above infer from F, B and $\neg B$. This is inconsistent. To solve such conflicts, we can try to use methods inspired by constraint programming, such as the use of negation by failure in Prolog or Solar. It is also possible to use a non-monotonic logic.

The first method, negation by failure, poses many theoretical and technical problems if you go further as the simple cases. These problems are often solved by adding properties to the formal system, properties that pose other problems. Therefore, we will use a classical non-monotonic formalism, the default logic of Reiter.

3.2 Causality and Default Logic

To resolve the conflicts seen above, the intuitive idea is to weaken the formulation of rules :

(1) If A trigger B, if A is true and it is possible that B, then B is true.

(2) If A blocks B, if A is true and it is possible that B is false then B is false.

The question then is to describe, what is formally $possible$. This question began to arise in artificial intelligence thirty years ago. In this type of reasoning, one has to reason with incomplete information, uncertain and subject to revision and sometimes false. On the other hand we must often choose between several possible conclusions contradictory. Here we use default logic [23]. This logic can be seen as an improvement and a generalization of the negation by failure in Prolog. It is also a generalization of ASP formalisms which appeared later [17]. With default logic the previous rules will be expressed intuitively :

(1) If A trigger B, if A is true and if B is not contradictory, then B is true.

(2) If A blocks B, if A is true and if $\neg B$ is not contradictory then $\neg B$ is true.

In default logic, these rules can be represented by normal defaults which are special, and specific, inference rules written :

$$d1 = \frac{A : B}{B} \quad \text{and} \quad d2 = \frac{A : \neg B}{\neg B}$$

- A is the prerequiste of default $d1$ and $d2$
- $: B$ (resp. $: \neg B$) are the justifications of $d1$ ($d2$)
- B (resp. $\neg B$) are the consequents of $d1$ ($d2$)

Therefore, the information is represented here using defaults theory $\Delta = \{W, D\}$ where W is a set of classical logic formula and D is the set of defaults.

3.3 Extension and choice of extension

The goal of default logic is to find extensions of a default theory $\Delta = \{W, D\}$. Simplifying, an extension E is a consistent set of formulas obtained by adding, under condition, to W a maximal set of consequents of D. An

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

extension can for example, represent a subgraph without conflict, of the gene network.

The classical definition of extension is based on the utilization of W and a subset of defaults D. An extension is built starting with W and subsequently it is added the maximum consistent set of consequences of D. The condition to use a default starts by checking if the prerequisites (here A for $d1$) are satisfied and the justification (here : B for $d1$) does not lead to contradiction. In a simple manner that means the negation of B is not verified. If this request is True we add the consequent B to W and the algorithm is restarted until all defaults has been used.

For example if we consider $\Delta = \{W, D\}$ with $W = \{A\}$ and $D = \{d1, d2\}$, we obtain two extensions :

$E1 = \{A, B\}$ if $d1$ is used.
$E2 = \{A, \neg B\}$ if $d2$ is used.

By using default logic, the conflict is resolved, but it is not possible to rank the extensions: B is true or false? In fact this will really depend on the context. For biologists, some times the positive interactions are preferred to negatives (or reverse). Another possibility is to use probabilistic or statistical methods or to weight each extension based on the evaluation of the knowledge. From an algorithmic viewpoint the ranking of extension could also be evaluated during the calculation of the extensions and even the off-line ranking could be preferred.

4. Completing the Signaling Networks by Default Abduction

Previously, we introduced a fun-filled example which sums up the question:

"How to block cancer by preventing B ?"

For this example, biological experiments have shown that a protein X could be a candidate for this block. Figures 3 and 4 show two types of interactions with X to hypothesize the blocking of B. Here the biologist completes the causal graph and, for the case of big data, it is necessary to automatize the process. The problem is thus complete in-silico network genes. Biological experiments are done to try to complete it, but these experiments are time consuming and expensive. We need to find, in-silico, a molecule (a future drug) which has a chance to act effectively. This is a problem of abduction.

Classical logic primarily uses three types of reasoning: deduction, induction and abduction. The purpose of deduction is to find out when a result R is inferred from a set of information C, written $C \vdash R$. Induction generalizes the deduction, whereas information is not complete in all its generality, but we know the special cases (examples, experiences..). It should then use these cases to discover general rules.

To simplify, abduction generalizes induction. Here, we do not share examples. The information is incomplete and make abduction amounts to adding to C a set of hypotheses

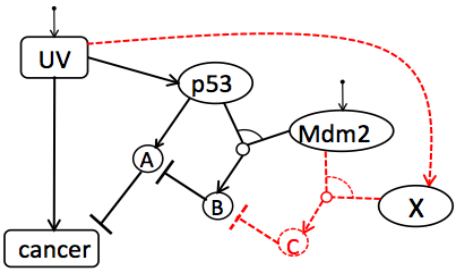


Fig. 3: mdm2 binds X.

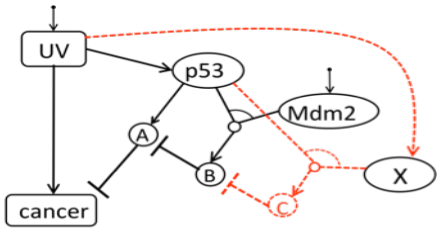


Fig. 4: p53 binds X.

H such that $C \wedge H \vdash R$ and H is consistent with C. In Artificial Intelligence, the notion of abduction is of paramount importance.

The trouble comes with implementation of the algorithms. Abduction algorithms are far too high computational complexity. Even limited to propositional calculus, the theoretical complexity revolves around $\sum_2^p$ which is totally unacceptable when we go beyond small examples. Many theoretical studies have been done on the complexity of the abduction and research sub-language of propositional calculus where complexity is reduced. These sub-languages most often cover the Horn clauses and renaming. But even here the complexity is too great for even, more or less, NP-complete. Conversely, existing polynomial classes provide only a low power of expression on issues to be addressed. On the other hand, for many real applications, experience shows that it is not necessary to use the full expressive power of logic. It seems that this is particularly the case for the study of gene networks.

For genes networks, abduction is used mainly to search missing interactions. These interactions would yield a result (for example "block cancer"). To search if one of these missing interactions:

$X \rightarrow Y$

can be used to obtain the result, it is possible to consider a default of type:

$$\frac{X : Y}{Y}$$

Then you must calculate extensions that contain the result and see the defaults used in these extensions.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

5. Logic Representation of Signaling Pathway to Reduce Computational Complexity

Today, programming language does not exist allowing abduction reasoning under the incomplete and uncertain information. We present the outline of a language dedicated to discovery of biological interactions answering these requests. This formalism uses the default logic and also has a dynamic approach by considering time as a succession of events. The syntax inspired from Prolog is described in the next section.

5.1 Clauses and Horn clauses.

In our representation, $product(P53)$ means that the protein p53 increases in concentration and $\neg product(P53)$ means that it is not possible to determine if the p53 concentration is increasing. The dynamic of the system can be, for example, specified by $concentration(p53, 100, T)$ which means the concentration of $p53$ at the time T is equal to 100 a.u. And $\neg concentration(P53, sup(200), T+3)$ says that at the time $T+3$, the concentration of $p53$ is not greater than 200 a.u.

The simplest formulas are the clauses. Formally, a clause is a disjunction of literals $l_1 \vee .. \vee l_n$. If the connectors are deleted, a clause is a set or a list of literals. For example $\{a, \neg b, \neg c, d\}$ or $a, \neg b, \neg c, d$ represents $a \vee \neg b \vee \neg c \vee d$. A Horn clause is a clause with a maximum of one positive literal. The clauses a and $\neg b \vee \neg c \vee d$ and $\neg b \vee \neg c$ are Horn clauses. And $a \vee b$ is not one. For the rest we use Horn clauses which are interesting for two reasons.

First using Horn clauses is a natural way to represent knowledge. In fact the formula $a \wedge b \wedge c \wedge d \rightarrow d$ is equivalent to the Horn clause $\neg a \vee \neg b \vee \neg c \vee d$. In the same time the formula $\neg(a \wedge b)$ (a and be cannot be True in the same time) is equivalent to the negative Horn clause $\neg a \vee \neg b$.

The second advantage of Horn clauses, fundamental here, is that their use drastically reduces computational complexity. Indeed, any logical formula can be rewritten as a set of clauses, so complexity problems may arise in terms of clauses. For propositional calculus the fundamental problem is whether a set of clauses is consistent or not. This is the problem SAT which is NP-complete. Otherwise all known algorithms are exponential in the worst case. On the other hand, if all clauses are Horn causes, algorithms can be linear proportional to the size of the data. For genes pathways, the use of Horn clauses provides practically usable algorithms .

Obviously Horn clauses can not represent all formulas. In particular $a \vee b$ is not a Horn clause. But in practice, this type of positive disjunctive information is quite rare. We have not really found it for the gene networks that we studied. If there are, most of the time you can use renaming techniques to solve the problem. Finally, if nothing works and it is impossible to use only Horn clauses, there are techniques to limit the combinatorial explosion. For example use strong backdoors, managing mutual exclusion and cardinality, recognition of symmetries [4] [5]. Here we are in the topic of practice solving NP-complete problems.

5.2 Language syntax

A rule is a triplet $(< type >, < corps >, < weight >)$.

• <type> can take 2 values : $hard$ or def. If the value is $hard$ the rule is an hard-rule and represents an Horn clause, which is sure and non-revisable. If the value is def the rule represents a normal default.

• <weight> weights the rule. These weights will make it possible to choose between the different extensions proposed by the algorithm.

• <corps> is a couple (L, R). The left element L is a set of literals $(l_1, ..l_n)$ perhaps empty. This set is identified to $l_1 \wedge .. \wedge l_n$. The right element R is either a single literal or empty. If the rule is hard, the couple (L, R) represents the formula $L \rightarrow R$. If the rule is a default, the couple represents a normal default $\frac{L:R}{R}$ An increased attention is done to these two cases.

Hard Rules

A hard rule (L, R) represents the formula $L \rightarrow R$ where L is a conjunction of literals and R a literal. How we decided to restrict our algorithm to Horn clauses all literals of L are positive. The literal R can be positive or negative. Here we have two special cases. :

1) L is empty. Therefore the rule represents a positive or negative unary clause. The unary clauses are elementary sources of information. They did not contain variables, they are ground clauses. This allows the decidability of the algorithm. However the other clauses can contain variables, leave the pure propositional calculus.

2) R is empty. For this empty-consequence, the rule $L \rightarrow \emptyset$ is equivalent to $\neg L$ equivalent to a negative clause. For example, we can use such a clause to represent a mutual exclusion "It is impossible to trigger and to block a protein at the same time".

Default Rules

If the rule (L, R) is a default, then it represents a normal default, the prerequisite is L, and R is the justification and also in the same time the consequent. If the prerequisite is empty, the default is without justification. By definition of the defaults it is impossible to have an empty consequence. Contrary of the hard rules the prerequiste R can contains negative literals.

5.3 Cell Signaling Pathway Representation

We have worked on the bibliographic data of the response to DSBs translated on a map of molecular interactions Figure 2 given by Pommier et al. [21]. A draw back of this map is that it is very difficult to add a new interaction or protein without full reassessment. In particular the management

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

of conflicts (for example simultaneous trigger and block interaction) is very difficult. So we worked on the translation of this map into our language. Initial results have translated this map and tested some algorithms [14], [15].

Today, the map is translated by 206 rules in a very natural way, without having to "tweak" the predicates or the rules. The rules are expressed in the syntax above. These rules can be hard rules or defaults. With our syntax it is very simple to change the nature of the rules to test different configurations. We can calculate the extensions in a very short time. We never needed to use non Horn clauses. This reinforces our opinion that it is possible to use a nonmonotonic logic and also abduction and also time, on real applications.

5.4 Rule Examples

In the context of cell pathways, a predicate can be an action on one or more protein. For example :

$product(P), binding(P, Q, R), block\text{-}bindind(P, Q),$
$stimulation(P), phosphorylation(P)$
$dissociation, transcription\text{-}activating..$

The predicate can also represent properties on protein concentration :

$concentration(P, > 1000),$
$incrasse(P)$ and $decrease(P)..$

We give here some examples of rules written for our example :

$hard : stimulate(dsb, dna)$

that is an elementary fact (a ground unary clause) who says that dsb stimulates ADN.

$def : stimulate(dsb, dna) \rightarrow product(altered\text{-}dna)$

that is default rule "Generally when dsb stimulates DNA, altered DNA is produced.

$hard : product(p\text{-}atm\text{-}atm\text{-}bound) \rightarrow \neg product(atm\text{-}atm)$

that is a negative clause.

$product(p\text{-}s15\text{-}p53\text{-}mdm2) \wedge product(p\text{-}chk1) \rightarrow phosphorylation(p\text{-}chk1, p\text{-}s15\text{-}p53\text{-}mdm2)$

Using a simple logic formalism can express much of what biologists are needed to represent.

6. Algorithm and implantation.

The algorithm is written with SWI Prolog.

A rule :

$(< type >, < corps >, < weight >)$

is represented by a unary Prolog clause:

$rule(< type >, < corps >, < weight >).$

Therefore, the rules and the algorithm are in the same Prolog program, which is very practical. Another advantage to use Prolog is that the unification, the backtracking and the lists management are well optimized. Of course Prolog is interpreted, so it is slower than compiled languages (but not that much). In the other hand Prolog programs are short and simple, which saves a lot of time to test programs and heuristics.

This algorithm calculates the extensions. As the clauses are Horn clauses and as the defaults are normal, the research tree is optimized. Particularly it is easy to calculate extensions without duplication (we do not calculate several times the same extension). For algorithms, we can also use a weak form of negation as failure [6,28].

For initial tests, given by the map of the entire network of Pommier [21], we can calculate all extensions in a short time. For example with most of the rules by default, there are two extensions. The calculation takes 500000 LIPS and 0.4 seconds of CPU time on MacBook. The temporal aspect of gene networks has been tested for small examples, but the scaling has not yet been done. For the abduction, it is almost the same. The algorithm has been tested on small examples and passing the scale remains to be done, but again that should be possible. There are no theoretical problems.

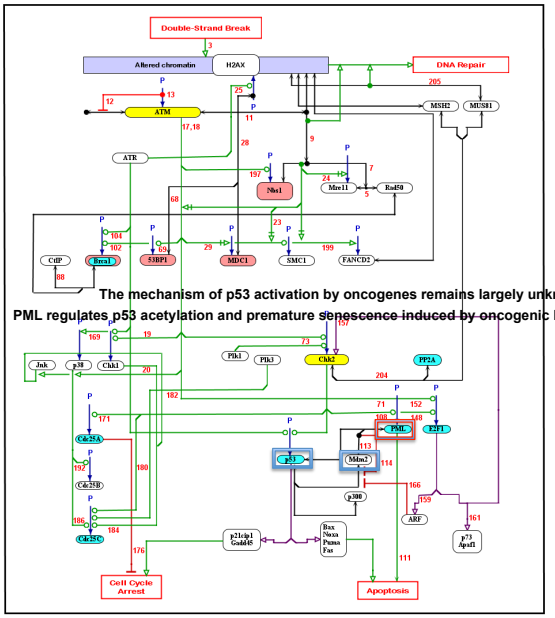


Fig. 5: DNA double strand break generated automatically by using the most relevant extension.

7. Results

Basically, many researchers are trying to complete the Signaling Map. In our approach the map is simplified which is very useful for biological experiments. Introducing time in defaults (the prerequisite considered at time t and the conclusion at time $t + 1$), we obtained a simplified map of Pommier. The most interesting result is the identification of the molecule "X" from figure 1 as be PML which regulates p53 acetylation and premature senescence induced by oncogenic Ras.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

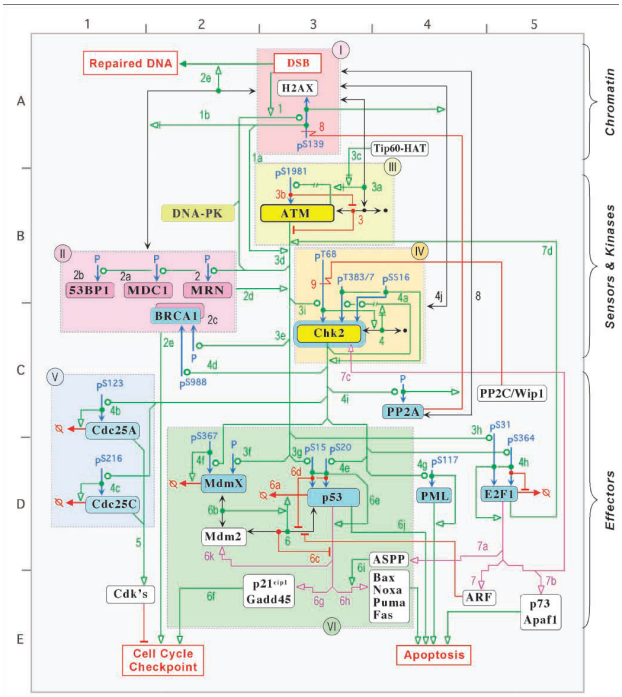


Fig. 6: The Molecular Interaction Map and Rationale for Chk2 build-up "manually" by Pommier [22] using biological cause-effect graph.

At first we tested the notion of production fields and a consequence finding algorithm for producing clauses [7], [19]. We tested also the SOLAR language that uses production field and this algorithm. The results are mixed in the case of "big" examples. We also looked at the ASP formalism [17]. Again the impression is mixed. Indeed ASP deal mainly with normal defaults without prerequisites. Getting all the power representation of defaults with prerequisite is possible by rewriting techniques. But you lose a lot of clarity and also efficiency.

By generating automatically the DNA double strand breaks map (Figure 5) we noticed that the protein p73 is not directly involved in activation of apoptosis as in the case of Pommier map (Figure 6). This result obtained from incomplete knowledge constitutes a theory formation framework for "Knowledge Discovery" using Default Logic.

8. Conclusion.

The A.I. challenge is to explain new phenomena using automatic causal discovery. To do that, we introduced formalism able to infer signaling pathway by using defaults approach and abductive reasoning. In this paper we define a new approach for the build up automatic the Double Strand Break Signaling Pathway. This map keeps only relevant proteins and it is very close to the bioregulatory network related to the histone γ-H2AX-ATM-Chk2-p53-Mdm2 pathway defined by Pommier.

References

[1] N. Tran, C. Baral, *Hypothesizing and reasoning about signaling networks.* Journal of Applied Logic, 7, 253-274,2007.

[2] CH. Bassing, FW Alt. *H2AX may function as an anchor to hold broken chromosomal DNA ends in close proximity.* Cell Cycle 2004; 3:149-53.

[3] J. Bartkova, Z. Horejsi,et al., *DNA damage response as a candidature anticancer barrier in early human tumorigenesis.* Nature 2005; 434:864-70.

[4] B. Benhamou, P. Siegel, *Symmetry and Non-Monotonic Inference.* Proc. Symco'08, Sydney, Australia, Sept. 2008.

[5] B. Benhamou,T. Nabhani, P. Siegel, *Reasoning by symmetry in non-monotonic logics.* Proc. 13th international workshop on Non-Monotonic Reasoning NMR 2010, Toronto, Canada, mai 2010.

[6] B. Benhamou,L. Paris, P. Siegel, *Dealing with Satisfiability and n-ary CSPs in a logical framework.* Journal of Automated Reasoning,Volume 48, Number3, Pages 391-417, 2012.

[7] J.M. Boi, E. Innocenti, A. Rauzy, P. Siegel, *Production Fields : A New approach to Deduction Problems and two Algorithms for Propositional Calculus.* Revue d'Intelligence Artificielle, 25(3) : 235-255, 1992.

[8] G. Bossu, P. Siegel. *Saturation, Nonmonotonic Reasoning and the Closed World Assumption.* Artificial Intelligence, 25(1) :13-63, 1985.

[9] M.O. Cordier, P. Siegel, *A Temporal Revision Model for Reasoning about World Change.* Proc KR 1992 p. 732-739, 1992.

[10] A. Doncescu, Y. Yamamoto, K. Inoue, *Biological systems analysis using Inductive Logic Programming.* Proc. of the 21st International Conference on Advanced Information Networking and Applications (AINA 2007), pages 690-695, IEEE Computer Society, 2007.

[11] A. Doncescu, K. Inoue, Y. Yamamoto, *Knowledge-based discovery in systems biology using CF-induction. New Trends in Applied Artificial Intelligence.* Proc. 20th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA / AIE 2007), Lecture Notes in Artificial Intelligence, volume 4570, pages 395-404, Springer, 2007.

[12] A. Doncescu, J. Weisman, G. Richard, G. Roux, *Characterization of bio-chemical signals by inductive programming.* Knowledge Based Systems, 15 (1), 129-137, 2002.

[13] A. Doncescu, T. Le, P. Siegel, *Default Logic for Diagnostic of Discrete Time Systems.* Proc. BWCCA-2013 - 8th International Conference on Broadband and Wireless Computing, Communication and Applications p. 488-493, Compiegne, France, Oct 2012

[14] A. Doncescu, P. Siegel, *The Logic of Hypothesis Generation in Kinetic Modeling of System Biology,* Proc. 23rd IEEE International Conference on Tools with Artificial Intelligence, p. 927-929, Boca Raton, Florida, USA, Nov. 2012

[15] A. Doncescu, T. Le, P. Siegel, *Utilization of Default Logic for Analyzing a Metabolic System in Discrete Time.* Proc.13th International Conference on Computational Science and Its Applications, ICCSA 2013, p. 130-136, Ho Chi Min, Vietnam, June 2013.

[16] D. Kayser, F. Levy, *Modeling symbolic causal reasoning,* Intellecta 2004, 1, 38, pp 291-232, 2004

[17] E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, H. Turner, *Non-monotonic causal theories.* Artificial Intelligence, No. 1-2 vol.153 pp.49-104, 2004.

[18] K. Inoue,A. Doncescu, H. Nabeshima. *Completing causal networks by meta-level abduction.* Machine Learning, 91 (2) :239-277, 2013.

[19] H. Nabeshima, K. Iwanuma, K. Inoue, O. Ray. *SOLAR: An automated deduction system for Finding consequence.* AI Commun, 23 (2-3): 183-203 (2010)

[20] R. Ostrowski, L. Paris, L. Sais, P. Siegel, *Computing Horn Strong Backdoor Sets Thanks to Local Search.* ICTAI'06, p. 139-143, IEEE Computer Society, Washington D.C., US, nov. 2006

[21] Pommier Y. and all. *Targeting Chk2 Kinase : Molecular Interaction Map and Therapetic Rationale.* Current pharmacy design, 11(22):2855-72, 2005.

[22] Pommier Y. and all. *Chk2 Molecular Interaction Map and Rationale for Chk2 Inhibitors* Clin Cancer Res. 2006 May 1;12(9):2657-61..

[23] R. Reiter *A Logic for Default Reasoning.* Artif. Intell. 13(1-2): 81-132 (1980).

[24] T Sato, Y Kameya. *PRISM: a language for symbolic-statistical modeling.* International Joint Conference on Artificial Intelligence 15, 1330-1339.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

A Hybrid Clustering Algorithm and Functional Study of Gene Expression in Lung Adenocarcinoma

Mohammad Shabbir Hasan and Zhong-Hui Duan
Department of Computer Science, University of Akron, Akron, Ohio, USA

Abstract - *DNA Microarray technology provides a convenient way to investigate expression levels of thousands of genes in a collection of related samples during different biological processes. Researchers from diverse disciplines such as computer science and biology have found it interesting as well as meaningful to group genes based on the similarity of their expression patterns. Hierarchical clustering and k-means clustering are commonly used algorithms to group genes with similar expression patterns. However, in spite of having some advantages such as producing tighter cluster than other algorithms, k-means clustering has some limitations also. The performance of k-means clustering algorithm largely depends on the selection of the value of k i.e., the number of clusters. In this research work, we proposed a new method to combine k-means clustering with hierarchical clustering to overcome the limitation. To test the algorithm, we used microarray data on lung adenocarcinoma, the most common type of non-small-cell cancers. We identified a number of representative genes from the group of normal tissue and from the group of KRAS mutation tissues. Genes for both of these groups were clustered using our proposed method. Finally we conducted functional investigation of the differentially expressed genes using Gene Ontology database to find changes in the enrichment of molecular functions of the genes contained in each cluster of both normal and KRAS positive groups. We discovered that our proposed method can group genes with similar expression pattern together and hence it can be used in future for clustering microarray data*.

Keywords: Gene Expression; Microarrays; K-means Clustering; Hierarchical Clustering; Gene Ontology

1 Introduction

In gene expression, gene products such as proteins or RNA are created from the inheritable information contained in a gene [1]. So far traditional molecular biology has focused on studying individual genes in isolation for determining gene functions. However it is not suitable for determining complex gene interactions as well as explaining the nature of complex biological processes. For this purpose, examining the expression pattern of a large number of genes in parallel is required [2]. DNA microarray technology which is one of the most important tools now-a-days for the analysis of gene expression patterns has made it possible to view thousands of genes expression levels in parallel [3]. This analysis is very useful to get information for diagnosis of different diseases and efficient algorithms are required to analyze DNA microarray datasets accurately. It is believed that a group of genes with similar gene expressions are likely to have related gene functions [4]. Hence identifying genes with similar expression patterns in different phases of the cell cycle or in different environmental conditions is an important task.

Clustering algorithms play an important role in gene analysis by separating a dataset of heterogeneous genes into homogeneous groups containing similar genes. It helps to analyze a group of genes instead of analyzing each one individually. After getting appropriate clusters, researchers can further investigated the clusters to find distinct pattern for each cluster as well as more information about functional similarities and gene interactions. A good number of algorithms have been developed for clustering DNA microarray data so far. These algorithms include k-means clustering [5], hierarchical clustering [6-8], self-organizing maps [9-11], support vector machines [12], Bayesian networks [13] and fuzzy logic approach [14]. Beside these algorithms, some algorithms use other genomic information along with gene expression data in order to improve clustering efficiency. Examples of such algorithms include [15] that use gene ontology data with gene expression data and [16-18] that clusters genes by using information of upstream regions of the coding sequences with gene expression profiles to get more biologically relevant clusters.

K-means clustering algorithm is computationally faster than hierarchical clustering and produces tighter clusters than hierarchical clustering. On the other hand, hierarchical clustering algorithm does not require the number of clusters to be known in advance and computes a complete hierarchy of clusters. Beside these advantages, however, both of these algorithms suffer from some limitations. The performance of *k*-means clustering depends on how effectively the initial number of clusters i.e. the value of *k* is determined. Moreover, these algorithms are computationally expensive which impede the wide use of these algorithms in gene expression data analysis [19-21]. To overcome these limitations, a combined hierarchical *k*-means clustering method has been proposed in [22] which firstly applies *k*-means algorithm in each cluster to determine *k* clusters and then feed those clusters to hierarchical clustering technique to shorten merging clusters time while generating a tree-like dendrogram. But still this algorithm suffers from the limitation of determining the initial value for *k*.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

In this paper we present a new algorithm that combines both hierarchical clustering and *k*-means clustering. The goal is to take the advantages of both algorithms to overcome the limitations of *k*-means clustering algorithm. We use the result of hierarchical clustering to decide the initial number of clusters and then feed this information to *k*-means clustering to obtain the final clusters. In microarray data analysis, clustering genes to find out the biologically relevant groups based on their expression profiles is one of the basic techniques. Similarity in gene expression profiles indicates similarity in their gene functionalities also [23]. After getting the new clusters, we explore the change in enrichment of molecular functionalities of the genes of each cluster for normal tissue and adenocarcinoma lung cancer tissue by using Gene Ontology (GO) annotations.

2 Materials and Methods

Lung adenocarcinoma is the most frequent type of non-small-cell lung cancers (NSCLC) and it accounts for more than 50% of NSCLC and the percentage is increasing [24]. Recent studies have shown that activation of the EGFR, KRAS and ALK genes defines 3 different pathways which are responsible for a considerable fraction (30%–60%) of lung adenocarcinoma [25-29]. The dataset used in this research contains expression profiles for 246 samples where 20 samples belong to normal lung tissue. Out of 226 lung adenocarcinomas samples, 127 are with EGFR mutation, 20 with KRAS mutation, 11 with EML4-ALK fusion and 68 samples are with triple negative cases. Platform used for this dataset is GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. This dataset was collected from GEO database (accession number GSE31210). The dataset contains 54675 genes and out of the 246 samples, for this research work, we considered 40 samples that consist of 20 samples from normal tissue and 20 samples from KRAS positive tissues.

To determine the differentially expressed genes, we performed paired Student *t*-test and Bonferroni correction followed by the calculation of the value of fold change of the genes. In this study, after performing Bonferroni correction, we selected the genes as the most differentially expressed which have adjusted *p*-values ≤ 0.05. In addition to that, we considered only those genes where the value of fold change (increase or decrease) is significant i.e. the average fold change between cancer and normal is greater than or equal to 2. Beside these preprocessing, we considered only those genes that are associated with molecular functions according to the Gene Ontology (GO). Figure 1 shows the flow diagram of the data preprocessing. After performing *t*-test, we obtained 21,880 genes which had significant *p*-value (≤ 0.05). We performed Bonferroni correction on these genes and found 1,988 genes which had a significant adjusted *p*-value (≤ 0.05). Adding the fold change criterion, we reduced the set of differentially expressed genes to 1,005. We then performed another step of filtering to keep only those genes that have Gene Ontology (GO) terms and responsible for molecular functions. Finally we came up with 464 genes in the dataset. The final dataset which is also termed as filtered dataset in this paper is given partially in Table 1 and the complete dataset is available in [30].

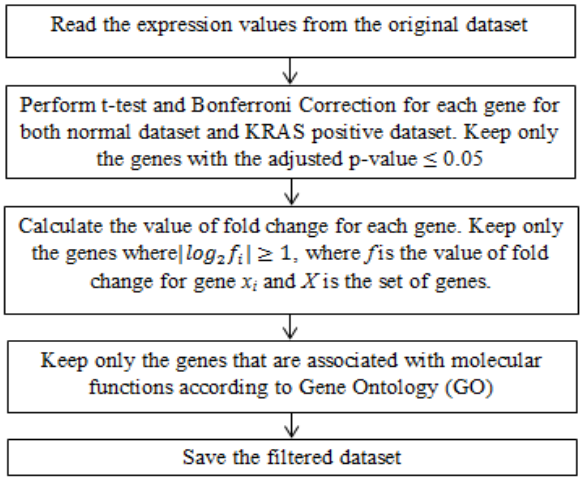


Figure 1: Flow diagram of data preprocessing

To overcome the limitation of k-means clustering algorithm, in our proposed method, we selected the value of *k* i.e. the number of clusters in a systematic way. Initially we use the hierarchical clustering algorithm for clustering the dataset and then check at which level the distance between two consecutive nodes of the hierarchy is the maximum and from this result the value of *k* is determined which is then used as the value of *k* for the *k*-means clustering. In both algorithms, Pearson correlation coefficient (*r*) is used as the similarity metric between two samples and 1-*r* is used as the distance metric.

Table 1: A brief overview of the final dataset

Affymatrix ID	Gene Symbol	Samples		
		GSM 773551	…	GSM 773784
1555579_s_at	PTPRM	3441.22	…	3569.13
211986_at	AHNAK	4395.68	…	7080.40
222392_x_at	PERP	21707.73	…	11350.53
236715_x_at	UACA	1303.01	…	1867.76
244704_at	NFYB	124.08	…	277.49
…	…	…	…	…
211237_s_at	FGFR4	22.41	…	11.07
203980_at	FABP4	257.25	…	920.44
207302_at	SGCG	47.09	…	9.61
210081_at	AGER	241.63	…	2001.28
217046_s_at	AGER	132.42	…	1016.05

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

3 Results and Discussions

Figure 2 shows the hierarchical clustering of normal tissue dataset. There are 463 interior nodes in the tree where each node is labeled based on the increasing order of its height. Therefore the root has its ID 463. To determine the number of clusters from the output of hierarchical clustering, we used a bar graph to show the difference of height between two consecutive interior nodes and it is shown in Figure 3.

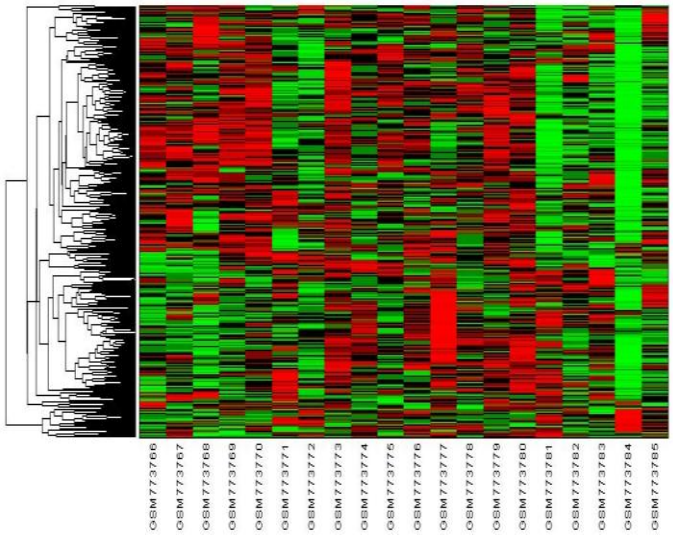


Figure 2: Hierarchical clustering of normal tissue dataset

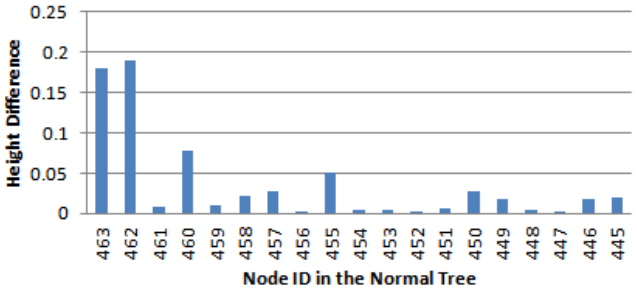


Figure 3: Height difference between two consecutive interior nodes in the hierarchical tree generated from the normal tissue. Since Pearson distance is used, the maximum height of the tree is 1.

From Figure 3 we can see that the difference is the maximum for node 461 and node 462. As there are total 463 nodes in the tree, node 461 is in level 3 from the top. So according to the approach we are discussing here, the total number of clusters for k-means clustering should be 4.

Similarly we can determine the number of clusters for the KRAS positive dataset. Figure 4 shows the hierarchical clustering of KRAS positive dataset and the height difference between two consecutive nodes is shown in Figure 5. The results indicate the number of clusters for KRAS positive dataset should be 4.

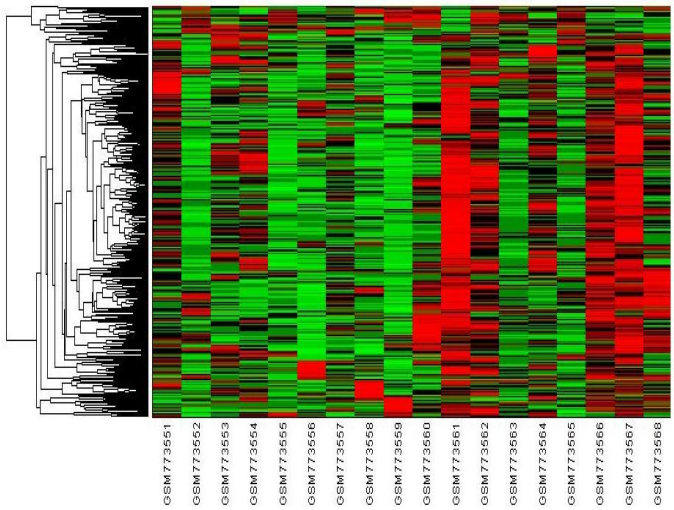


Figure 4: Hierarchical Clustering of KRAS positive dataset

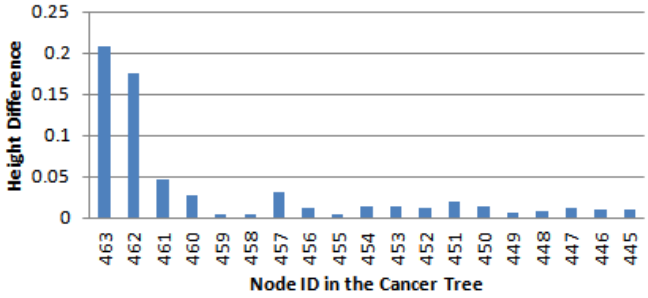


Figure 5: Height difference between two consecutive nodes in the hierarchical tree generated from KRAS positive dataset.

Clearly we see different clusters formed from normal tissue and cancer tissue. We explore their common features (genes) and explain the change of molecular function of the genes captured in the clusters of both normal tissue and KRAS positive datasets using Gene Ontology (GO) annotations. For comparing the molecular function of the clusters of normal tissue and KRAS positive tissues, we took one cluster from normal tissue dataset and one from KRAS positive dataset which have maximum number of common genes. Table 2 shows the clusters we have selected for comparing their molecular functions with the number of genes they have in common.

Table 2: List of the clusters to be compared for the alteration in molecular function

Clusters to compare		Number of genes in common
Normal Tissue	KRAS positive	
Cluster 1	Cluster 1	20
Cluster 2	Cluster 3	52
Cluster 3	Cluster 4	46
Cluster 4	Cluster 2	69

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

We explain the molecular functions of the genes in each cluster using GO annotations and their relationship are represented using a directed acyclic graph (DAG) which is also termed as GO graph in this paper. To generate these graph, we used a web based tool Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) [31]. This graph displays enriched Gene Ontology IDs (GOIDs) and their hierarchical relationships in Molecular Function GO categories. Here boxes represent GO terms, labeled by its GOID and term definition. Note that significantly enriched GO terms are marked yellow. The degree of color saturation of each node is positively correlated with the significance of enrichment of the corresponding GO term. Non-significant GO terms within the hierarchical tree are shown as white boxes. In this graph, edges stand for connections between different GO terms. Edges with red color stand for relationship between two enriched GO terms, black solid edges stand for relationship between enriched and un-enriched terms; black dashed edges stand for relationship between two un-enriched GO terms.

Figure 6 and 7 shows the GO graph for the cluster 1 of normal tissue dataset and cluster 1 of KRAS positive dataset respectively.

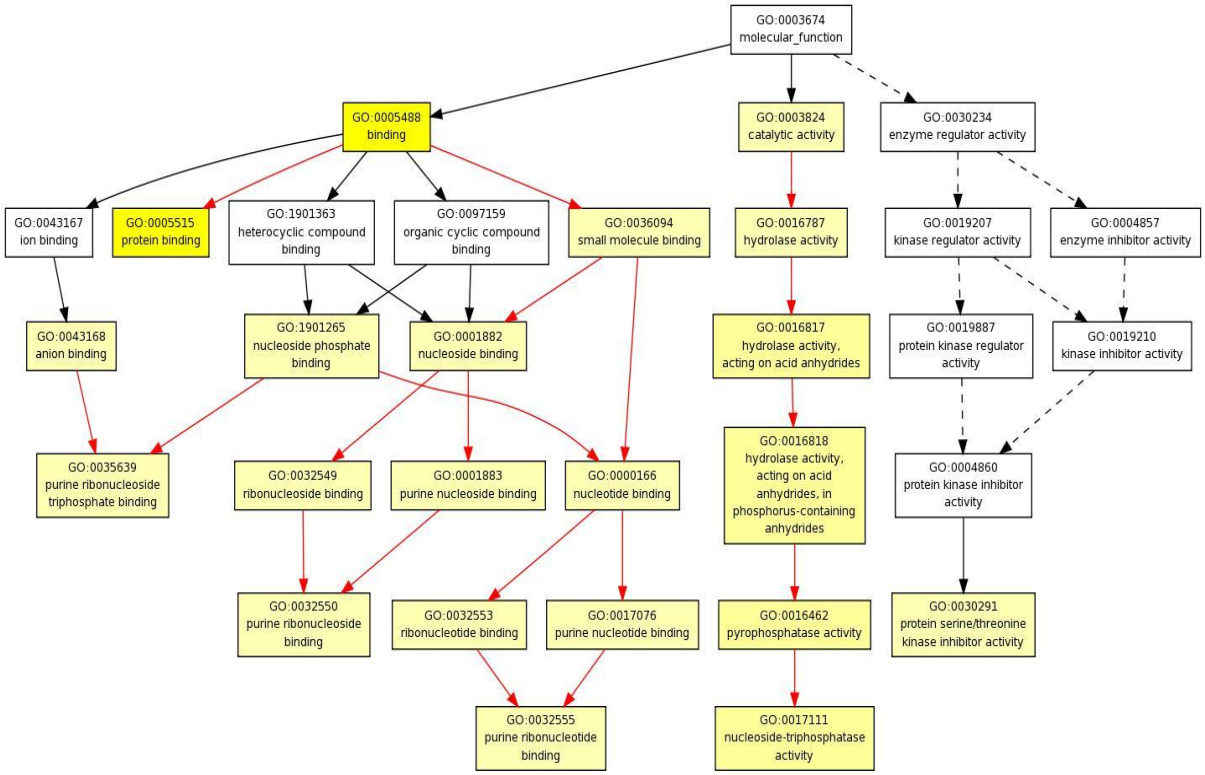


Figure 6: GO graph for cluster 1 of normal tissue data set.

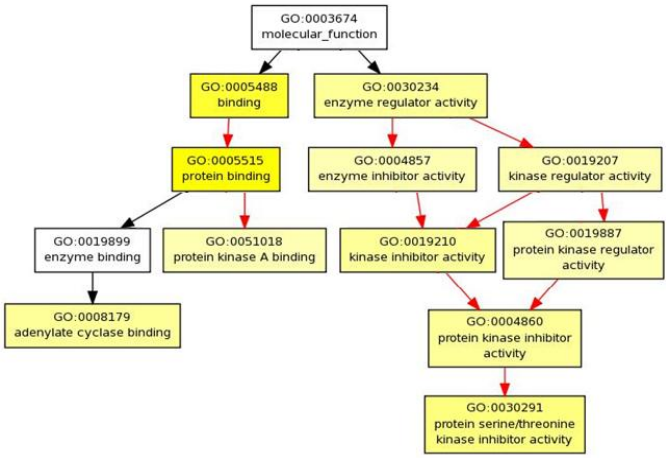


Figure 7: GO graph for cluster 1 of KRAS positive data set

In brief, from these two figures we see that, the significant GO terms GO: 0005488 (binding) and GO: 0005515 (protein binding) remain same in both clusters. GO terms such as GO: 0030234 (Enzyme Regulator Activity), GO: 0019207 (Kinase Regulator Activity), GO: 0019210 (Kinase Inhibitor Activity), GO: 0019887 (Protein Kinase regulator Activity) and GO: 0004860 (Protein Kinase Inhibitor Activity) which are un-enriched in normal tissue, become highly enriched in the KRAS positive tissues.

For better comparing the enrichment status of the two clusters, we used Multi-GOEAST which is an advanced version of GOEAST and it is helpful to identify the hidden correlation between the two clusters [31]. Figure 8 shows the comparative GO graph of the clusters discussed above.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

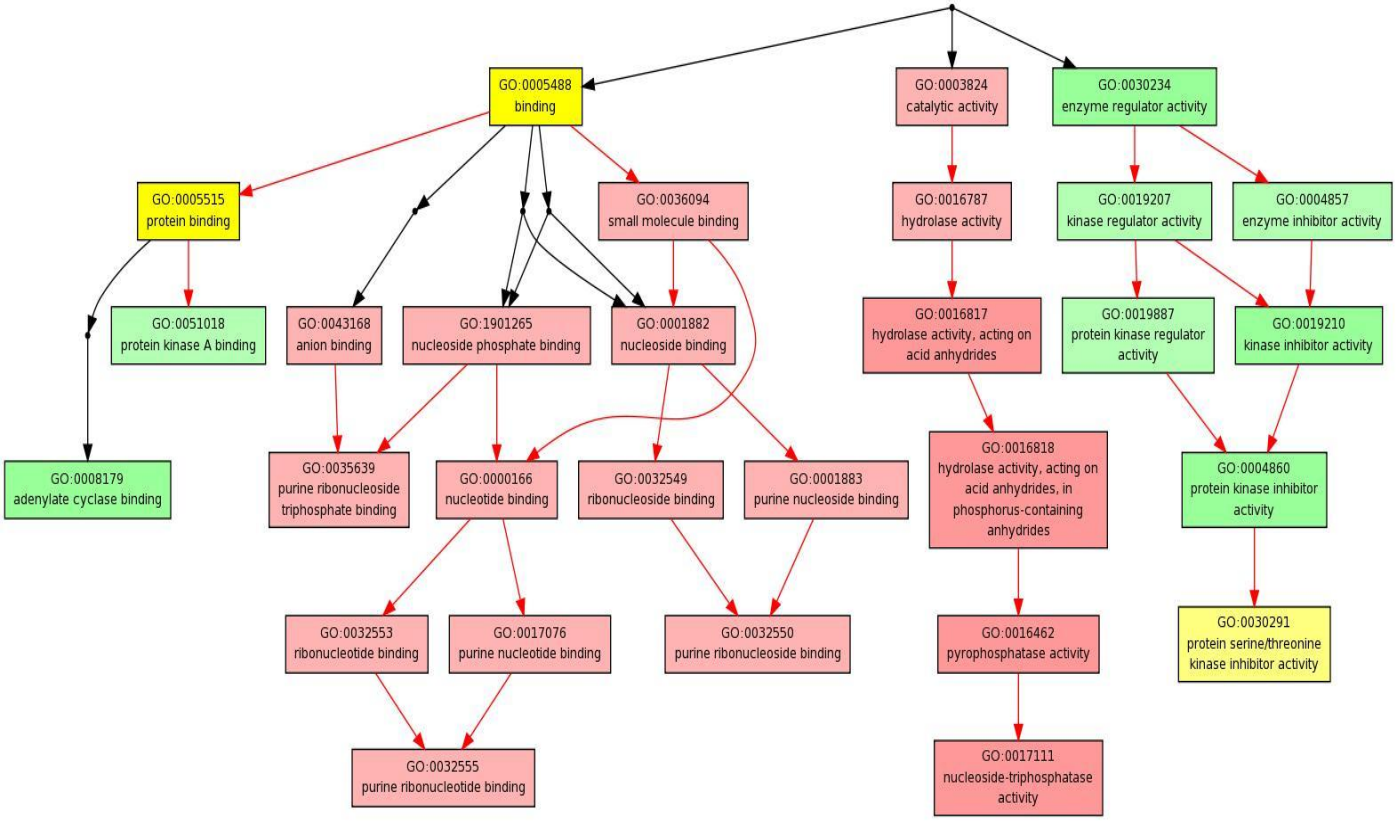


Figure 8: Comparative GO graph for comparing GO enrichment status of Cluster 1 of normal tissue dataset and Cluster 1 of KRAS positive dataset.

In the comparative GO graph, significantly enriched GO terms in both clusters are marked yellow, light yellow color indicates the GO terms which are enriched in both clusters. Nodes marked with coral pink indicate the GO terms which are enriched in normal tissue dataset but not in KRAS positive dataset. In addition to that, nodes with green color represent the GO terms which are un-enriched in normal tissue but enriched in KRAS positive tissues. Note that, the degree of color saturation of each node is positively correlated with the significance of enrichment of the corresponding GO term.

Table 3 lists the genes associated with the GO terms which are enriched in the cluster 1 of KRAS positive tissue dataset but not enriched in the cluster 1 of normal tissue dataset and these GO terms which are marked with green color in the comparative GO graph shown in Figure 8. We believe these are responsible for the alteration of the molecular activity in the cell and are linked to the development of KRAS lung cancer. Similarly we can generate and compare the GO enrichment graph for the rest of the clusters.

4 Conclusions

In this paper we proposed a combined clustering algorithm to cluster genes in a microarray dataset based on their expression levels. In the algorithm the number of clusters, i.e. the value of k which is required for k-means clustering algorithm is determined from the output of hierarchical clustering. Using this systematic way of determining the value of k, this approach overcomes the limitation of k-means clustering. This proposed method of clustering takes the advantage of hierarchical clustering to get a complete hierarchy of clusters and uses this information to determine the number of clusters to be used in k-means clustering for producing tighter cluster.

In this study we examined 40 samples and 464 genes from the dataset of KRAS lung denocarcinoma which is one of the most frequent types of non-small-cell lung cancers. Out of the 40 samples, 20 were from normal tissue and 20 were from KRAS positive tissues. We applied t-test, Bonferroni correction and fold change cutoff to find the significantly differentially expressed genes and among them only the genes having GO terms and responsible for molecular functions were included in the final dataset.

After applying the clustering algorithms, we obtained 4 clusters for both normal tissue dataset and KRAS positive dataset. Hereafter, we examined the genes contained in each cluster with respect to their molecular functions based on Gene Ontology (GO) annotation to see what are the changes

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

in the enrichment of the molecular functions of the genes took place from normal tissues to KRAS positive tissues.

In summary, we presented a coherent approach to examine alterations of molecular activities in different environmental settings such as in cancer cells. Furthermore, the proposed clustering algorithm can be generalized for clustering other types of large datasets.

Table 3: GO Terms and pathways which are enriched in molecular functions of the genes of Cluster1 of KRAS positive tissue but un-enriched in the genes of Cluster1 of normal tissue dataset

GO ID	GO Term	Associated Genes	Pathway
GO:0030234	Enzyme Regulator Activity	TIMP3	Matrix_Metalloproteinases
		CDKN1C	G1_to_S_cell_cycle_Reactome
		PAK1	Integrin mediated_cell_adhesion_KEGG
		ECT2	-----
		RALGPS2	-----
		SFN	Calcium_regulation_in_cardiac_cells
			Smooth_muscle_contraction
GO:0019207	Kinase Regulator Activity	CDKN1C	G1_to_S_cell_cycle_Reactome
		SFN	Calcium_regulation_in_cardiac_cells
			Smooth_muscle_contraction
GO:0004857	Enzyme Inhibitor Activity	TIMP3	Matrix_Metalloproteinases
		CDKN1C	G1_to_S_cell_cycle_Reactome
		SFN	Calcium_regulation_in_cardiac_cells
			Smooth_muscle_contraction
GO:0019887	Protein Kinase Regulator Activity	CDKN1C	G1_to_S_cell_cycle_Reactome
		SFN	Calcium_regulation_in_cardiac_cells
			Smooth_muscle_contraction
GO:0019210	Kinase Inhibitor Activity	CDKN1C	G1_to_S_cell_cycle_Reactome
		SFN	Calcium_regulation_in_cardiac_cells
			Smooth_muscle_contraction
GO:0004860	Protein Kinase Inhibitor Activity	CDKN1C	G1_to_S_cell_cycle_Reactome
		SFN	Calcium_regulation_in_cardiac_cells
			Smooth_muscle_contraction
GO:0051018	Protein Kinase A Binding	AKAP12	G_Protein_Signaling
GO:0008179	Adenylate Cyclase Binding	AKAP12	G_Protein_Signaling

5 References

[1] L. Hunter, "Artificial intelligence and molecular biology," in Proceedings of the tenth national conference on Artificial intelligence, 1992, pp. 866-868.

[2] G. S. Michaels, D. B. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi, "Cluster analysis and data visualization of large-scale gene expression data," in Pacific symposium on biocomputing, 1998, pp. 42-53.

[3] N. Speer, C. Spieth, and A. Zell, "A memetic co-clustering algorithm for gene expression profiles and biological annotation," in Evolutionary Computation, 2004. CEC2004. Congress on, 2004, pp. 1631-1638.

[4] D. W. Mount, "Sequence and genome analysis," New York: Cold Spring, 2004.

[5] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," Nature genetics, vol. 22, pp. 281-285, 1999.

[6] F. Luo, K. Tang, and L. Khan, "Hierarchical clustering of gene expression data," in Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on, 2003, pp. 328-335.

[7] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proceedings of the National Academy of Sciences, vol. 95, pp. 14863-14868, 1998.

[8] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, et al., "Large-scale temporal gene expression mapping of central nervous system development,"

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Proceedings of the National Academy of Sciences, vol. 95, pp. 334-339, 1998.
[9] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," Proceedings of the National Academy of Sciences, vol. 96, pp. 2907-2912, 1999.
[10] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of gene expression data using self-organizing maps," FEBS letters, vol. 451, pp. 142-146, 1999.
[11] J. He, A.-H. Tan, and C.-L. Tan, "Self-organizing neural networks for efficient clustering of gene expression data," in Neural Networks, 2003. Proceedings of the International Joint Conference on, 2003, pp. 1684-1689.
[12] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," Proceedings of the National Academy of Sciences, vol. 97, pp. 262-267, 2000.
[13] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," Journal of computational biology, vol. 7, pp. 601-620, 2000.
[14] P. J. Woolf and Y. Wang, "A fuzzy logic approach to analyzing gene expression data," Physiological Genomics, vol. 3, pp. 9-15, 2000.
[15] H. Wang, F. Azuaje, and O. Bodenreider, "An ontology-driven clustering method for supporting gene expression analysis," in Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on, 2005, pp. 389-394.
[16] I. Holmes and W. J. Bruno, "Finding regulatory elements using joint likelihoods for sequence and expression profile data," in Ismb, 2000, pp. 202-210.
[17] Y. Barash and N. Friedman, "Context-specific Bayesian clustering for gene expression data," Journal of Computational Biology, vol. 9, pp. 169-191, 2002.
[18] J. Kasturi, R. Acharya, and M. Ramanathan, "An information theoretic approach for analyzing temporal patterns of gene expression," Bioinformatics, vol. 19, pp. 449-458, 2003.
[19] G. Garai and B. Chaudhuri, "A novel genetic algorithm for automatic clustering," Pattern Recognition Letters, vol. 25, pp. 173-187, 2004.
[20] K. Ushizawa, C. B. Herath, K. Kaneyama, S. Shiojima, A. Hirasawa, T. Takahashi, et al., "cDNA microarray analysis of bovine embryo gene expression profiles during the pre-implantation period," Reproductive Biology and Endocrinology, vol. 2, p. 77, 2004.
[21] N. Bolshakova, F. Azuaje, and P. Cunningham, "An integrated tool for microarray data clustering and cluster validity assessment," Bioinformatics, vol. 21, pp. 451-455, 2005.
[22] T.-S. Chen, T.-H. Tsai, Y.-T. Chen, C.-C. Lin, R.-C. Chen, S.-Y. Li, et al., "A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray," in Intelligent Signal Processing and Communication Systems, 2005. ISPACS 2005. Proceedings of 2005 International Symposium on, 2005, pp. 405-408.
[23] F. Azuaje, J. Dopazo, and J. Wiley, Data analysis and visualization in genomics and proteomics: Wiley Online Library, 2005.
[24] H. Okayama, T. Kohno, Y. Ishii, Y. Shimada, K. Shiraishi, R. Iwakawa, et al., "Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas," Cancer research, vol. 72, pp. 100-111, 2012.
[25] W. Pao and N. Girard, "New driver mutations in non-small-cell lung cancer," The lancet oncology, vol. 12, pp. 175-180, 2011.
[26] R. S. Herbst, J. V. Heymach, and S. M. Lippman, "Lung Cancer," The New England Journal of Medicine, vol. 319, pp. 1367 - 1380, 2008.
[27] F. Janku, D. J. Stewart, and R. Kurzrock, "Targeted therapy in non-small-cell lung cancer—is it becoming a reality?," Nature Reviews Clinical Oncology, vol. 7, pp. 401-414, 2010.
[28] G. Bronte, S. Rizzo, L. La Paglia, V. Adamo, S. Siragusa, C. Ficorella, et al., "Driver mutations and differential sensitivity to targeted therapies: a new approach to the treatment of lung adenocarcinoma," Cancer treatment reviews, vol. 36, pp. S21-S29, 2010.
[29] D. E. Gerber and J. D. Minna, "ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time," Cancer cell, vol. 18, pp. 548-551, 2010.
[30] (2013, 06/18/2013). Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31210
[31] Q. Zheng and X.-J. Wang, "GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis," Nucleic acids research, vol. 36, pp. W358-W363, 2008.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Gene Collector V1.1: a user-friendly bilingual gene data mining software

C. Pacheco[1]*, S.G. Lopes[2], W.P. Silva[2], J.L.C. Silva[2], R.P. Vasconcelos[2], A.A. Araujo-Neto[2], S.M.S. Felipe[1], M.M.D.C. Soares[1], J. O. Alves[1], Soares, P.M. [1], D. P. Carvalho[3], and V.M. Ceccatto[1]

[1]Instituto Superior de Ciências Biomédicas, Universidade Estadual do Ceará, Fortaleza, CE, Brazil.
[2] Universidade Federal Rural do Semi Árido, Angicos, RN, Brazil.
[3]Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro (IBCCF/UFRJ), Rio de Janeiro, RJ, Brazil.
*christinaosvaldo@yahoo.com.br

Abstract: *The amount of available biological information in public databases has grown, and in an exponential manner. In the light of this growth, scientists' data analysis requirements have rapidly changed. We developed a computer software capable of capturing and summarizing information about a gene set from public databases."Gene Collector v.1.1" is a bilingual data mining tool to aid in biomedical researches and also useful for educational uses. Using NCBI and Ensembl IDs as input, it is capable of gathering information on nomenclature (official symbol, aliases and full name), species, location (band, coordinates and strand), number of alternative transcripts and gene product type. The data set can be downloaded as a spreadsheet or as a local database. Some obtained parameters are summarized in a simple report. Using the program's output, further analyses can be performed aiming to answer specific biological questions, making useful links with other bioinformatic tools.*

Keywords: genetic data collection, bilingual data mining, biological database searches.

Paper submitted to BIOCOMP'14, the 2014 International Conference on Bioinformatics & Computational Biology.

1 Introduction

Wet-lab geneticists, molecular biologists, biomedical scientists and other life science researchers are producing and dealing with ever increasing volumes of genetic data. The amount of available biological information in public databases has grown, and in an exponential manner. In the light of this growth, scientists' data analysis requirements have rapidly changed in the last decades. In the current "*omics*" era, scientists that study molecular biology, disease mechanisms, medical genetics, structural biology and population genetics, among other areas, have found the need to comprehend basic bioinformatics, focusing in resources and tools directed towards their needs [1].

Biological databases, such as NCBI gene [2] and Ensembl Genome Browser [3], offer access to a huge amount of information about genes, transcripts and their products, encompassing sequence, location, structural and functional data. With the advances in science, approaches that deal with large amounts of information, such as data mining and statistical methods, are now seen as complementary procedures [4]. Data mining can be defined as a process that handles large quantities of information, selecting, exploring and modeling data in search for regularities or relations in large data sets. The process envisages providing comprehensible and useful information for the user [5].

Information processing and data mining tools have, therefore, become key components in modern biomedical research, especially in genetics and protein biology [4, 6]. A great number of bioinformatics tools have been developed in the past few years, in order to address a wide variety of biological research problems. The current scenario for the use of tools for data mining includes the search and processing of biological information. This field includes the discovery and accumulation of information, such as IDs, aliases, official name, location, and genetic description parameters in general. "Gene Collector" is being presented as a new approach for collection and utilization of basic genetic data in a local database.

Even simple molecular biology projects, nowadays, tend to study a gene set rather than a single gene. As is well known that database searches can be time-consuming, especially when dealing with a large number of genes, our multi-disciplinary group aimed to develop a user-friendly tool for data mining basic genetic information on large gene sets. Our main objective was the development of a user friendly software capable of capturing and summarizing information about a gene set from the NCBI Gene database and Ensembl Genome Browser through a bilingual interface.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2 Methods

Parsing, a syntactic analysis of HTML pages which transforms text into a binary tree, was used in order to create the data structure using the API JSoup, a Java library for the extraction and manipulation of HTML (http://jsoup.org). Afterwards, it was possible to access the tags and save the information in the program's variables.

Using this approach, the Gene Collector software was created in Java language, as shown in the pipeline in Figure 1. Gene Collector is capable of finding, extracting and classifying genetic information of interest from NCBI Gene (http://www.ncbi.nlm.nih.gov/gene/) [2] and Ensembl Genome Browser (http://www.ensembl.org/) [3].

It is possible to input a list of genes for the software to download data, and after data collection it brings the option of saving the spreadsheet as an ".xlms" file. Another function that Gene Collector brings is to allow that the gathered data on the gene set is saved in a local database, which, in turn, allows for searches and manual alterations. After collecting all the information, modest data mining approaches were performed, creating the "Gene Set Summary" section, bringing a summary of the collected data.

In order to evaluate the efficacy of the searches, as a quality control measure, we used a comparative method. After performing searches on a gene set both manually and using the program, we compared the table produced by Gene Collector with a data set (with 1000 genes) previously manually characterized by our group (through exhaustive DB searches), revealing the percentage of correctly and incorrectly filled cells. Data mining procedures (the gene set summary) revealing aspects of the data set such as number of genes per gene type, division of gene set by organisms, as well as chromosome distribution were also evaluated by comparative methods.

Additionally, to characterize some biological aspects of gene sets, the program output data was further evaluated with other bioinformatic online tools, using 13 correlated genes for this purpose (cholesterol transporter activity - GO:0017127). Another set of 622 genes was used for evaluation of the output data with co-localization and gene synteny analyses.

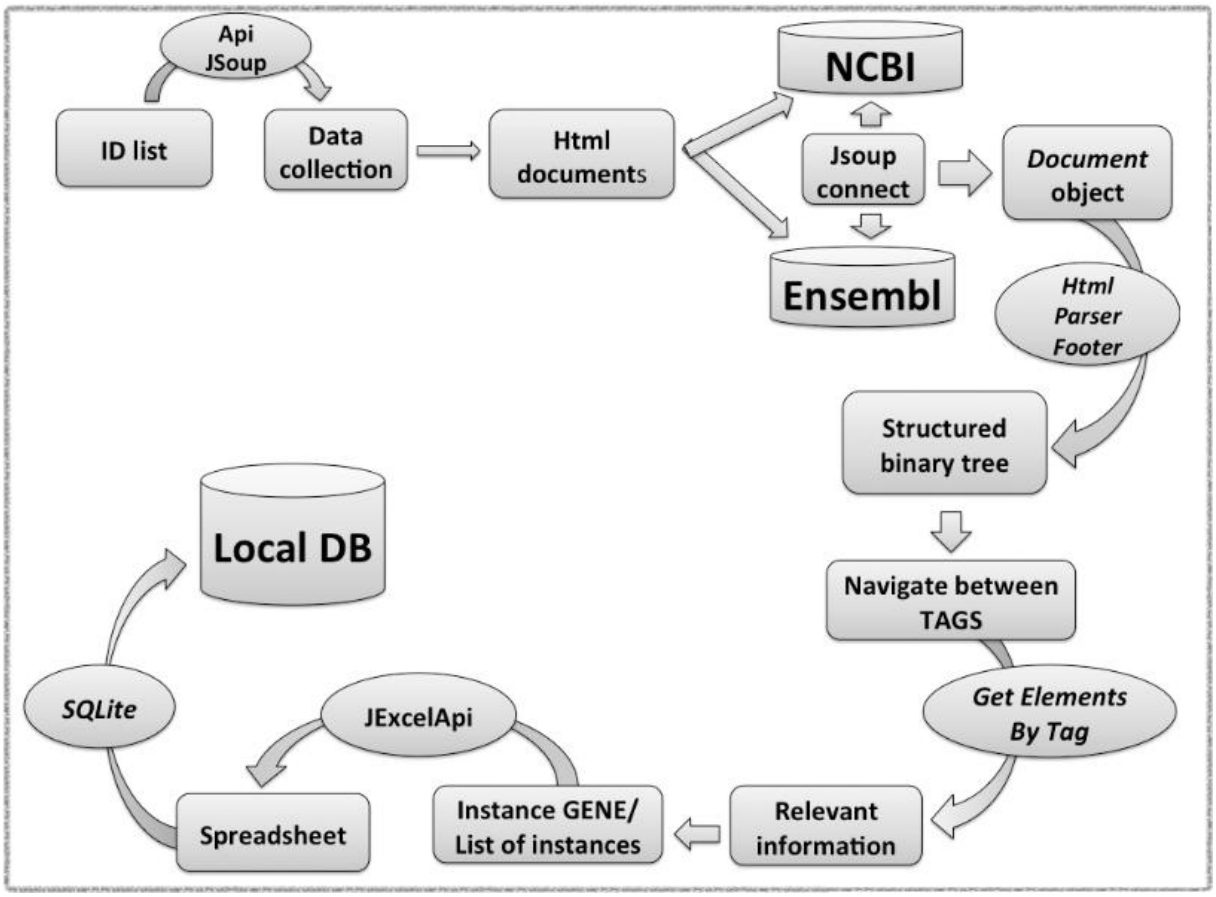


Figure 1. Program development pipeline of the "Gene Collector" software.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

3 Results and discussion

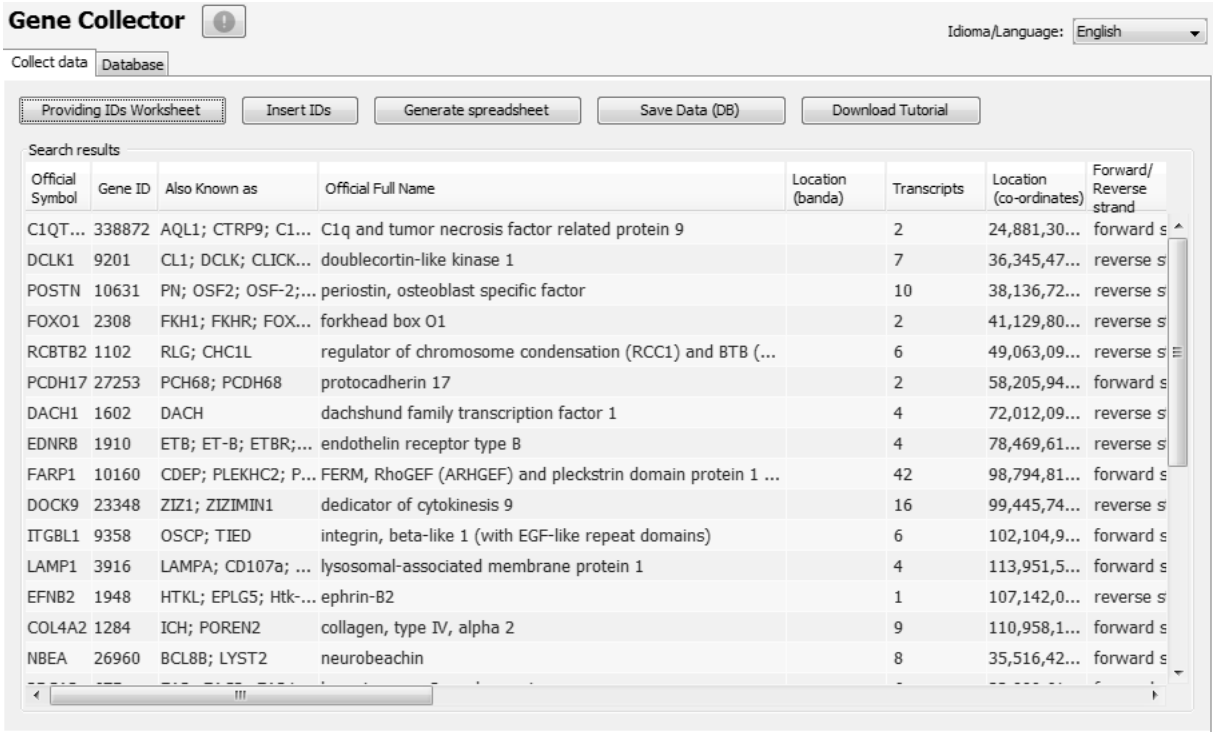
In the post-genomic era, the data explosion that is allowing for huge scientific progress is also threatening to drown scientists with ever increasing amounts of information [7]. Bioinformatics and Genomic approaches are helping in the understanding of biological systems with the development of computational tools that aim to solve biomedical problems [8]. Thousands of bioinformatics tools have already been developed, for a wide variety of biomedical research ends. However, there are still some questions that can only be solved through specific tools for the biomedical researcher.

Gene Collector is a bilingual data mining computer software capable of capturing and summarizing information about a gene set from the National Center for Biotechnology Information Gene database and Ensembl Genome Browser. Using NCBI and Ensembl IDs as input, the created algorithm is capable of gathering information on nomenclature (official symbol, aliases and full name), species, location (band, coordinates and strand), number of alternative transcripts and gene product type.The resulting data set can be downloaded as a spreadsheet or as a local database. Gene Collector's user friendly data gathering interface can be seen in Figure 2. The program is capable of showing some parameters of gene nomenclature and identification (such as official symbol, gene IDs, aliases, official full name, description, name, synonyms).

One of the most common difficulties found by molecular biology researchers is the ambiguity in gene nomenclature (aliases). Scientific literature searches require text mining methods that are able to recognize the same gene in different databases, occurring when these aliases are not associated with the same subset of documents. These problems involve a) different names associated with the same gene, and b) one name associated with multiple genes or even with non-gene meanings [9].

The interface of the Gene Collector database is shown in Figure 3. This is a reasonably simple visualization of the results obtained by the search and allows to save the effected changes and generate a useful spreadsheet. The generated local database facilitates data maintenance and review.

The evaluation of the software involved a comparison with a manually filled spreadsheet produced from data collected from 77 papers. We used a set of 1000 genes obtained from the literature which showed genes expressed in human skeletal muscle. This evaluation of the current version of the software revealed that our newly developed software was reasonably efficient in the data gathering stage. The efficiency in this stage of the process must improve in latter versions of the program. The previously characterized gene set comprising 1000 genes was investigated using the newly developed tool. A significant proportion of genes (183 genes, 18.3% of the searched gene set) were not found by the program. Of the remaining 817 genes, we found, through comparisons between the manually filled table with the generated spreadsheet, that 96.0% of the automatically collected information matched data in the manually filled spreadsheet, with some columns being filled more accurately than others, as shown in figure 4. The number of transcripts and gene type columns, for example, were filled 99.9% correctly, whereas the aliases information was only identical for 74.7% of the genes.



Official Symbol	Gene ID	Also Known as	Official Full Name	Location (banda)	Transcripts	Location (co-ordinates)	Forward/ Reverse strand
C1QT...	338872	AQL1; CTRP9; C1...	C1q and tumor necrosis factor related protein 9		2	24,881,30...	forward s
DCLK1	9201	CL1; DCLK; CLICK...	doublecortin-like kinase 1		7	36,345,47...	reverse s
POSTN	10631	PN; OSF2; OSF-2;...	periostin, osteoblast specific factor		10	38,136,72...	reverse s
FOXO1	2308	FKH1; FKHR; FOX...	forkhead box O1		2	41,129,80...	reverse s
RCBTB2	1102	RLG; CHC1L	regulator of chromosome condensation (RCC1) and BTB (...		6	49,063,09...	reverse s
PCDH17	27253	PCH68; PCDH68	protocadherin 17		2	58,205,94...	forward s
DACH1	1602	DACH	dachshund family transcription factor 1		4	72,012,09...	reverse s
EDNRB	1910	ETB; ET-B; ETBR;...	endothelin receptor type B		4	78,469,61...	reverse s
FARP1	10160	CDEP; PLEKHC2; P...	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 ...		42	98,794,81...	forward s
DOCK9	23348	ZIZ1; ZIZIMIN1	dedicator of cytokinesis 9		16	99,445,74...	reverse s
ITGBL1	9358	OSCP; TIED	integrin, beta-like 1 (with EGF-like repeat domains)		6	102,104,9...	forward s
LAMP1	3916	LAMPA; CD107a; ...	lysosomal-associated membrane protein 1		4	113,951,5...	forward s
EFNB2	1948	HTKL; EPLG5; Htk-...	ephrin-B2		1	107,142,0...	reverse s
COL4A2	1284	ICH; POREN2	collagen, type IV, alpha 2		9	110,958,1...	forward s
NBEA	26960	BCL8B; LYST2	neurobeachin		8	35,516,42...	forward s

Figure 2. Gene Collector's user friendly interface.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

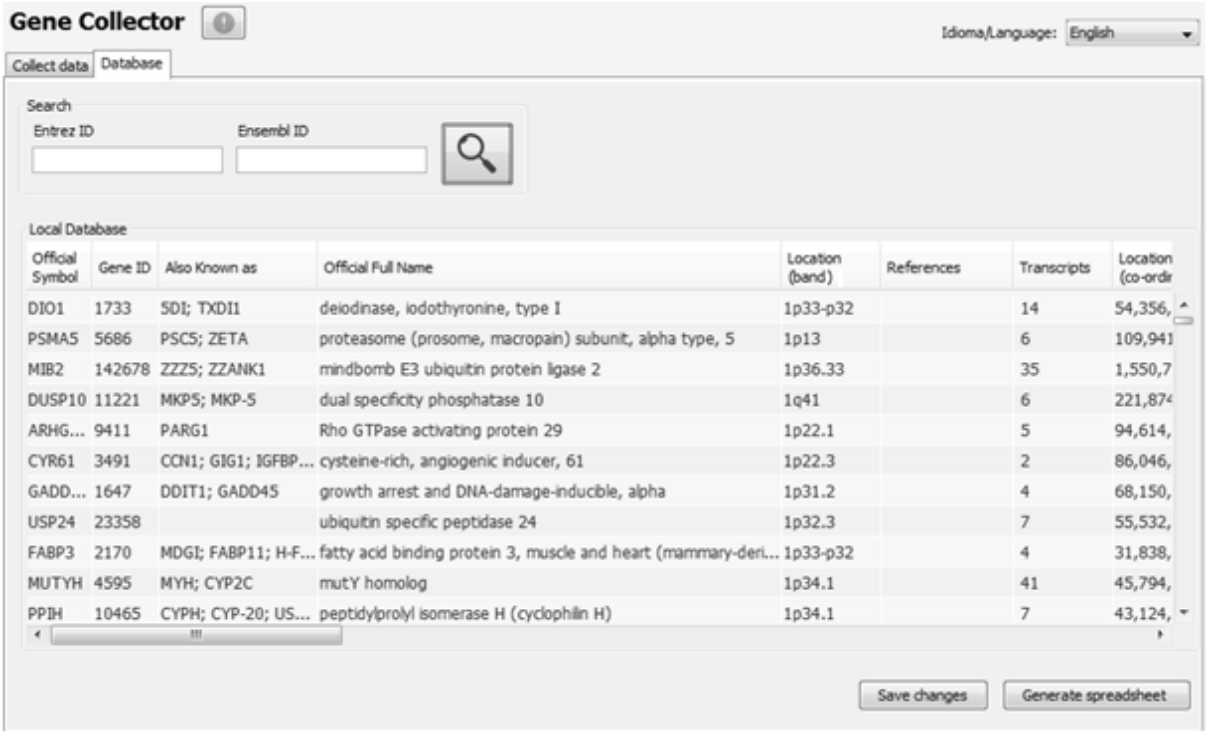


Figure 3. Local database generated by the Gene Collector software.

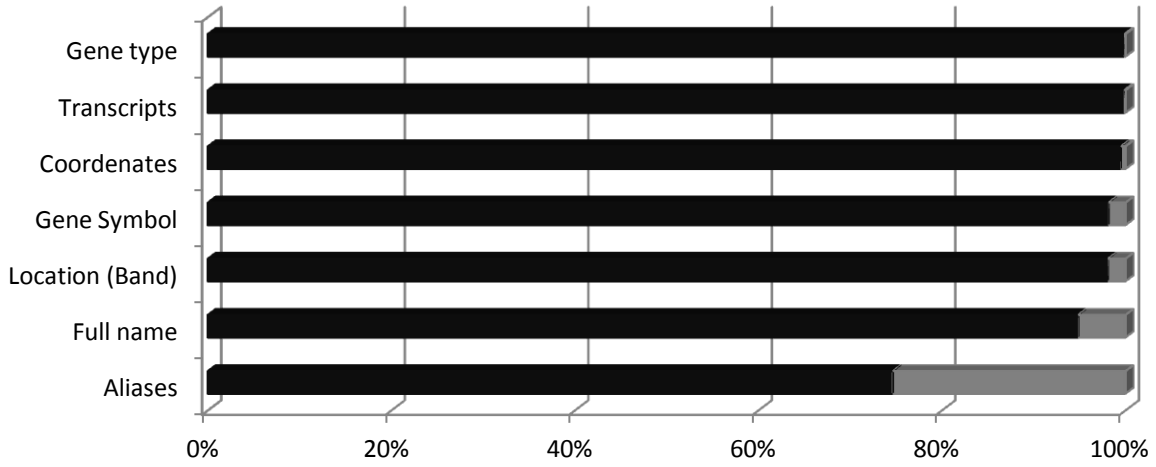


Figure 4. Data gathering quality control. Shown in black are the percentages of correctly collected data, while in grey are the percentages of incorrectly filled data.

The data summary report ("Gene Set Summary") was limited to a few parameters, such as number of genes per gene type, division of gene set by organisms, as well as chromosome distribution of the studied gene set. This is a simple but significant output which aims to summarize the collected data.

3.1. Further biological evaluations from Gene Collector v.1.1 output

One of the challenges of physiology nowadays arise from conducting large-scale analyses of the transcriptome of a tissue and the large scale of the data sets. Comparative studies of transcriptomes from different tissues or life stages from an organism allow for inferences of spatial or temporal regulation of gene expression, which may hold some additional clues to the complexity of biological systems.

Acquiring and understanding experimental data is the first step to turn large datasets useful and understandable in the search for biological patterns. The dry-lab must be accessible and comprehensible for small and large volumes of biological data. It is necessary to: a) define the questions to be made to the dataset; b) produce inferences concerning

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

biological responses and c) understand the inferences obtained. The biological question that led to the production of Gene Collector involved the need to obtain information on genes obtained by various proposals. Large datasets obtained by the use of microarray or RNA-Seq are now a reality for the biomedical researcher [11].

Some examples in biological, biomedical and physiological research are questions like: what are the metabolic pathways involved in the differentially expressed genes in a metabolic pathway? These differentially expressed genes are located in specific groups, jointly regulated? What are the connections and interactions between these genes? What is the location of these genes on chromosomes?

From the generated output it was possible to further evaluate the studied gene set and answer specific biological questions, as exemplified in figure 5. Two of the most relevant questions to physiology are a) ontology structure and b) chromosomal co-localization of genes and synteny. The bioinformatic tools used to perform in-depth gene set analyses were Gene ontology [12], Genemania [13], Enrichnet [14], Kerfufle [15] and Idiographica [16]. The results were satisfactory taking into account the objectives of the software. The sorting and organizing of data for later analysis in a user friendly manner is one of the desires of wet-lab professionals. Future requirements may occur with the advent of new bioinformatics tools.

An set of 13 genes linked to cholesterol transporter activity (GO:0017127) was used to present the possible biological uses of Gene Collector's output (figure 5). Cholesterol is the principal sterol of vertebrates and the precursor of many steroids, including bile acids and steroid hormones. The cholesterol transporter activity enables the directed movement of cholesterol into, out of or within a cell, or between cells [17]. The software produced a local database and a spreadsheet, in which the data organization enabled a comparative study of genes and a starting point for further assessments using other available tools, such as Gene Ontology, Genemania and Enrichnet [12, 13, 14].

Shared or conserved synteny describes preserved co-localization of genes on chromosomes which can reveal coregulation in genomic assembly or regulation clusters [18]. Using the software's output table on a sample gene set it was possible to obtain both the chromosome location depicted in the graphic idiogram format [16] and information on gene co-localization, which may be connected to genetic co-regulation [15].

3.2. Educational uses of Gene Collector v.1.1

English is the universal language in all fields of science and technology. Scientific language for molecular genetics is an extensive vocabulary to learn [19]. For non-English speaking countries such as Brazil, this is a continuous challenge for undergraduate students in biological and technical areas. Using the language in which the students feel most comfortable could benefit, especially when teaching bioinformatic approaches.

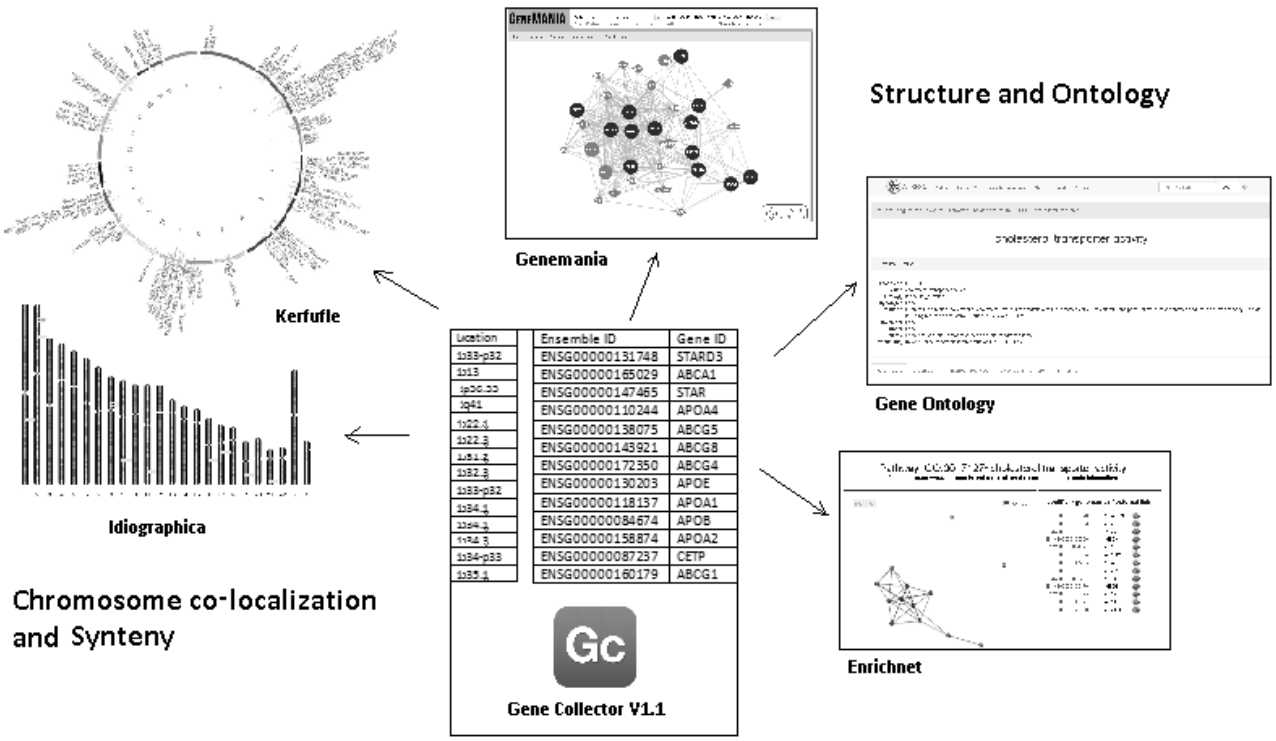


Figure 5 – Some examples of user latter from the output of the Gene Collector v.1.1.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Gene Collector can be used as an educational tool for undergraduate students, with a bilingual interface, with the option to choose the parameters, as well as the tutorials, in English and Brazilian Portuguese. Since the searched databases are in English, the returned data will still be in English, although the column header will be shown in the chosen language, facilitating the comprehension. Considering curricular issues, courses in biological, biomedical, health and correlated sciences could be improved with bioinformatics practical classes [20]. Bypassing the language barrier with this new tool can ease this process.

Gene Collector is currently undergoing registration process and will be made freely available for download at Laboratory of Biochemistry and Genetic Expression (ISCB/UECE – Fortaleza/CE - Brazil) web page at *http://www.uece.br/cmacf/index.php/salas-e-laboratorios/274*.

4 Concluding remarks and future developments

As far as we know, Gene Collector v1.1 is relatively unique and partially solves some problems concerning the organization and visualization of data from important databases. Some improvements are still needed, especially in the efficiency of gene recognition. After inputting *Entrez* and *Ensembl* gene IDs, Gene Collector is capable of gathering and summarizing information about nomenclature, species, location, alternative transcripts and gene product type. These are questions concerning the various gene expression laboratories among others demands.

In future versions of the program we aim to improve the efficiency in gene recognition, as well as include more data to be collected. Our challenges include obtaining more relevant information and reducing error percentages. Future versions may also include links with exemplified bioinformatics tools.

5 Acknowledgements

We gratefully acknowledge the financial support from "*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*" (CAPES), "*Conselho Nacional de Desenvolvimento Científico e Tecnológico*" (CNPq) and "*Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico*" (FUNCAP).

6 References

[1] Jennifer M. Williams, Mary E. Mangan, Cynthia Perreault-Micale, Scott Lathe, Neeraj Sirohi and Warren C. Lathe. "OpenHelix: bioinformatics education outside of a different box"; Briefings in Bionformatics (Oxford Journals), 11, 06, 598 – 609, Aug 2010.

[2] Donna Maglott , Jim Ostell , Kim D. Pruitt , Tatiana TatusovaT. "Entrez Gene: gene-centered information at NCBI"; Nucleic Acids Research (Oxford journals), 39, suppl1, D52-D57, Oct 2010.

[3] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos Garcıa Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Kahari, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J.P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbinoand Stephen M.J. Searle. "Ensembl 2014"; Nucleic Acids Research (Oxford Journals), 42, D1, D749 – D755, Dec 2013.

[4] Riccardo Bellazzi, Marianna Diomidous, Indra Neil Sarkar, Katsuhiko Takabayashi, Andreas Ziegler, Alexa T. McCray. "Data Analysis and Data Mining: Current Issues in Biomedical Informatics"; Methods of Information in Medicine (Schattauer), 50, 6, 536-544, Dec 2011.

[5] Paolo Giudici. "Applied data mining statistical methods for business and industry". Wiley & Sons, 2003.

[6] Eirini Papageorgiou, Ioanna Kotsioni, Athena Linos. "Data mining: a new technique in medical research"; Hormones (Greek Endocrine Society), 4, 210 – 212, Oct 2010.

[7] Peter Schattner. "Genomes, browsers, and databases. Cambridge University Press, 2008.

[8] Jack Y Yang, Mary Qu Yang, Mengxia (Michelle) Zhu, Hamid R Arabnia, Youping Deng. "Promoting synergistic research and education in genomics and Bioinformatics"; BMC Genomics (BioMed Central), 9, 11, Mar 2008.

[9] Barbara Bennani-Baiti and Idriss M Bennani-Baiti."Gene symbol precision"; Gene (Elsevier), 491, 2,103-109, Oct 2012.

[10] Eleonora de Klerk, Johan T. den Dunnen and Peter A. C. Hoen. "RNA sequencing: from tag-based profiling to resolving complete transcript structure"; Cellular and Molecular Life Sciences (Springer). May 2014.

[11] Fatih Ozsolak and Patrice M. Milos. "RNA sequencing: advances, challenges and opportunities"; Nature Review Genetics (Nature Journals), 12, 2, 87-98, Feb 2011.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[12] The Gene Ontology Consortium. "Gene ontology: tool for the unification of biology". Nature Genetics (Nature Journals). 25, 1, 25-29. May 2000.

[13] Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function"; Genome Biology (BioMed Central). 6, Supp 1, S4, Jun 2008.

[14] Enrico Glaab, Anaïs Baudot, Natalio Krasnogor, Reinhard Schneider and Alfonso Valencia. "EnrichNet: network-based gene set enrichment analysis"; Bioinformatics (Oxford Journals). 28,18, i451–i457. Sept 2012.

[15] Robert Aboukhalil, Bernard Fendler, and Gurinder S Atwal. "Kerfuffle: a web tool for multi-species gene colocalization analysis"; BMC Bioinformatics (BioMed Central), 14, 22, Jan 2013.

[16] Taishin Kin, and Yukiteru Ono. "Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat"; Bioinformatics (Oxford Journals), 23, 21, 2945-2946, Sept 2007.

[17] Curtis D. Klaassen and Lauren M. Aleksunes. "Xenobiotic, Bile Acid, and Cholesterol Transporters: Function and Regulation"; Pharmacological Reviews (The American Society for Pharmacology and Experimental Therapeutics), 62, 1, 1-96, Mar 2010.

[18] Giacomo Cavalli. "Chromosome kissing"; Current Opinion in Genetics & Development (Elsevier) 17, 443-450, Oct 2007.

[19] Chad E. Campbell and Ross H. Nehm. "A Critical Analysis of Assessment Quality in Genomics and Bioinformatics Education Research"; CBE Life Sciences Education (The American Society for Cell Biology), 12, 3, 530–541, Feb 2013.

[20] Louisa Wood and Philipp Gebhardt. ''Bioinformatics Goes to School—New Avenues for Teaching Contemporary Biology"; PLOS Computational Biology (PLOS), 9, 6, e1003089, Jun 2013.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

An algorithmic and computational approach to Open Reading Frames in short dsDNA sequences: Evaluation of "Carr's Conjecture"

[1,2,4,*] Steven M. Carr, [2]H. Todd Wareham, and [3]Donald Craig
[1]Departments of Biology and [2]Computer Science,
[3]eHealth Research Unit (Faculty of Medicine),
Memorial University of Newfoundland, St John's NL A1B 3X9 Canada,
[4]Terra Nova Genomics, Inc., St John's NL A1C 2R4
*Correspondence author, e-mail scarr@mun.ca

ABSTRACT - *In the genomic data-mining era of genetics and bioinformatics, a frequent task is the exploration of new and (or) unannotated double-stranded DNA (dsDNA) sequence data for the occurrence of protein-coding regions. The characteristic expectation is that five of the six possible three-letter reading frames for amino acids will be "closed" by one or more "stop" triplets in the Genetic Code: the sixth will be an "Open Reading Frame" (ORF) without "stops" that specifies a polypeptide sequence. The same constraint, which we designate the "5&1" condition, occurs in short dsDNA exemplars (ca. 15 ≤ L ≤ 25 bp) used in genetic and bioinformatic education. We describe an algorithmic and computational evaluation of "Carr's Conjecture," that no dsDNA sequence of L~10 or less exists that satisfies the "5&1" condition. We show there are no solutions for L ≤ 10, 96 for L=11, and that the number of solutions thereafter increases exponentially towards the upper bound 4^L. Enumeration is practically limited by CPU time beyond L=25. We describe implementation of the algorithm as a web application that generates appropriately constrained dsDNA exemplars of length L ≤ 100 bp, and their pedagogic utility and application.*

Keywords: DNA, mRNA, Genetic Code, stop codons, Open Reading Frames, data mining, webapps in genetics and bioinformatics

1 Introduction

The "cracking" of the Genetic Code by means of a rapid series of experiments and logical inferences is arguably the first instance of a "big science" approach in the history of molecular genetics [1]. Theoretical considerations had already indicated that any nucleic acid code words must comprise a minimum of three letters [2]. After demonstration in 1961 that an artificial poly-U RNA template directs incorporation of the amino acid proline into a polypeptide, and thus that UUU was the "code" for PRO, Marshall Nirenberg's lab had by 1963 deduced an incomplete "dictionary" of 50 three-letter "code words" [3], and a substantially complete Genetic Code table by 1965 [4] (Fig. 1). The iconic 4x4x4 table is now a standard feature of biology textbooks, and has been incorporated as a fundamental feature of bioinformatic computational schemes.

We consider here properties of short segments of Genetic Code that are of interest both theoretically as an unexplored computational challenge, and practically, as a pedagogic challenge for students and instructors. Taken together, solution of these challenges at the intersection of computational and biological science provides reciprocal illumination to each as an example of biological computation in 2014.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

TABLE 3. NUCLEOTIDE SEQUENCES OF RNA CODONS

1st Base	2nd Base U	C	A	G	3rd Base
U	PHE*	SER*	TYR*	CYS*	U
	PHE*	SER*	TYR*	CYS	C
	leu*?	SER	TERM?	cys?	A
	leu*, f-met	SER*	TERM?	TRP*	G
C	leu*	pro*	HIS*	ARG*	U
	leu*	pro*	HIS*	ARG*	C
	leu	PRO*	GLN*	ARG*	A
	LEU	PRO	gln*	arg	G
A	ILE*	THR*	ASN*	SER	U
	ILE*	THR*	ASN*	SER*	C
	ile*	THR*	LYS*	arg*	A
	MET*, F-MET	THR	lys	arg	G
G	VAL*	ALA*	ASP*	GLY*	U
	VAL	ALA*	ASP*	GLY*	C
	VAL*	ALA*	GLU*	GLY*	A
	VAL	ALA	glu	GLY	G

Figure 1: The Genetic Code, 1965

2 Molecular genetic and bioinformatic considerations

2.1 Molecular genetics of DNA ➔ RNA ➔ Protein

Deoxyribonucleic Acid (DNA) is famously a double-stranded molecule (dsDNA) that comprises two polymeric sequences of four bases (A, C, G, and T) in an aperiodic order that conveys bioinformation. The two strands are arranged in anti-parallel 5'➔3' directions that are implicit in the deoxyribose component. The strands are held together by non-covalent hydrogen bonds between paired A + T or C + G "base pairs". The anti-parallel arrangement and base pairing rules ensures that the alternative strands are complementary to each other. This relationship is the basis of DNA as a self-replicating molecule.

One DNA strand, designated the template strand, serves as a template for 5'➔3' synthesis (transcription) of a complementary messenger RNA (mRNA) molecule, where RNA differs from DNA in being single-stranded and substituting base U for T. The mRNA molecule is translated in the 5'➔3' direction into a polymer comprising a sequence of amino acids (a "polypeptide"), according to a Genetic Code (Fig. 1). In the Code, each of the 64 possible three-letter base sequences ("codons") read 5'➔3' specifies a particular amino acid, except that three codons (UAA, UAG, and UGA) do not specify any amino acid, and therefore serve as terminators ("stops") to polypeptide synthesis. A common Genetic Code is universal for the nuclear genomes of all organisms.

2.2 Bioinformatic data-mining

Because the mRNA sequence is complementary to that of the DNA template strand, it necessarily has the same base sequence in the same 5'➔3' direction as the DNA strand complementary to the template strand, except for the substitution of U for T. This DNA strand, designated the "sense" strand, may therefore be "read" directly from the Genetic Code table, substituting "T" for "U". As a bioinformatic process, it is straightforward to read the polypeptide sequence directly from the DNA sense strand, without the intermediate molecular steps of mRNA transcription and subsequent translation via tRNA (see below).

Any dsDNA molecule may be read from six potential starting points, designated "reading frames." Reading frames are three-base windows that commence at the 1st, 2nd, or 3rd base from the 5' end of one strand, after which each frame repeats, or from the 5' end of the other strand starting at the opposite end of the molecule. Full-length DNA sequences of several hundred to more than a thousand bases that specify protein sequences hundreds of amino acids long are expected to show that only one of these

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

reading frames is an "open" reading frame (ORF), that is, that it does not include a "stop" triplet over the required length of the polypeptide. [By definition, "codons" occur only in mRNA: the equivalent three-letter sequences in the DNA sense strand may be designated "triplets"]. As three out of 64 triplets are stops, the five alternative reading frames are expected to include multiple random stops at expected intervals of ~20 triplets: the first occurrence of a stop closes the reading frame. Commercial DNA software performs this process as a matter of routine, either from novel data or data mined from resources such as GenBank (e.g., Sequencher: Gene Codes, Ann Arbor MI). The sequence data are depicted one strand at a time, in a conventional left-to-right, 5'➔3' screen or text presentation with the inferred polypeptide sequences of one, two, or all three reading frames. The software must also be able to re-orient the data "upside down and backward" as a reverse complement simulacrum of the complementary strand in order to maintain this convention.

2.3. Pedagogic considerations: "Carr's Conjecture"

Introduction to the theory of data mining for ORFs typically begins with the propounding of a short dsDNA sequence of length L=15~25 base pairs. The exemplar sequence is constructed such that five reading frames are closed and only one is open (the "5&1" condition). The task for students is to identify the ORF and infer the correct polypeptide sequence from the Genetic Code. The challenges for instructors include construction of exemplars from scratch, where placement of multiple mutually compatible stop triplets in exactly five reading frames over a short distance is non-trivial. The double-strand nature of the DNA molecule means that specification of letters in one strand to create stops and ORFs mandates complementary changes in the other that may unintentionally create or destroy stop triplets and (or) ORFs.

The senior author therefore asked the second author for a computational algorithm that would provide exemplars that satisfied the "5&1" condition, and evaluate "Carr's Conjecture." By inspection, no solution to the "5&1" condition exists for L=5, which is the minimum-length dsDNA with three (single-triplet, single amino acid) reading frames on either strand. Then, there must exist an upper limit to L for which no solution exists, noting that exemplars of L=15 are common in teaching, and that for L=11 there are three (triple-triplet, or tri-peptide) reading frames on either strand.

3 Algorithmic & programming considerations

A practical algorithmic generator of ORF exemplars must be able to access the entire space of dsDNA sequences that satisfy the "5&1" condition for a specified L, sample that space in an at least approximately random manner, and be efficient in terms of both CPU runtime and required memory space.

As with a hand calculation, the most direct computational method would be to first generate all possible DNA sequences of length L, and then sample randomly from this generated set. Given 4^L possible sequences, this remains computationally impracticable in terms of memory and (or) runtime. Even if such a process were made more space-efficient by implementing enumeration in a recursive process that terminates as soon as an ORF exemplar is found, on inspection the small proportion of ORF exemplars relative to 4^L suggests that

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

the time required to encounter such a sequence would be prohibitive.

```
function GenerateSkeleton(S, ORFNum, rfn)
 if rfn == 7 then
  return CompleteSequence(ORFNum, S)
 elif rfn == ORFNum
  return GenerateSkeleton(S, ORFNum, rfn + 1)
 else
  res = Null
  for (pos, stopCodon) pair in a randomization of the
     list of all possible such pairs for reading frame rfn
     of S do
   St = place stop codon stopCodon at position pos relative
       to reading frame rfn of S
   if that stop placement is possible then
    res = GenerateSkeleton(St, ORFNum, rfn + 1)
    if res is not equal to Null then
     exit for-loop
  return res

function CompleteSequence(ORFNum, S)
 if S has no more unfilled bases then
  return S
 else
  res = Null
  pos = position of unfilled base in S
  for base in randomization of list ["A", "G", "C", "T"] do
   St = S
   St[pos] = base
   if the number of stops in open reading frame of St equals 0 then
    res = CompleteSequence(ORFNum, St)
    if res is not equal to Null then
     exit for-loop
  return res

function generateRandomORF(seqLen)
 let S be a sequence of length seqLen of unfilled bases
 ORFNum = random selection from reading-frame numbers 1 .. 6
 return GenerateSkeleton(S, ORFNum, 1)
```

Figure 2: Pseudocode for the two-part recursive search algorithm

Instead, we developed a method that invokes a two-level recursive search that first generates a dsDNA "skeleton" with at least one stop codon in each of five frames, and then completes the remainder of the dsDNA sequence by adding bases at random to the skeleton so as to produce an ORF exemplar in which the "5&1" condition is maintained, i.e., the sixth frame remains open. The two levels of this search are described in the algorithms *GenerateSkeleton* and *CompleteSequence*, respectively, for which pseudo-code are given in Fig. 2. As required, access to the entire space of dsDNA sequences satisfying the "5&1" condition for a specified L is complete and random by virtue of the randomization of the lists of stop triplets and stop triplet positions at each stage of the recursion in *GenerateSkeleton* and the randomization of the DNA base to be added to the skeleton completion at each stage of the recursion in *CompleteSequence*. The ORF exemplars are then produced by *GenerateRandomORF*. Note that these are not fully random, as certain sequences may be generated multiple times through appropriate completions relative to different skeletons, and will hence be over-represented in the sampled sequence space. They are, however, sufficiently random for heuristic purposes. The algorithms were implemented in the Python 2.7 programming language.

4 Mathematical considerations

We would like to derive an accurate estimate of the total number of ORF exemplars of length L ($NORF_L$), and (or) an efficient enumeration of that number.

The upper bound on $NORF_L$ is simply 4^L, the total number of sequences of length L. To derive a lower bound, we note that a subset of the ORF exemplars for any L>10 must include the ORF examples for L=11 (96: see Fig. 4), supplemented by a completion of the open frame that does not contain any of the three stop triplets. There are of course 61 such completions. Note that it is immaterial for present purposes if such a completion generates additional stop codons in the closed reading frames. The size of this subset, which sets a lower bound on $NORF_L$, is

$$
\begin{aligned}
&= 96 * 61^{\{((L - 11)/3) - 1)\}} \\
&= 2^{6.585} * (2^{(5.93)})^{\{((L - 11)/3) - 1)\}} \\
&= 2^{\{(5.93 * (((L - 11)/3) - 1)) + 6.585\}} \\
&= 2^{\{((1.976 * ((L - 11)) - 5.93) + 6.585\}} \\
&= 2^{(1.976L - 21.08)}
\end{aligned}
\qquad (1)
$$

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

The lower bound thus increases exponentially in L, which validates our use of a smart random ORF generator as described, rather than a naive all-strings random ORF generator. The recursive search algorithm, modified to an all-exemplars search that stores results and removes duplicate exemplars prior to enumeration, would therefore be expected to succumb to memory limitations. Alternatively, an exhaustive "brute force" algorithm that enumerates exemplars as they arise, without storing them, would be expected to succumb to CPU limitations, even when optimized to run on multiple machines. We implemented both alternative algorithms.

5 Results

5.1 Behavior of search algorithms with respect to L

The recursive and exhaustive algorithms show that there are in fact no solutions for L = 5~10, and 96 for L=11. Enumerations from the two methods agree for 11≤L≤19, at which point the recursive algorithm succumbs to memory limitations. For L<22, CPU usage for the exhaustive algorithm was measured on a single quad-core PC. For L≥22, CPU usage was measured over a network of such machines, and by L=25 exact CPU usage is obscured by competing demands from other users on the same network. Calculation of $NORF_L$ for L>25 with the resources available to us would require several days (Fig. 3).

5.2 Properties of $NORF_L$ with respect to 4^L

$NORF_L$ increases exponentially, and appears to converge on 4^L (Figs. 3 and 4). Although power curves calculated for values of L over the range L=19~N appear to provide close approximations of $NORF_L$ up to the (N-1) ($r^2 > 0.999$), as in Fig. 3 all such curves begin to diverge from 4^L rather than approach it, and all perform poorly as *a posteriori* predictors for L > (N-1).

The number of solutions for RFs 1-3 are symmetrical to those for RFs 4-6 respectively in all enumerations up to L=25 (Fig. 4). All RFs for L=12, 13, and 14 can encode tetra-peptides, but L=12 only for the 1st, L=13 only the 1st and 2nd, and L=14 for all three. Ordered alternatively, L=11 can encode tri-peptides in all three frames, whereas L=12 encodes a tetra-peptide only in the first, and L=13 in the first and second. Stated formally, if (L-2) and the number of amino acids are congruent modulo 3, equal numbers of amino acid residues are

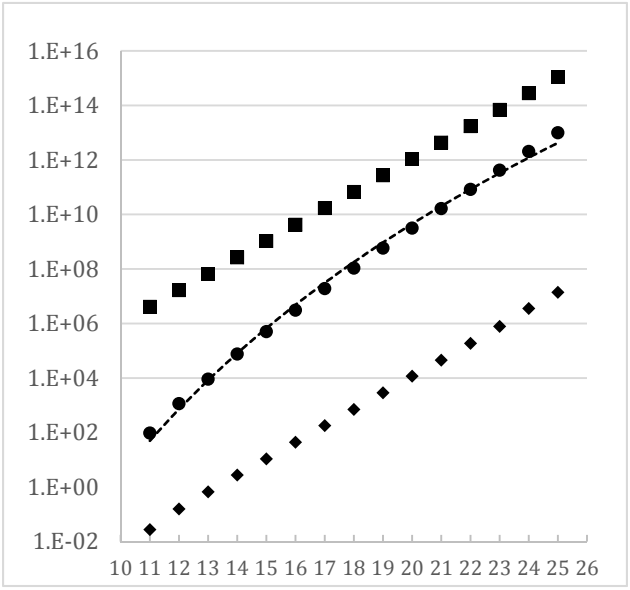


Figure 3: Semi-logarithmic plot of the enumerated number of ORF exemplars of length ($NORF_L$) for L = 11 ~ 25 (•). The power curve of best fit is $NORF_L = 6x10^{-31} \times (L)^{30.659}$ (dotted line). The upper limit on $NORF_L$ is the total number of possible dsDNA sequences of length L, 4^L (■). Required CPU time for the exhaustive algorithm is given in seconds (⧫); CPU is log-linear with respect to L, as CPU = $0.613(\log_{10} L) - 11.736$ ($r^2 = 0.99978$).

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

generated in the polypeptides from all reading frames. It is therefore counterintuitive that for L=11,14,17 there are more ORF exemplars in RFs 1 & 4, and that for L=12,15,18 and L = 13,16,19 there are more ORF exemplars in RFs 2 & 5 and 3 & 6, respectively. We suspect this is due to irregularities in strand-specific base composition imposed by the asymmetric base composition of TAA, TAG, & TGA stops.

L	RFs 1&4	RFs 2&5	RFs 3&6	Total
11	48	0	0	96
12	128	320	128	1,152
13	1,024	1,024	2,560	9,216
14	20,768	8,912	8,480	76,320
15	67,072	118,528	69,952	511,104
16	422,912	410,624	727,808	3,122,688
17	4,330,112	2,707,232	2,605,712	19,286,112
18	15,141,696	22,959,360	16,147,968	108,498,048
19	85,099,776	83,415,552	125,783,808	588,598,272
20	654,746,480	480,181,328	467,512,496	3,204,880,60
21	2.386139E+	3.311441E+0	2.565512E+0	1.652618E+1
22	1.249009E+	1.235346E+1	1.699009E+1	8.366729E+1
23	8.200810E+	6.561595E+1	6.438249E+1	4.240131E+1
24	3.058337E+	4.008699E+1	3.297290E+1	2.072865E+1
25	1.525479E+	1.518219E+1	1.969322E+1	1.002604E+1

Figure 4: Distribution of ORFs over reading frames

Similarly, the asymmetry of ORFs between RFs 1 and 3, and 4 and 6, is also unexpected. We suspect that this is influenced by asymmetries in the stop sequences, such that the [G+C] content of the two DNA strands will differ from each other. Although bases used in the skeleton completions in the webapp algorithm are selected at random, the fraction of completions with [G+C] < 0.4 for the dsDNA fluctuates over smaller values of L.

5.2 Web Application

A webapp that generates dsDNA sequence exemplars that satisfy the "5&1" condition for L ≤ 100 is available at [http://www.ucs.mun.ca/~donald/orf/biocomp/] The Python implementation was converted into a webapp with the use of HTML/CSS and JavaScript so as to allow the random ORF generator to run as a self-contained, client-side application inside a web browser. Once the page is downloaded, no further communication with a web server is necessary. Although the web application does not require a separate plug-in, JavaScript must be enabled in the web browser used to run the application.

The webapp (Fig. 5) displays a color-coded dsDNA sequence, with the top strand oriented left–to-right in the 5'➔3' direction, and reading frames (RFs) 1-3 commencing at the 1st, 2nd, and 3rd bases, respectively. The lower DNA strand is oriented right-to-left in the anti-parallel, 5'➔3' direction, and RFs 4-6 commence at the 1st, 2nd, and 3rd read from the right. The ORF is highlighted. The conventional IUPAC single-letter abbreviations for amino acids are centered over the middle base of the triplet; stop codons are indicated by asterisks (*).

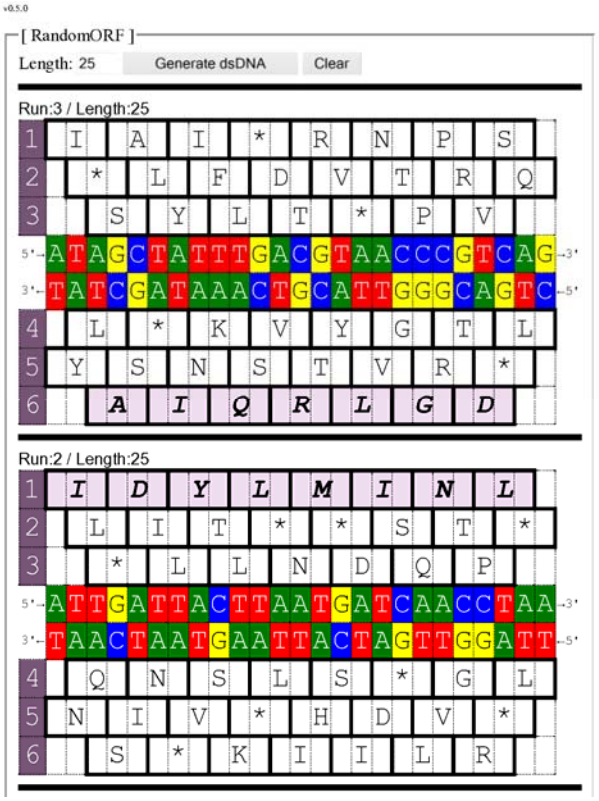


Figure 5: Screen capture of webapp. 5'➔3' ORFs in frames 1 (bottom) and 6 (top) run left-to-right and right-to-left, and encode polypeptides ***N***-IDYLMINL-***C*** and ***N***-DGLRQIA-***C***, respectively, where **N** and ***C*** are the amino and carboxyl termini, respectively.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

6 ORF exemplars in genetic and bioinformatic education

The standard "Central Dogma" rubric for extraction of information from DNA (DNA makes RNA makes Protein) requires identification of a 5'➔3' ORF in one strand of a dsDNA molecule, use of the antiparallel, complementary DNA strand as a template to transcribe an antiparallel, complementary 5'➔3' mRNA, and "translation" of the mRNA codons 5'➔3' into an amino acid polypeptide, by means of the Genetic Code.

We have found it more useful to demonstrate and emphasize the equivalence of the DNA sense strand to the mRNA, which are oriented in the same and have identical sequences except for substitution of U in the mRNA for T in the DNA. It is then more obvious that the polypeptide may be inferred logically from the DNA sense strand, with mental substitution of U for T in reading the code table. This is of course the same logic used by standard computer programs (e.g., Sequencher) to display an amino acid sequence aligned left-to-right from the input of a co-linear, single-stranded DNA sequence. The webapp presented here also introduces the information content of the dsDNA molecule, and reinforces the bioinformatic logic of data mining in a way that the standard rubric does not.

We provide a more complete discussion of the pedagogical applications of the web application in [5]. The webapp may also be a useful research tool. We are, for example, exploring the occurrence of randomly-generated medium-length exemplars in real-life data as periodic 'punctuation' that could ensure that closed frames stay closed.

7 Acknowledgements

SM Carr and HT Wareham were supported by NSERC Discovery Grants during the preparation of this MS. We thank Terra Nova Genomics, Inc., for hosting the web application. SMC dedicates this paper to Prof. William D. Stansfield of Cal Poly, San Luis Obispo, in recognition of his long service in genetics education.

8 References

[1] H. Judson, The Eighth Day of Creation, 2nd ed. Cold Spring Laboratories, Cold Spring Harbor, New York, 1996.

[2] F. H. C. Crick, The genetic code, yesterday, today, and tomorrow. Cold Spring Harbor Symposia in Quantitative Biology 31 (1966) 3-9.

[3] M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, C. O'Neal, RNA codewords and protein synthesis VII. On the general nature of the RNA code. Proceedings of the National Academy of Science (USA) 53 (1965) 1161-1168.

[4] M. Nirenberg, T. Caskey, R. Marshall, R. Brimacombe, D. Kellogg, et al., The RNA code and protein synthesis. Cold Spring Harbor Symposia in Quantitative Biology 31 (1966) 11-24.

[5] S. M. Carr, D. Craig, and H. T. Wareham. A web application for generation of DNA sequence exemplars with open and closed reading frames in genetics and bioinformatics education. CBE – Life Sciences Education, in press (2014)

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Gene Expression Profile Classification in Random Feature Space

X. Hang[1]
[1]Department of Electrical & Computer Engineering, California State University, Northridge, California, USA

Abstract - *In this study, gene expression profile classification is done via sparse representation in the random feature Space, which is obtained by either random projection or nonlinear random mapping used in Extreme learning machine* (*ELM*)*. The numerical experiment shows that sparse representation has slightly better performance than ELM.*

Keywords: Gene expression profile, Random feature, Sparse representation, Extreme learning machine

1 Introduction

A typical system for gene expression profile classification consists of two major components: feature selection and classifier. Feature selection is mainly used to reduce the dimensionality of feature space. Some methods, called gene selection, can also discover gene candidates for biological markers [1]. These methods, however, usually are data dependent, and may be affected by quality of training datasets. Many methods also have been proposed for classifier, and an excellent review can be found in [2].

Random projection has been served as a general tool for dimensionality reduction by projecting original feature vectors into reduced random feature space [3-5]. Recently it is also used as feature selection for gene profile classification [6].

The extreme learning machine (ELM) [7], as a generalized feed-forward network, is an emerging machine learning technique, where classification also takes place in random feature space. In ELM, all the weights between input nodes and hidden nodes are randomly selected as well as all the biases for hidden nodes. As a result, the outputs of all the hidden nodes can be treated as vectors in random feature space. The main difference between random projection and ELM for feature selection lies in the fact that a nonlinear function, such as sigmoid, is applied to each hidden node in ELM. Hence, the feature selection in ELM can be treated as a nonlinear mapping technique. ELM has found successful application in gene expression profile classification in [8], showing comparable or slightly better performance than support vector machines with much reduced computing complexity.

Compared with traditional feature selection methods, both random projection and nonlinear random mapping in ELM are data independent and not affected by the quality of training datasets. In this study we investigate gene expression profile classification in random feature space. We use spare representation technique [9-10] for classifier, and compare the performance with ELM.

2 Random feature space

Let $\mathbf{x} \in R^d$ denote a gene expression profile column vector with d as the number of genes. Assume the dimensionality of the random feature space is m.

2.1 Radom projection

Random projection can be achieved by

$$\tilde{\mathbf{x}} = \mathbf{R}\mathbf{x} \tag{1}$$

where $\tilde{\mathbf{x}} \in R^m$ is the vector in the random feature space, and $\mathbf{R} \in R^{m \times d}$ is a random matrix following either standard normal distribution or the following Bernoulli distribution [11]

$$R_{i,j} = \begin{cases} 1/\sqrt{m} & \text{with probability } 0.5 \\ -1/\sqrt{m} & \text{with probability } 0.5 \end{cases} \tag{2}$$

2.2 Nonlinear random mapping in ELM

The output of a hidden node is given as

$$\tilde{x}_i = g(\mathbf{a}_i \cdot \mathbf{x} + b_i), i = 1,2,\cdots m \tag{3}$$

where g is the active function for the hidden nodes, $\mathbf{a}_i \in R^d$ is the weight vector between input nodes and the i-th hidden node which are selected as a random vector following standard normal distribution, and b_i is the bias for the i-th hidden node which is generated as a random number following the same distribution.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

3 Classifier

Consider a training dataset $\{(\mathbf{x}_i, l_i); i = 1,..., n\}$, $\mathbf{x}_i \in R^d, l_i \in \{1,2,\cdots N\}$, where $\mathbf{x}_i$ represents the *i*-th sample, a d-dimensional column vector containing gene expression values with *d* as the number of genes, and l_i is the label of the *i*-th sample with *N* as the number of categories.

3.1 Sparse representation

Each training sample, $\mathbf{x}_i$, and a test sample **y** are mapped to $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{y}}$ in the random feature space by either (1) or (2). Thus $\tilde{\mathbf{y}}$ can be expressed as a linear combination of the entire training samples in the random feature space $\{\tilde{\mathbf{x}}_i; i = 1,..., n\}$

$$\tilde{\mathbf{y}} = c_1\tilde{\mathbf{x}}_1 + c_2\tilde{\mathbf{x}}_2 + \cdots + c_n\tilde{\mathbf{x}}_n. \quad (3)$$

Define a matrix by putting $\tilde{\mathbf{x}}_i$ as the *i*-th column: $\tilde{\mathbf{A}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \cdots \tilde{\mathbf{x}}_n]$, and a column vector $\mathbf{c} = [c_1, c_2 \cdots, c_n]^T$. We can convert (1) into

$$\tilde{\mathbf{y}} = \tilde{\mathbf{A}}\mathbf{c}. \quad (4)$$

The sparse solution of **c** can be obtained by solving the following l_1-regularized least square problem [12]

$$\min_{\mathbf{c}} \left\|\tilde{\mathbf{A}}\mathbf{c} - \tilde{\mathbf{y}}\right\|_2 + \lambda\|\mathbf{c}\|_1 \quad (5)$$

where $\lambda > 0$ is the regularization parameter, and $\|\mathbf{c}\|_1 = \sum_i |c_i|$ is the l_1-norm of **c**. A truncated Newton interior-point method (TNIPM) proposed in [12] can be used to solve the above optimization problem.

Let $\hat{\mathbf{c}}$ denote the sparse representation obtained by l_1-regularized least square. Ideally, the nonzero entries in $\hat{\mathbf{c}}$ are associated with the columns in $\tilde{\mathbf{A}}$ corresponding to those training samples of the same category as testing sample $\tilde{\mathbf{y}}$. However, noises may cause the nonzero entries to be linked with multiple categories. Simple heuristics, such as assigning $\tilde{\mathbf{y}}$ to the category with the largest entry in $\hat{\mathbf{c}}$, is not dependable. Instead, we define *N* discriminate functions

$$g_k(\tilde{\mathbf{y}}) = \left\|\tilde{\mathbf{y}} - \tilde{\mathbf{A}}\hat{\mathbf{c}}_k\right\|_2, k = 1,2,\cdots N \quad (6)$$

where $\hat{\mathbf{c}}_k$ is obtained by keeping only those entries in $\hat{\mathbf{c}}$ associated with category *k*, and assigning zeros to other entries. Thus g_k represents the approximation error when $\tilde{\mathbf{y}}$ is assigned to category *k*, and we can assign $\tilde{\mathbf{y}}$ to the category with the smallest approximation error. Please refer to [10] for more details.

3.2 ELM

Each training sample is mapped to $\tilde{\mathbf{x}}_j$ in the random feature space by (2). Let $\boldsymbol{\beta}_i \in R^N, i = 1,2,...m,$ denote the output weight vector between the *i*-th hidden node and *N* output nodes, and $\mathbf{t}_j \in R^N, j = 1,2,...n,$ as the *j*-th target vector whose only non-zero element is 1 at the l_j-th position. Then all the weights between the hidden nodes and the output nodes can be obtained by solving

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T} \quad (7)$$

where

$$\mathbf{H} = \begin{pmatrix} \tilde{\mathbf{x}}_1^T \\ \bullet \\ \bullet \\ \bullet \\ \tilde{\mathbf{x}}_n^T \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1^T \\ \bullet \\ \bullet \\ \bullet \\ \beta_m^T \end{pmatrix},$$

and

$$\mathbf{t} = \begin{pmatrix} \mathbf{t}_1^T \\ \bullet \\ \bullet \\ \bullet \\ \mathbf{t}_n^T \end{pmatrix}.$$

The above linear system can be efficiently solved by using Moore-Penrose generalized inverse. Please refer to [7] for details.

4 Numerical experiment

Two datasets used in the numerical experiment are described below.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

- Brain_Tumor1 [13]: The dataset comes from a study of 5 human brain tumor types: medulloblastoma, malignant glioma, AT/RT, normal cerebellum, and PNET, including 90 samples. Each sample has 5920 genes.
- Brain_Tumor2 [14]: There are 4 types of malignant glioma in this dataset: classic glioblastomas, classic anaplastic oligoden-drogliomas, non-classic glioblastomas, and non-classic anaplastic oligodendrogliomas. The dataset has 50 samples, and the number of genes is 10367.

Leave-one-out (LOO) cross validation is used to assess the performances among ELM, Sparse Representation with Gaussian random projection (SR+GRP), Sparse Representation with Bernoulli random projection (SR+BRP), Sparse Representation with Gaussian nonlinear random projection (SR+GNRP), Sparse Representation with Bernoulli random projection (SR+BNRP).

Table 1 shows the number of accurate predictions out of 90 samples in LOO for Brain_Tumor1 when the dimensionality of the random feature space ranges from 100 to 1000. And Table 2 is for Brain_Tumor1 when the total sample is 50. In both cases, the approach based on SR is slightly better than ELM, especially when the dimensionality is small. It seems that the performances are very similar between random projection and nonlinear random projection. In addition, there is no significant difference between two distributions

5 Conclusion

In this study, gene expression profile classification is done via sparse representation in the random feature space, which is obtained by either random projection or nonlinear random mapping used in ELM. The numerical experiment shows that sparse representation has slightly better performance than ELM.

6 References

[1] Y. Saeys, I. Inza, and P. Larraañga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, vol. 19, 2007, pp. 2507-2517.

[2] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarray data", *Computational Statistics & Data Analysis*, vol. 48**:**869-885, 2005.

[3] S. Kaski, "Dimensionality reduction by random mapping," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 1, 1998, pp. 413–418.

[4] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 245–250.

[5] A. Yang, J. Wright, Y. Ma, and S. Sastry, "Feature selection in face recognition: A sparse representation perspective." (Preprint, 2007)

[6] X. Hang, "Gene expression profile classification using random projection and sparse representation," in *Machine Learning and Applications (ICMLA)*, 2013. pp. 411-414.

[7] .G-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme Learning Machine: Theory and Applications," *Neurocomputing*, vol. 70, 2006, pp. 489–501.

[8] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multi-Category Classification Using Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, 2007, pp. 485-495.

[9] J. Wright, A. Yang, A. Ganesh, S. Shastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31(2), 2009, pp. 210-217.

[10] X. Hang and F.-X. Wu, "Sparse representation for classification of tumors using gene expression data," J. Biomed. Biotechnol., vol. 2009, pp. 1–6, 2009, DOI: 10.1155/2009/403689.

[11] R. Baraniuk et al., "A Simple Proof of the Restricted Isometry Property for Random Matrices," *Constructive Approximation*, vol. 28, no. 3, 2008, pp. 253–263.

[12] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares", *IEEE J. Select. Topics in Signal process.*, Vol. 1, No. 4, 2007, pp. 606-617.

[13] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression", *Nature*, Vol. 415, No. 6870, 2002, pp. 436–442.

[14] C.L. Nutt, D.R. Mani, R.A. Betensky, P. Tamayo, J.G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M.E. McLaughlin, T.T. Batchelor, P.M. Black, A. von

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Deimling, S.L. Pomeroy, T.R. Golub, and D.N. Louis, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification," *Cancer Res.*, vol. 63, 2003, pp. 1602–1607.

Table 1. LOO results for Brain_Tumor1

	100	200	300	400	500	600	700	800	900	1000
SR+BRP	80	81	80	81	81	80	81	81	82	82
SR+GRP	77	80	81	82	83	79	82	81	82	82
SR+BNRP	79	79	79	81	82	81	81	80	82	81
SR+GNRP	76	78	77	81	80	81	82	82	80	80
ELM	28	60	74	71	79	77	79	79	79	79

Table 2. LOO results for Brain_Tumor2

	100	200	300	400	500	600	700	800	900	1000
SR+BRP	34	34	34	39	38	38	39	38	38	36
SR+GRP	37	37	34	37	41	39	37	41	38	35
SR+BNRP	34	37	38	39	39	40	39	40	40	39
SR+GNRP	28	32	30	36	37	36	37	39	36	38
ELM	22	25	32	31	28	33	32	35	37	33

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Grid Computing-based Sequence Alignment

M. Shalaby[1] **and Y. Kamal**[2]
[1] Department of Information system, Modern university for information and technology, El-Hadaba El Wosta, Zone 5, Cairo, Egypt
[2] Department of Communication and Electronics, Giza Higher Institute of Engineering & Technology Giza, Egypt

Abstract - *Sequence alignment plays an important role in bioinformatics. There are two types of sequence alignment, global and local alignment. Needleman-Wunsch algorithm is the most famous example of global alignment. It is based on sequential computing so it has a problem of being slow. In this paper we present a new global alignment algorithm that can be implemented using parallel computing (such as grid computing) rather than sequential computing. The grid computing-based sequence alignment has the advantage of being as fast as heuristic algorithms and as sensitive as dynamic programming algorithms. We also present a comparative study between our approach and Needleman-Wunsch algorithm according to time and spaces complexity*.

Keywords: Sequence Alignment, Dynamic Programming, Grid Computing

1 Introduction

Sequence alignment is used to detect the functional, structural, or evolutionary similarities between two DNA, RNA, or protein sequences [1]. Detection of similarities in viruses and bacterium can help scientists find treatments for diseases.

There are several methods that have been implemented to solve the sequence alignment problem. The sequence alignment methods are either global or local. Global alignment conducts an end to end alignment between the query sequence and the subject sequence and is used to compare homologous genes, while local alignment aligns a substring (or the whole string) of the query sequence to a substring (or the whole string) of the subject sequence and is used to discover homologous portions in non-homologous genes.

Needleman-Wunsch algorithm [2] is the most famous example of global alignment. This algorithm is based on dynamic programming, it gives accurate results, however it is time consuming. Smith-Waterman algorithm [3] is another dynamic programming algorithm which is very similar to Needleman-Wunsch algorithm, however, it is used for local alignment. Other examples of local alignment algorithms are FASTA [4] and BLAST [5] which are heuristic-based pairwise local aligner. Heuristic methods are much faster than dynamic programming methods, however they are less sensitive.

In this paper we present a pairwise aligner algorithm that aims at obtaining as accurate results as Needleman-Wunsch and Smith-Waterman algorithms and is as fast as FASTA and BLAST. Unlike Needleman-Wunsch algorithm which is sequential in nature, we present an algorithm that can be implemented using grid computing and hence it overcomes the time consumption problem of dynamic programming methods.

In section 2, we briefly describe the Needleman-Wunsch algorithm. In section 3, we introduce the concept of grid computing. In section 4, we present the proposed grid computing-based sequence alignment algorithm. In section 5, we discuss the results of applying our approach and present a comparative study between this approach and the traditional Needleman-Wunsch algorithm. Finally in section 6, we conclude this work.

2 Global alignment using Needleman-Wunsch algorithm

Needleman-Wunsch algorithm (also called optimal matching algorithm) is a dynamic programming algorithm that is used to align globally two Protein sequences. Let $SEQ1$ be a sequence of residues (symbols) with length M and $SEQ2$ be a sequence of residues with length N, then the algorithm constructs a two dimensional matrix Q of dimension $(M+1) \times (N+1)$. We denote the entry of row m and column n as Q_{mn}, where Q_{mn} is calculated according to the scoring matrix [6-10] and the gap penalty scheme. Here, S denotes the scoring matrix, $S(i,j)$ denotes the similarity score between the residue of index i in $SEQ1$ and the residue of index j in $SEQ2$, and d denotes the gap penalty. The elements Q_{mn} are calculated as follows

$$Q_{0n} = d \times n \quad (1)$$

$$Q_{m0} = d \times m \quad (2)$$

$$Q_{mn} = max \begin{cases} Q_{(m-1)(n-1)} + S(i,j) \\ Q_{(m-1)n} + d \\ Q_{m(n-1)} + d \end{cases} \quad (3)$$

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

To calculate the elements Q_{mn} where $0 < m \leq M, 0 < n \leq N$, the elements $Q_{(m-1)(n-1)}, Q_{(m-1)n}, Q_{m(n-1)}$ (source elements) must exist and hence the algorithm is executed sequentially. After the calculation of all elements Q_{mn}, the algorithm starts from the bottom right element (Q_{MN}) and traces back the elements Q_{mn} which leaded to this score. Let $SEQ1(m)$ be the residue of index m in $SEQ1$ and $SEQ2(n)$ be the residue of index n in $SEQ2$. If the source element equals to $Q_{(m-1)(n-1)}$ the residue $SEQ1(m)$ is aligned to $SEQ2(n)$, if the source element equals to $Q_{(m-1)n}$ the residue $SEQ1(m)$ is aligned with gap, if the source element equals to $Q_{m(n-1)}$ the residue $SEQ2(n)$ is aligned with gap. Fig. 1 shows the pseudo code of Needleman-Wunsch algorithm, the inputs of this algorithm are $SEQ1$ and $SEQ2$, and the outputs are the aligned sequences A_SEQ1 and A_SEQ2.

```
Needleman-Wunsch algorithm
Input: SEQ1, SEQ2
for i=0 to length(SEQ1)
  Qi0 ← d*i
for j=0 to length(SEQ2)
  Q0j ← d*j
for i=1 to length(SEQ1)
  for j=1 to length(SEQ2)
  {
    Match ← Q(i-1)(j-1) + S(i,j)
    Delete ← Q(i-1)j + d
    Insert ← Qi(j-1) + d
    Qij ← max(Match, Insert, Delete)
  }
A_SEQ1 ← ""
A_SEQ2 ← ""
i ← length(SEQ1)
j ← length(SEQ2)
while (i > 0 or j > 0)
{
   If(i>0 and j>0 and Qij == Q(i-1)(j-1) +
S(i,j))
  {
    A_SEQ1 ← SEQ1(i) + A_SEQ1
    A_SEQ2 ← SEQ2(j) + A_SEQ2
    i ← i - 1
    j ← j - 1
  }
  elseif(i>0 and Qij==Q(i-1)j+d)
  {
    A_SEQ1 ← SEQ1(i) + A_SEQ1
    A_SEQ2 ← "-" + A_SEQ2
    i ← i - 1
  }
  else (j>0 and Qij==Qi(j-1) + d)
  {
    A_SEQ1 ← "-" + A_SEQ1
    A_SEQ2 ← SEQ2(j) + A_SEQ2
    j ← j - 1
  }
}
Output: A_SEQ1, A_SEQ2
```

Figure 1: Needleman-Wunsch algorithm.

3 Grid computing

Grid computing is software and hardware structure that provides high computing capabilities by sharing distributed resources like computers, storage space, and software applications [11]. The resources of grid computing structure must be scheduled [12]. The scheduling process has three stages: resources discovery, resource selection, and job submission [13].

Scheduling is either static or dynamic. In the static mode, the information of all resources and the tasks of the application are available when the application is scheduled [14, 15]. Unlike static mode, the dynamic scheduler allocates tasks to resources while the application executes. This case is useful when the number of iterations in a loop cannot be determined and jobs arrive in real time [16-20].

The overhead cost is the time required to schedule tasks and communicate the results of finished distributed tasks. The overhead cost should be minimized so that it does not negate the benefits of using grid computing [21].

In section 5 we show the results of applying grid computing to the concurrent tasks of our proposed approach.

4 Proposed approach

This approach aligns the two sequences $SEQ1$, and $SEQ2$. It assumes that length of $SEQ1$ is greater than or equal length of $SEQ2$. The algorithm searches for the maximum length consecutive match in $SEQ1$ and $SEQ2$. Here, the consecutive match is defined as the consecutive sequence of residues in $SEQ1$ and $SEQ2$ such that the cumulative score increases as long as one residue of the first consecutive sequence is compared with its corresponding residue of the second consecutive sequence. After finding the maximum length consecutive match, $SEQ1$ and $SEQ2$ are split into three sub-sequences each, the left side of consecutive match (L_SEQ1, L_SEQ2), the maximum consecutive match (M_SEQ1, M_SEQ2), and the right side of consecutive match (R_SEQ1, R_SEQ2). Again, search for the maximum consecutive match is applied recursively for the (L_SEQ1, L_SEQ2) and (R_SEQ1, R_SEQ2). For each recursive iteration, we add the gaps required to keep the maximum consecutive match in each part aligned, then we concatenate these three parts.

4.1 Proposed Approach Sub-modules

The proposed approach can be divided into three main sub-modules: Consecutive match sub-module (CM), split sequence sub-module (SS), and combine sub-sequences sub-module (CS).

The CM algorithm obtains all possible consecutive matches (we denote the consecutive matched residues of $SEQ1$ as $WORD1$ and the consecutive matched residues of $SEQ2$ as

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

$WORD2$) and saves the parameters of each match in a parameters table (P_TABLE). These parameters include $WORD1$ start index, $WORD2$ start index, the consecutive match length (W_LENGTH), and the consecutive match score ($SCORE$). CM algorithm starts with finding i, j such that the similarity score between $SEQ1(i)$ and $SEQ2(j)$ is positive. Therefore $WORD1$ is initialized to $SEQ1(i)$ with start index i, and $WORD2$ is initialized to $SEQ2(j)$ with start index j. After that, the next residues of index $(i+1)$ and $(j+1)$ are checked, they are added to $WORD1$ and $WORD2$ respectively if the similarity score between them are positive. The last step is repeated as long as the last residues similarity check is positive. The parameters table (P_TABLE) is initially empty and constructed only in the first run of CM algorithm, and hence $FIRST\_TIME$ parameter is initialized to TRUE. Eventually, P_TABLE is searched to obtain the maximum consecutive match (M_SEQ1, M_SEQ2) as well as other maximum consecutive match parameters (M_SEQ1 start index, M_SEQ2 start index, W_LENGTH). Figure 2 shows the pseudo code of CM algorithm.

```
CM algorithm
Input: SEQ1, SEQ2, P_TABLE, FIRST_TIME
If FIRST_TIME
{
 FIRST_TIME ← FALSE
 For k=0 to length(SEQ1)
 {
   SCORE ← 0
   For j=0 to length(SEQ2)
   {

 SEQ1_INDEX ← j + k (modulo length(SEQ1))
    If S(SEQ1_INDEX,j) > 0
    {
       If W_LENGTH=0
     Get WORD1_START_INDEX,
           WORD2_START_INDEX
        SCORE ← SCORE+
S(SEQ1_INDEX,j)
        Add SEQ1(SEQ1_INDEX) to
WORD1
        Add SEQ2(j) to WORD2
        WLENGTH ← WLENGTH+1
        Add the match parameters to
     P_TABLE
     }
    Else
    {
        SCORE ← 0
        WORD1 ← ""
        WORD2 ← ""
        W_LENGTH ← 0
     }
   }
 }
}
Get the maximum consecutive match
parameters from match parameters table
Output: P_TABLE , Maximum
consecutive match parameters
```

Figure 2: The pseudo code of CM algorithm.

The SS algorithm obtains (L_SEQ1, L_SEQ2) and (R_SEQ1, R_SEQ2) and hence it exploits the match parameters table (P_TABLE) to generate match parameter tables (LP_TABLE, RP_TABLE) for (L_SEQ1, L_SEQ2) and (RSEQ1, RSEQ2), respectively. LP_TABLE and RP_TABLE are generated such that LP_TABLE only considers the consecutive matches in L_SEQ1 and L_SEQ2, and RP_TABLE only considers the consecutive matches in R_SEQ1 and R_SEQ2. We note that, the start index and word length parameters of any consecutive match may be modified if it overlaps with the maximum consecutive match. Figure 3 shows the pseudo code of SS algorithm.

```
SS algorithm
Input: Maximum consecutive match
parameters, P_TABLE
Get L_SEQ1,L_SEQ2
Get R_SEQ1, R_SEQ2
For each record in P_TABLE
  If match parameters belong to
L_SEQ1, L_SEQ2
   {
     If (L_SEQ1 overlaps M_SEQ1) or
   (L_SEQ2 overlaps M_SEQ2)
       Modify parameters
     Add these parameters to
   LP_TABLE
 }
  else if match parameters belong
to LSEQ1, LSEQ2
   {
     If (RSEQ1 overlaps MSEQ1) or
   (RSEQ2 overlaps MSEQ2)
       Modify parameters
     Add these parameters to
   RP_TABLE
    }
Output: L_SEQ1, L_SEQ2, R_SEQ1,
R_SEQ2,LP_TABLE, RP_TABLE
```

Figure 3: The pseudo code of SS algorithm.

CS algorithm simply concatenates (L_SEQ1, M_SEQ1, R_SEQ1) and concatenates (L_SEQ2, M_SEQ2, R_SEQ2). For each iteration, gaps might be added to either L_SEQ1 or L_SEQ2 to make their length equal and keep maximum length consecutive match aligned. Similarly, gaps might be added to either R_SEQ1 or R_SEQ2 for the same reason.

4.2 Example

As an example, we use Blosum50 scoring matrix [22] and consider the two sequences,

$SEQ1$: *BHYYXALKRHHQWWHHQWW*

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

$SEQ2$: *HYYQCBBALKRHHQXHQWY*

In the first iteration, the algorithm fetches the maximum length consecutive match and obtains the following results,

M_SEQ1: *ALKRHHQ*, L_SEQ1: *BHYYX*, R_SEQ1: *WWHHQWW*

M_SEQ2: *ALKRHHQ*, L_SEQ2: *HYYQCBB*, R_SEQ2: *XHQWY*

In the second iteration, the algorithm fetches the maximum consecutive match in both left side and right side sub-sequences. For the left side sub-sequence, we obtain the results,

M_SEQ1: *HYY*, L_SEQ1: *B*, R_SEQ1: *X*

M_SEQ2: *HYY*, L_SEQ2: , R_SEQ2: *QCBB*

At this stage, the algorithm combines these results and adds the necessary gaps. Therefore, the aligned sub-sequences are

BHYY - - - X

- HYYQCBB

For the right side sub-sequences in the second iteration and taking into account that the similarity between the two residues ('W','Y') is positive, we obtain the results

M_SEQ1: *HQWW*, L_SEQ1: *WWH*, R_SEQ1:

M_SEQ2: *HQWY*, L_SEQ2: *X*, R_SEQ1:

Again, at this stage the results are combined and the aligned sub-sequences are

WWHHQWW

X - - HQWY

Back to the first iteration, the two aligned sub-sequences are concatenated so we obtain the following results

BHYY - - - XALKRHHQWWHHQWW

- HYYQCBBALKRHHQX - - -HQWY.

5 Results

In this section we apply the grid computing trend to the proposed approach presented in section 4. We use a network of three personal computers (with dual core 2.66 GHz processor, and 3 Gigabyte RAM each), one coordinator and two agents. The network topology is star where the coordinator controls the two agents. We use User Datagram Protocol as the data communication protocol (UDP). The tasks that can be executed concurrently are processed in parallel by the two agents. The first parallel task that is dispatched by coordinator is creating the lookup table of consecutive matches. The coordinator aggregates the results obtained by the two agents to find the maximum length consecutive match. The second parallel task is done after splitting the two sequences where searching for the maximum length consecutive match in left side sub-sequences is done in parallel (by the two agents) with searching for the maximum length consecutive match in right side sub-sequences.

On one hand, we use the environment described above to align many pairs of Protein sequences of different lengths. On the other hand, we align the same pairs of sequences using the typical Needleman-Wunsch algorithm which is sequential in nature using one personal computer (with dual core 2.66 GHz processor, and 3 Gigabyte RAM each).

To present a comparative study between our Grid Computing-based approach and Needleman-Wunsch approach, we analyze the time and space complexity of their main modules (which are constructing the lookup table in the proposed approach and filling the two dimensional matrix Q in Needleman-Wunsch approach). Let M be the length of the first sequence and N be the length of the second sequence, then the time complexity of the construction of the lookup table in the proposed approach is $O(MN)$ and so does the time complexity of filling the matrix Q in Needleman-Wunsch approach. Each record of the lookup table in the proposed approach occupy 16 bytes of memory so the space complexity of constructing the lookup table is $O(16L) \approx O(L)$, where L is the number of records in the lookup table. The space complexity of the matrix Q in Needleman-Wunsch approach is $O(MN)$.

Figure 4 shows the execution time of the two approaches for different values of MN. The difference between the execution time of the two approaches increases dramatically as MN increases.

Empirical results show that the proposed grid computing environment is not efficient for small sequence lengths (where $MN < 20000$) because the non-distributed devices outperform the distributed devices as a result of the overhead cost.

6 Conclusions

We presented a grid computing-based sequence alignment algorithm that can be used for the global alignment of a pair of Protein sequences. The proposed algorithm differs from the traditional Needleman-Wunsch algorithm in that some modules can be executed concurrently, and hence it is convenient to implement such algorithm using grid computing. The proposed approach has been implemented using a star network of three computers (one coordinator that controls two agents) and UDP data communication protocol.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Unlike the time complexity which is equal in the two approaches, the space complexity of the proposed approach is less than its counterpart for Needleman-Wunsch algorithm.

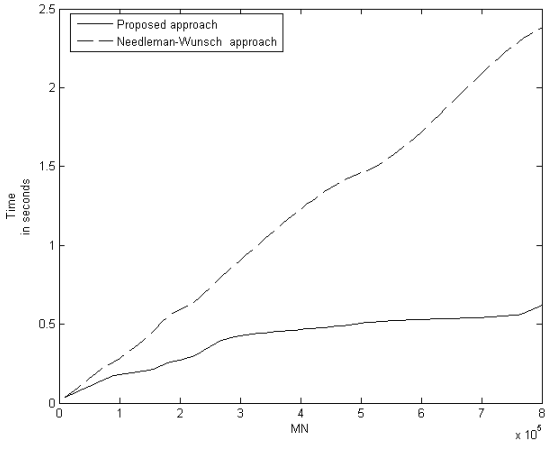


Figure 4: The execution time of the proposed approach and Needleman-Wunsch approach, where MN is the product of the length of the first sequence and the length of the second sequence.

Empirical results show that the proposed approach is exponentially faster than Needleman-Wunsch algorithm. They also show that applying the proposed approach on short-length sequences using one computer is more convenient than using a grid of computers because the overhead cost negates the benefits of using grid computing in short-length sequence alignment.

7 References

[1] D. Mount. "Bioinformatics: Sequence and Genome Analysis (2nd ed.)". Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-608-7, 2004.

[2] S. B. Needleman and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins"; J. Mol. Biol. 48, 443-453, 1970.

[3] T. F. Smith and M. Waterman. "Identification of common molecular subsequences"; J. Mol. Biol. 147, 195-197, 1981.

[4] D. J. Lipman and W. R. Pearson. "Rapid and Sensitive Protein Simlarity Searches"; Science 227, 1435-1441, 1985.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman. "A Basic Local Alignment Search Tool"; J. Mol. Biol., 215, 403-410, 1990.

[6] T. Müller, S. Rahmann, and M. Rehmsmeier; "Non-symmetric score matrices and the detection of homologous transmembrane proteins". Bioinformatics (Ox-ford, England). 17 Suppl 1: S182–9. PMID 11473008, 2001.

[7] D. W. Rice, D. Eisenberg. "A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence"; Journal of Molecular Biology, 267, 4, 1026–38. doi:10.1006/jmbi.1997.0924. PMID 9135128, 1997

[8] Gong, Sungsam, Blundell, and L. Tom. "Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures"; In Levitt, Michael. PLoS Computational Biology 4 (10): e1000179. doi:10.1371/journal.pcbi.1000179. PMC 2527532. PMID 18833291, 2008.

[9] N. C. Goonesekere, B. Lee. "Context-specific amino acid substitution matrices and their use in the detection of protein homologs"; Proteins, 71, 2, 910–9. doi:10.1002/prot.21775. PMID 18004781, 2008.

[10] Y. M. Huang, C. Bystroff. "Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions"; Bioinformatics, 22, 4, 413–22. doi:10.1093/bioinformatics/bti828. PMID 16352653, 2006.

[11] I. Foster and C. Kesselman. "The Grid: Blueprint for a Future Computing Infrastructure". Morgan Kaufmann Publishers, 1999.

[12] Y. Zhu. "A Survey on Grid Scheduling Systems", Department of Computer Science, Hong Kong University of science and Technology, 2003.

[13] J. Schopf. "Ten Actions When SuperScheduling", document of Scheduling Working Group, Global Grid Forum, http://www.ggf.org/documents/GFD.4.pdf, 2001.

[14] R. Braun, H. Siegel, N. Beck, L. Boloni, M. Maheswaran, A. Reuther, J. Robertson, M. Theys, B. Yao, D. Hensgen, and R. Freund. "A Comparison of Eleven Static Heuristics for Mapping a Class of Independent Tasks onto Heterogeneous Distributed Computing Systems"; J. of Parallel and Distributed Computing, 61, 6,810-837, 2001.

[15] H. Casanova, A. Legrand, D. Zagorodnov, and F. Berman. "Heuristics for Scheduling Parameter Sweep Applications in Grid Environments"; Proc. of the 9th hetero-geneous Computing Workshop (HCW'00), pp. 349-363, Cancun, Mexico, 2000.

[16] K. Kurowski, B. Ludwiczak, J. Nabrzyski, A. Oleksiak, and J. Pukacki. "Improving Grid Level Throughput Using Job Migration And Rescheduling"; Scientific Programming, 12, 4, 263-273, 2004.

[17] A. Takefusa, S. Matsuoka, H. Casanova, and F. Berman. "A Study of Deadline Scheduling for Client-Server Systems on the Computational Grid"; Proc. of the 10th IEEE International Symposium on High Performance Distributed

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Computing (HPDC-10'01), pp. 406-415, San Francisco, California USA, 2001.

[18] H. Chen and M. Maheswaran. "Distributed Dynamic Scheduling of Composite Tasks on Grid Computing Systems"; Proc. of the 16th International Parallel and Distributed Processing Symposium (IPDPS), pp. 88-97, Fort Lauderdale, Florida USA, 2002.

[19] N. Muthuvelu, J. Liu, N. L. Soe, S. Venugopal, A. Sulistio, and R. Buyya. "A Dynamic Job Grouping-Based Scheduling for Deploying Applications with Fine-Grained Tasks on Global Grids"; Proceedings of the 3rd Australasian Workshop on Grid Computing and e-Research, Newcastle, Australia, 2005.

[20] G. Malathi, S. Sarumathi. "Survey On Grid Scheduling"; Journal of Computer Applications, 3, 3, 22-29, 2010.

[21] F. Azzedin and M. Maheswaran. "Metagrid: A scalable framework for wide-area service deployment and management"; Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRIDí02), Berlin, Germany, 2002.

[22] S. Henikoff, and J. G. Henikoff. "Amino acid substitution matrices from protein blocks"; Proc. Natl. Academy Science 89, 915-919, 1992.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

SESSION

COMPUTATIONAL BIOLOGY AND MEDICAL APPLICATIONS + HEALTH INFORMATICS AND RELATED ISSUES

Chair(s)

TBA

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Intentionally-Linked Entities: a database system for health care informatics

V. Kantabutra[1],[2]
[1]Department of Informatics and Computer Science and
Department of Electrical Engineering
Idaho State University, Pocatello, ID, U.S.A.
[2]This work is supported by the U.S. National Science Foundation, award no. NSF-0941371
J.B. Owens, V. Kantabutra, D.P. Ames, R. Jones, investigators.

Abstract— *This paper introduces the Intentionally-Linked Entities or ILE database system and the possibility of using ILE to more efficiently and accurately manage data in health care information systems. ILE links together data entities in a more robust and efficient way than the Relational database system. ILE also keeps data in a more organized fashion than the Relational or the graph database system, and is capable of expressing more general relationships among data entities than the Object-Oriented database system. All these positive qualities of ILE present the possibility of improving the reliability and correctness of health care databases, and may lead to improved and more efficient patient care and health care database analysis.*

Keywords: ILE, EHR, epidemiology, health care, analytics, linked data

1. Introduction

This paper introduces a database system called Intentionally-Linked Entities (ILE), which links together data entities in a more robust and efficient way than the Relational database system. ILE also keeps data in a more organized fashion than the Relational system, and is more efficient at searches. These positive qualities of ILE present the possibility of improving the reliability and correctness of health care databases, and may therefore lead to improved and more efficient patient care and database analysis. By storing information more accurately and in a more organized fashion than the Relational database system, ILE may help health care providers avoid mistakes that can compromise patient health, patient confidentiality, or other aspects of good quality patient care.

A major motivation for developing the ILE database system is the importance of implementing data linking in a better, more robust way. The importance of data linking is eloquently described in [1, page 55]:

> Data linking is all about the way data in one part of a patient's record relates to data in another part of the record. When data linking fails, the data in an EHR for a patient is at war with itself. The simplest way to ensure that data is well-linked is to try and ensure that data is always linked correctly, ...

The authors also pointed out that when the information from a health care database is used for making dangerous decisions such as drug dosing and administration, then even a seemingly minuscule error rate such as 0.02% may mean several tragic errors because of the volume of cases handled in a large medical facility. Patients with the same or similar names are routinely confused in health care facilities. Such confusion may lead to serious health risks such as improperly handling drug allergies, inappropriate medicine or medical/surgical procedures, a compromise of patient confidentiality, or inefficient or inaccurate health care database analysis. This is why linking errors have to be eliminated to the greatest extent possible.

The Relational database system, first developed by Codd in the early 1970's, remains the most popular type of database system for health care informatics today [2]. Despite its name, a significant problem with the Relational database system is relationship linkage, which refers to the physical linking of data entities that are supposed to be related to each other via a relationship. More specifically, the Relational database system determines whether two data entries are the same by comparing data field values, values that are entered separately by the users, often without a stringent check to make sure that entries that supposed to match actually match each other.

The problem with that is that any misspellings, including the inclusion of blanks or invisible control characters, can cause an absence of linkage. Less likely but possible is the situation where two entries are inadvertently spelled the same and therefore linked when they shouldn't be.

Another problem with the Relational database system is that it is not a natural means of modeling complex data. Researchers and practitioners have noticed this fact since the 1980's, including in CAD [3] and in electronic medical records [4]. As stated correctly in [5, page 134], the Relational database model is appropriate for use when data logically match the idea of many identically structured records with relatively simple structure, or a collection of such structures.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

The Relational database system is also well known to suffer from data redundancy and data fragmentation. Data redundancy can be reduced by not permitting duplicate entries and by normalization. However, checking for duplicates reduces efficiency, and normalization can be quite difficult if it is to be done correctly. Additionally, many users and database practitioners prefer non-normalized databases because the tables of a non-normalized database can be much easier to use, having all or much of the desired information in one table.

Using an Object-Oriented database (henceforth "ODB," but more commonly "OODB," for object-oriented database) would solve the above-mentioned linking problem inherent in the Relational database system. Using a Network database, such as its new form, the graph database, would solve the linking problem as well. This is because both the OODB and Network database systems use object references or pointers for linkage, and linking to a non-existent object is automatically not permitted. From [6] it can be seen that Object databases have often become the database paradigm of choice for bioinformatics, or least for genomic computations.

However, the OODB system has two limitations that make it far less than optimum for health care informatics, or at least the part of it that deals with the treatment of diseases by health care providers working with patients and institutions like hospitals. One limitation is the lack of a query language in the original versions of the OODB. It is argued that the OODB is more often used in programs, and therefore what's important is an API, not a query language. However, by the turn of the millennium OODB vendors have realized that users normally require a query language, and thus the vendors responded by providing query languages. In fact, vendors went further and permitted Relational features in their systems, resulting in the Object-Relational database system. The other, more important limitation of the OODB is that it only supports relationships with two roles, that is, binary relationships, or relationships with arity = 2. Basically, the only means the OODB has to relate objects is to use object-valued attributes holding references to other objects [7], [8]. Thus, researchers came up with various ways to augment the OODB system to handle relationships of higher arity, that is relationships with 3 or more roles played by entities [8].

As mentioned earlier, the graph database would also solve the linkage problem. However, like the OODB, this style of database system also restricts the arity of the relationships to 2. Many-many relationships are not natively supported. Additionally, there are no physical objects representing the entity sets, and likewise no physical objects representing the relationship sets.

This paper introduces the author's Intentionally-Linked Entities (ILE) data model and the corresponding ILE database system as a better alternative for the EHR (Electronic Health Record) than the Relational, Objected-Oriented, or the Graph database system. The ILE database system was reported by this author in [9]. How ILE can be used in social network analysis and other applications in the Digital Humanities was explored in [10] and in [11].

ILE uses object-reference-based data structures to link entities that play the various roles in each relationship. The ILE database system natively supports relationships of all arity, meaning binary, ternary, quaternary, etc. In addition to that there is no restriction on how many other entities each entity could be related to in a relationship. So, for example, a many-many relationship, which is binary, is supported. Additionally, each role in a relationship can be played by an entire set of entities if that capability is needed. No other database system or model supports such general types of relationships, and that fact makes ILE more suited to the complex relationships found in health care situations. ILE was conceived as a direct implementation of the E/R (Entity/Relationship) database model. However, as it stands now ILE is more general in some important ways than E/R.

At the time that the paper [9] was written, ILE was not implemented. Since then a prototype has been implemented in Ruby, an object-oriented programming language that is particularly good for handling complex data structures and algorithms. However, lately the Treetop library for interpreting the commands and queries is broken. The Treetop package is not easy to work with anyway, and so the author decided to rewrite the command interpreter. The ILE code itself, implementing all the data structures, remains intact, and a new command interpreter is being written using a different package, Parslet, which, in the author's opinion, is much easier to work with than Treetop.

2. Introducing ILE for Health Care Applications

The best way to think of ILE is that it is a direct, straightforward implementation of the E/R model. There are differences and extensions, but we can discuss that as they come up.

As can be concluded from above, the Relational model favors relatively simple data models, even when these models are not necessarily as realistic as one would like. As an example, let's consider a Relational model from a database actually implemented at JMTZ Bee Healthcare, Inc., of a relationship between a Provider (doctor in this case) and a patient, shown in Fig. 1 [12].

Suppose we want to model the fact that a relationship between a provider and a patient comprises a set of visits. There are database models for the patient-provider relationship where the two people are related by a "visit" relationship. In such a representation, each visit is a separate relationship and there is nothing that really binds all the visits of one patient to the same provider together.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

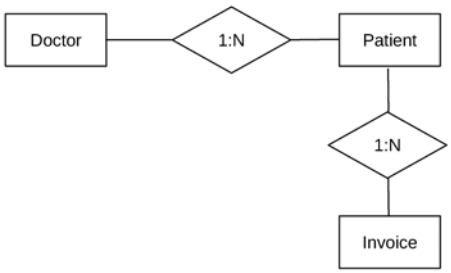


Fig. 1: A partial E/R diagram showing the provider-patient relationship in a Relational database actually used in the industry, by JMTZ Healthcare, Inc.

In ILE, we can easily model both individual visits and the longer-term relationship between a provider and a patient. The most natural way to do this is shown in Fig. 2. This can

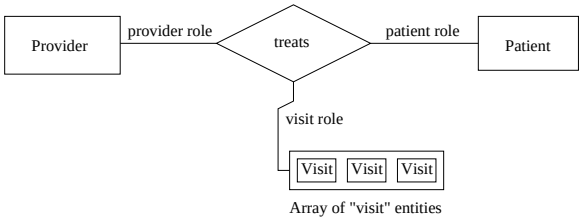


Fig. 2: A Possible Provider-patient relationship in ILE

be easily implemented in ILE as a ternary relationship, where the roles are patient, provider, and visits. The third role, visits, is actually a set or an array. In Relational databases, arrays are usually not permitted. For example, MySQL does not permit array datatypes. Workarounds are necessary, for example, see [13]. Oracle does have array data types [14], but the elements don't appear to be full-fledged entities that can be conveniently linked in relationships as individuals.

As another example to use in comparing the various kinds of databases, we can look at prescriptions. In JMTZ's Relational database, a prescription is an entity with two binary relationships, as shown in Fig. 3. One of these relationships is with an invoice, and the other with one or more medicines. An invoice may have 0 or 1 prescription. The Relational data model used by JMTZ allows for relationships with arbitrary arity. However, many designers of Relational databases favor binary relationships because in a binary relationship, entities can be linked together directly without an extra table representing the relationship, and also because joins can be expensive, especially joins of more than two tables. Even query optimization can take considerable time. If this situation were to be modeled using a graph database or a pure object-oriented database, then the type of relationships used would most likely be binary because only binary relationships are natively supported.

ILE, as opposed to these other database schemes, comfort-

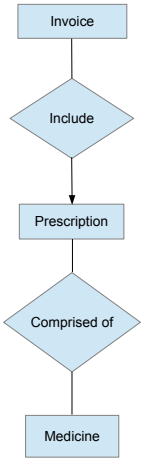


Fig. 3: Representing a prescription at JMTZ Healthcare, Inc.

ably and natively supports relationships of practically any finite arity. Fig. 4 shows how we can model a prescription

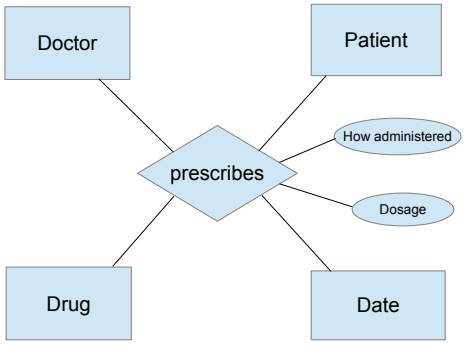


Fig. 4: The *Prescription* relationship in ILE

in ILE as a relationship of arity 4. In the next section the paper will cover how such relationships are implemented in ILE.

3. Inside the ILE Database System

In this section we look deeper into the ILE system.

Before we do that, let us note that the main idea is to be able to represent each relationship as an object that links the related entities together by means of links (object references or pointers) that go in both directions, namely, from the relationship object to each entity and from each entity to the relationship object. For example, the prescription relationship illustrated earlier in Fig. 4 would be linked as illustrated in Fig. 5.

While the idea behind this relationship linkage is simple enough, the actual implementation is not so simple. This is because each entity may be involved in more than one relationship, and maybe even more than one type of relationship. Therefore, instead of using one object reference to go from

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

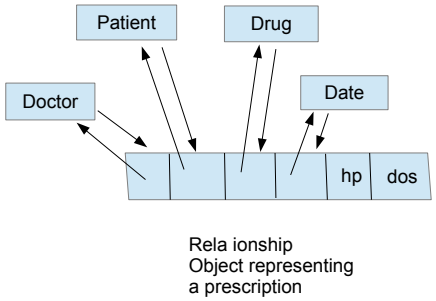


Fig. 5: Showing links used in the *Prescription* relationship as in Fig. 4 in ILE

an entity to a relationship object, we will use an aggregate data structure to hold such object references. This and other details will be the subject of the rest of this section.

The top level structure of ILE is actually a database set, implemented as a database set object. The database set contains its databases in a hash. This setup will permit cross-database searches within a database set, to be implemented in the future if there appears to be a need for such searches. Each database is also implemented as an object. The database set object has a hash whose data values are references to the database objects representing the databases of the database set. Each database object also has a reference back to the database set object.

Within an ILE database there are 4 different sets of components, namely,

1) entity sets
2) entities
3) relationship sets, also known as relsets
4) relationships

Each of these components is implemented as an object. The database object keeps track of the database's entity sets and relationship setss, and each entity set and each relset in turn has direct references back to the database to which they belong.

Note that the current version of ILE is value-oriented, not object-oriented, though arbitrary classes of entities are permitted simply because ILE is implemented in an object-oriented language. In the near future it may be a good idea to extend ILE to allow full object-oriented features. Being a value-oriented database system, ILE requires the entities of each entity set to have at least 1 field serving as the primary key. This concept should be familiar to readers who use the Relational database model.

The principal function of an entity set object is to keep all the entities belonging to the entity set together, as well as to facilitate searches for members of the entity set. The components of the entity set object are shown in Fig. 6, and will now be explained:

- Entity set name. This is any string, but of course it is meant to describe the entities in the entity set.
- Link to database. This is just an object reference back to the database to which this entity set belongs
- Key field names. The current version of ILE is value-oriented like the Relational database system, not object-oriented, even though it is implemented with an object-oriented language and permits use of non-elementary objects such at Date, or anything for that matter. Anyway, the fact that the ILE system is value-oriented means that all entities in an entity set have key fields. The entities may also have non-key fields just like the entities in a Relational database.
- Multilevel hash or index to the entities belonging to this entity set. Every entity set uses a multilevel hash as the primary way to access the entities. Another indexing scheme could be used, especially if the database is large enough that the entities mainly reside in secondary storage or somewhere on a computer network. The current index, the multilevel hash, is structured so that each level corresponds to a key field.
- A hash of relsets to whose members the members of this entity set are involved. This is a useful aid for performing queries on the database.

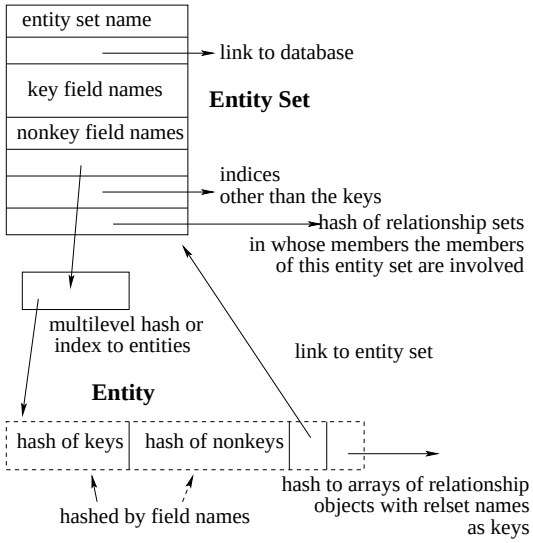


Fig. 6: An entity set object and an entity object in ILE, detail of contents

An entity object contains the following fields:

- Hash of key values.
- Hash of non-key values.
- Link back to the entity set object representing the entity set to which this entity belongs.
- A hash to arrays of relationship objects with relset names as hash keys. For each relset, there may be more than one relationship in which this entity participates, thus explaining the use of an array of relationship

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

objects for each relset.

We now explain relationship set (relset) objects and relationship objects as shown in Fig. 7. A relationship set object,

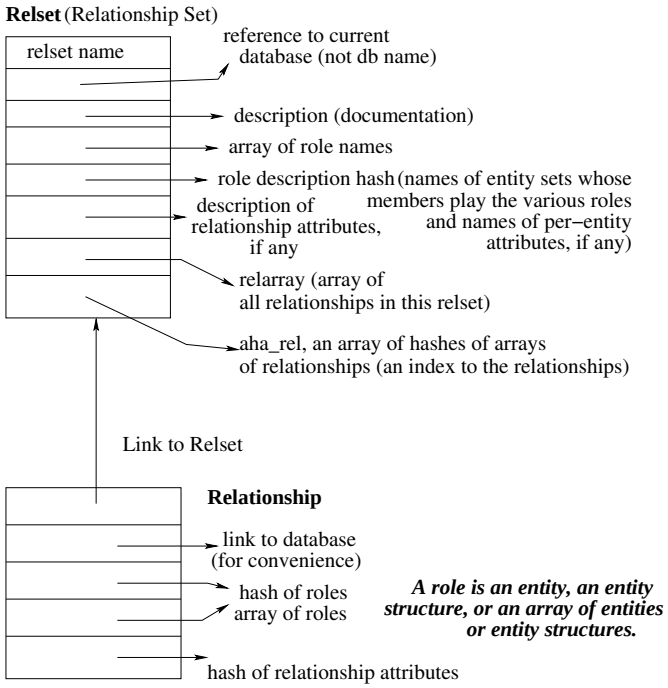


Fig. 7: A relset object and a relationship object in ILE, detail of contents

or relset object, contains the following fields:

- A field for the relset name, which is any string but should represent a descriptive name.
- An object reference back to the database object.
- A description for the relset.
- A set (Ruby Array) of role names.
- A description of relationship attributes, if any. This should be upgraded in later versions of ILE to be a more structured container for such attributes.
- Relarray, an array of all relationships in this relset.
- A data structure called aha_rel, which is an array of hashes of arrays of relationships in this relset. This is a complex and efficient index of all the relationships.

A relationship object has the following fields:

- A link to the database.
- A hash of roles, hashed by role names.
- An array of roles, so that one could access the roles in numerical order. Since both the hash and the array really contain object references, it is not much of a waste of space to have both the hash and the array.
- A hash of relationship attributes, if any.

A role belonging to a relationship can be of different types. See Fig. 8. The simplest type of role is just an entity object, representing the entity that plays that role. However,

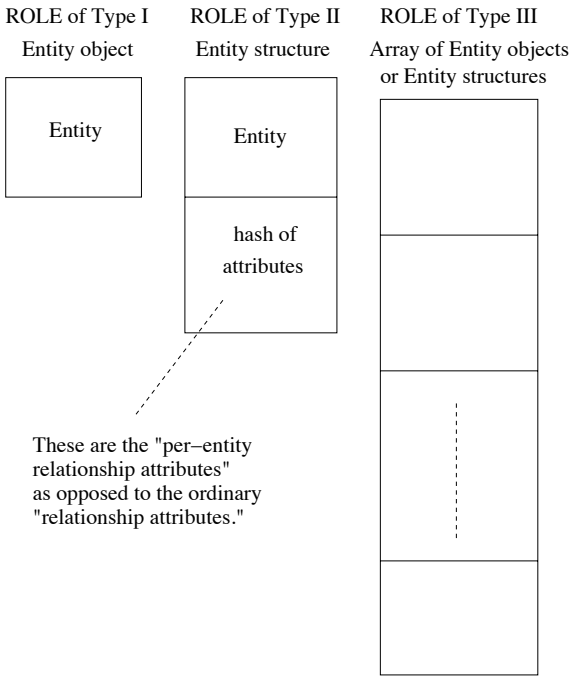


Fig. 8: The 3 types of role objects in a relationship

in case there is a per-entity relationship attribute, then we need such attributes to be attached to the entity, forming an *entity structure* object.

But what is a per-entity relationship attribute? The concept behind this kind of attribute was inspired by a historical database, where merchants and their clients came together to sign a contract in the presence of a notary back in the 1500's in Spain. There were some cases where the person wasn't present, but had someone sign the contract instead. The contract was modeled as a relationship in ILE, and each merchant and each client played roles in this relationship. A Boolean flag was used to indicate whether or not the person was present in person to sign the contract. This flag was modeled as a per-entity relationship attribute, so that each person had a different flag. The flags could have been modeled as a bit array serving as an ordinary relationship attribute. However, it appears that the modeling of these flags as per-entity relationship attributes was a "cleaner choice."

We have now discussed the first two possibilities for a role as shown in Fig. 8. The third possibility is that a role can be an array (representing a set) with each array element being either an entity object or an entity structure object.

As an example of a relationship linking roles, Fig. 9 shows how 3 roles can be linked to a relationship object in ILE.

4. Example of Importance of an EHR implemented in ILE

We close this paper by examining how the sophisticated data modeling afforded by ILE can make for an Electronic

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

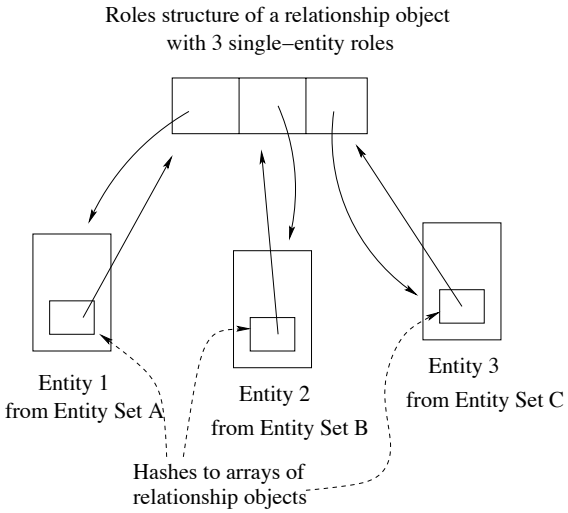


Fig. 9: A role structure

Health Record for a hospital admission that would make it easier for hospital epidemiologists or infection control practitioners to track down how an infectious disease or agent, such as staph, tuberculosis, pneumonia, etc., is being spread or may be spread in a hospital. Nosocomial infections are one of the hazards of hospitalization. Minimizing the chances of harmful nosocomial infections being spread is essential to patient health and health care administration and a robust ILE patient EHR can help with this.

One solution is to model hospital admissions as a relationship set. Each hospital admission then becomes a relationship in this set. This relationship is very complex because it relates entities from many entity sets. In particular, it relates a patient with a set of doctors, a set of nurses, CNAs, technicians of various sorts, equipment, doctor's orders, patient diagnosis or diagnoses, rooms occupied, what diet the patient is eating, etc. If the patient has an infectious disease, the room and patient could be flagged to indicate this fact with a per-entity relationship attribute. If a particular infectious disease agent is spreading, the epidemiologist or infection control practitioner can look at the database of ILE Electronic Health Records to see if there is a common factor or factors, such as a particular health care worker, etc., that is found in many of the patients contracting a nosocomial infection. Other database system types would have difficulty modeling such a complex set of entities and relationships about health care admissions and may be more error prone and present more difficulties with analyzing the health care data.

5. Conclusions

This paper introduced ILE, or Intentionally-Linked Entities, and explains why ILE may be a better database system than the existing Relational, Graph, and Object-Oriented database systems for managing health care information. ILE links together data entities in a more robust and efficient way than the Relational database system. With the embodiment of the entity sets, entities, relationship sets, and relationships as objects, ILE keeps data in a more organized fashion than the Relational or the graph database system. ILE is also capable of expressing more general relationships among data entities than the Object-Oriented database system because the pure Object-Oriented database system only natively supports binary relationships when more complex relationships may need to be represented, such as in the above hospital epidemiology example. Object-Relational database systems have the same pitfall in linking general relationships of arity higher than 2 as the Relational database system does. ILE is designed to make more complex relationships between data entities easy to represent and analyze. All these desirable properties of the ILE database system may lead to better management of health care information and therefore possibly improved patient care, more efficient health care operation, and better health care database analysis.

References

[1] F. Trotter and D. Uhlman, *Hacking Healthcare*. O'Reilly, 2013.
[2] K. A. Wager, F. W. Lee, and J. P. Glaser, *Managing health Care Information Systems*. Wiley, 2005.
[3] F. Stajano, "A gentle introduction to relational and object-oriented databases," 1998, ORL Technical Report TR-98-2.
[4] V. Speckauskiene and A. Lukosevicius, "The use of object-oriented technologies for medical data storing and retrieving," *European Journal for Biomedical Informatics*, 2008.
[5] I. J. Kalet, *Principles of Biomedical Informatics*, 2nd ed. Elsevier, 2014.
[6] A. B. Chaudhri and R. Zicari, *Succeeding with Object Databases*. Wiley, 2001.
[7] W. Kim, "Object-oriented database systems: Promises, reality, and future." in *Proc. of the 19th VLDB*, 1993, pp. 652–687.
[8] J. Schlegelmilch, "An advanced relationship mechanism for object-oriented database systems," 1996, department of Computer Science, University of Rostock, Germany.
[9] V. Kantabutra, "A new type of database system: Intentionally-Linked Entities-a detailed suggestion for a direct way to implement the Entity-Relationship data model." in *CSREA EEE*, 2007, pp. 258–263.
[10] V. Kantabutra, J. B. Owens, D. P. Ames, C. N. Burns, and B. Stephenson, "Using the newly-created ILE DBMS to better represent temporal and historical gis data," *Transactions in GIS*, vol. 14, pp. 39–58, 2010.
[11] V. Kantabutra, J. B. Owens, and A. Crespo-Solana, "Intentionally-linked entities: a better database system for representing dynamic social networks, narrative geographic information, and general abstractions of reality," in *Spatio-temporal Narratives: HGIS and the study of Trading Networks (1500 - 1800)*, A. Crespo Solana and D. Alonso García, Eds. Cambridge, U.K.: Cambridge Scholars Press, 2014.
[12] Y. Jin, "Healthcare management system (jmtz bee healthcare, inc.)," PDF document distilled from Powerpoint slides, 2000. [Online]. Available: http://www.angelfire.com/ny4/yjin/Healthcare/Healthcare-ppt.pdf
[13] "MySQL 5.1 reference manual." [Online]. Available: http://dev.mysql.com/doc/refman/5.1/en/data-type-overview.html
[14] "ARRAY (definition of, in the Oracle database system)." [Online]. Available: http://psoug.org/definition/ARRAY.htm

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Determining biological trends using machine learning techniques

by Regie Felix [1],Sky Adams [2], Nick Beck[3], Sophie D'Arcy [4]
[1]: Bioinformatics, CSU San Bernardino, San Bernardino, CA USA
[2]: Computer Science, UC Santa Barbara, Santa Barbara, CA USA

Abstract
Within biological data, there are significant trends that could provide pertinent information about a particular condition or disease. A well-known method in analyzing these trends is machine learning, a process where an algorithm learns data, studies its characteristics, and produces a detailed output. Machine learning includes techniques such as artificial neural networks (ANN), decision trees (DT), support vector machine (SVM), and Bayesian networks (BN). One application of this method is studying the behaviors within EEG results of alcoholic and non-alcoholic patients with the use of time series classification; with this study, one would be able to understand which area of the brain and to what extent does alcohol affect a patient's brain. Another application of this method is examining the epistatic interactions among single nucleotide polymorphisms (SNPs), gene expression, and phenotype using Bayesian Networks; with this study, one would be able to distinguish which genomic compartments causes a certain disease.

1. Introduction

When conducting a laboratory experiment, scientists often spend an extensive amount of time waiting for results to occur. This large amount of time could be further extended due to errors of the experimenter. Therefore, there are two significant factors when conducting an experiment: accuracy and efficiency. Ideally, there should be an even balance between these two experimental factors. For example, DNA (deoxyribonucleic acid) is the hereditary material that has the ability to store a massive amount of information and explain the complexities of life. For many organisms, this material is composed of millions of base pairs (Adenine, Thymine, Cytosine, and Guanine) that are ordered in various sequences. If one would want to process a set of genes that are possibly involved with the leukemia, one must have an appropriate algorithm that is able to accommodate for large sample sizes, yet accurately pinpoint a subset of genes that significantly influence leukemia. This is the type of bioinformatics research that will be discussed in this paper.

Bioinformatics is the area of science that incorporates biology, chemistry, and computer science. Its goal is to develop powerful yet efficient algorithms that solve problems in biology and chemistry. One of the commonly used tools for bioinformatics research is machine learning. The basis of this tool is that a particular algorithm learns the trends of the given data and uses that information to analyze and interpret the data. The machine learning techniques that will be discussed in this paper will include decision trees, Bayesian networks, and artificial neural networks.

2. Time Series Classification

A. Time Series Analysis

A sequence of data that is recorded in a given time period is called time series and the relationship between the data points is called time series analysis. Because most biological experiments are performed in a particular allotment of time, this type of analysis is efficient in analyzing the data within each time period. The initial experiment for time series analysis utilized the time series of the number of air passengers recorded each year from January 1949 to August 1960 (approximately 11 years).

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Figure 1 : The Box-Jenkins Approach to Modeling

Figure 1: This is the approach that is commonly used for producing a model that best represents the given data. In this case, we want to create a model for data in a form of a time series.

The method that has been commonly used for time series analysis is the Box-Jenkins Approach. It consists of three steps: model identification, coefficient estimation, diagnostic checking. The goal of this method is to determine the autocorrelation, which is the relationships between the sequential data points[1]. Before applying the Box-Jenkins approach, one should ensure that analysis of the given data is feasible. One must make sure that the data is set is a time series and its variability is consistent throughout the data. If the data varies drastically at a certain time period, it could possibly create noise in the autocorrelation model. After these prerequisites are met, the Box-Jenkins approach can be performed on the data. We first graph the ACF (autocorrelation function) and PACF (partial autocorrelation function); these two functions will determine the coefficients p and q respectively. Then, these coefficients will build a model that will analyze the autocorrelation of the time series. Once the model is produced, we have to check whether or not this model truly represents the data. If so, then the corresponding model is the one that "best fits" the data; if not, one would have to preprocess the data further to remove any more noise or change the p and/or q coefficients.

B. Time Series Classification

Another type of analysis with time series is called time series classification. Classification is the task of learning function *f* that maps each attribute to a predefined class. These different classes will separate and categorize each of the attributes. The classifier is a systematic tool that distinguishes between objects of different classes. There are many different classifiers that are commonly used for time series classification. The classifiers that this paper will focus on will be decision trees (DT) and artificial neural networks (ANN) to classify time series.

Figure 2 : Classification Process

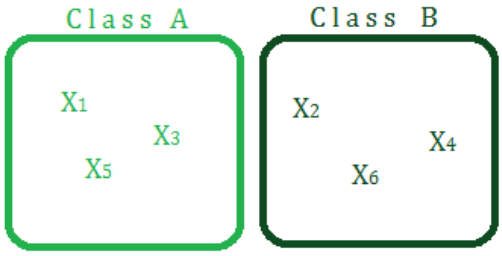


Figure 2: The image above depicts the process of classification; this process includes a vector of data which is separated into different classes based on its specific attributes.

Decision trees embody the structure of a tree; it consists of roots and branches which classifies the attributes in the data[6]. Decision trees start with a root node with no incoming edges and splits into multiple branches that are composed of internal and terminal nodes (Figure 3).

Figure 3 : Structure of a Decision Tree

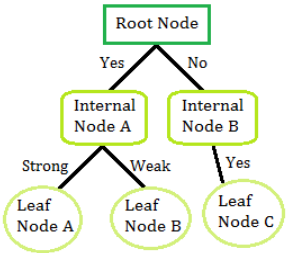


Figure 3: The image illustrates the structure of a decision tree, consisting of nodes and classifications.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Artificial Neural Networks (ANN) is analogous to the signaling neurons within the brain[6]. It is made out of an interconnected set of units which are all used to produce a single input (Figure 4). An example of an ANN would be the ALVINN driving system; this system is in charge of driving a car by gathering input from a camera placed in the car.

Figure 4: Structure of Artificial Neural Networks

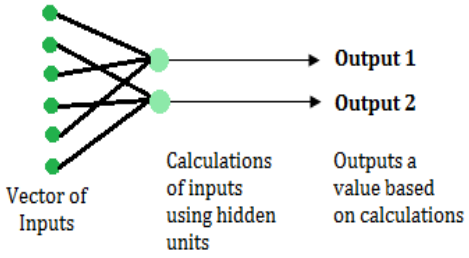
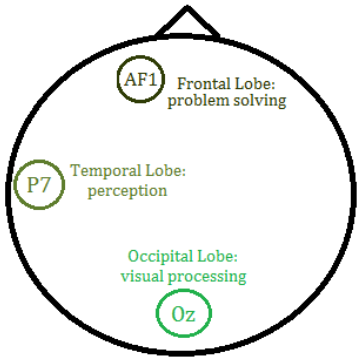


Figure 4: The image demonstrates the structure of an ANN, which consists of a vector of inputs that are analyzed and results in output 1 or output 2.

My time series classification experiment utilized a dataset from UC Irvine KDD[7] that studied the effect of alcoholism on a patient's EEG. A total of 20 patients (10 alcoholic and 10 non-alcoholic) were exposed to three different stimuli. While attached to a 64 electrode channel EEG, the patients were shown one image (s1), two images that match (s2), and three images that do not match (s3).

Figure 5: Arrangement of Isolated Electrodes



Our goal was to produce an efficient model that is able to accurately detect alcoholism based on the given EEG results. First, we determined the parts of the brain most affected by alcoholism, which are the frontal lobe, temporal lobe, and occipital lobe.

Then, in order to generate a model, we isolated the EEG values associated those areas and performed time series classification by utilizing decision trees and ANN. In this experiment, we preprocessed the data (organizing data into directories, applying DWT, creating an ARFF file) using Java and R[9] and classified the data with decision trees and ANN in the Weka software[5]. This experiment yielded intriguing results. When classifying the various electrode sensors correlated with alcoholism, we consistently had 50-60% classification accuracy. When comparing the efficiencies of the two different classifiers (ANN, decision trees), ANN had a higher accuracy with the AF1 sensor and decision trees had a higher accuracy with the OZ and P7 sensors.

3. Modeling Epistatic Interactions using Bayesian Networks

I. *Biological Background*

Genetics plays a significant role in the cause of a particular disease or phenotype. This relates back to the central dogma; double stranded DNA gets transcribed to single stranded RNA, which in turn is translated into a protein. This process describes how genetic information is used to express the particular characteristic of an organism.

Figure 6: The Central Dogma

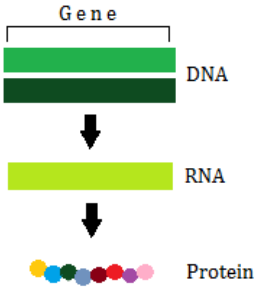


Figure 6: The central dogma illustrates how DNA is transcribed to RNA which is later translated to protein. The process that our research focuses on is the conversion from DNA to RNA.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

There are two factors that we focus on SNP (single nucleotide polymorphism), which is a change in one base pair within a DNA strand, and gene expression levels, which is the amount of RNA present from a particular gene.

II. Models

A. Two Layer SNP-Phenotype Model

This model illustrates the associations between SNP and the phenotype. This is a simple model where only one variable affects the organism's phenotype. In this experiment, we utilized a 20 SNP and 1000 SNP synthetic datasets, where the phenotype was binary (0: normal and 1: diseased).

Figure 7: Two Layer SNP-Phenotype Model

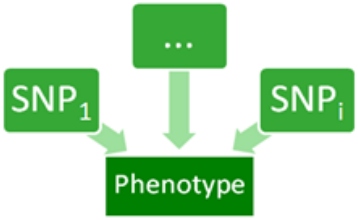


Figure 7: This model explains that changes in the DNA sequence are directly correlated to changes in the phenotype.

The purpose of this experiment was to determine which methods perform the best in performance and computation time. A synthetic dataset is initially used because we are able to evaluate the accuracy of the method since we know which SNPs are associated with the disease. There are five methods that we implemented in this experiment: Bayesian networks with multiple scoring systems, decision trees, Multifactor Dimensionality Reductions (MDR), Support Vector Machine (SVM), and Consistency subset analysis with resampling (listed below).

B. Two Layer SNP-Gene-Phenotype Model

This model illustrates the associations between SNP/gene expression and the phenotype. Since this model incorporates these two genomic factors, it demonstrates a more accurate representation of genomic activity. In this experiment, we used real datasets (listed below) with a large number of SNPs and gene expression values. Some of the datasets are preprocessed, but some of them are still in their original format.

Figure 8: One Layer SNP-Gene-Phenotype Model

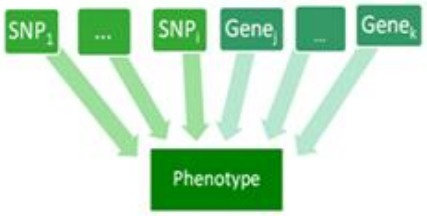


Figure 8: This model explains that changes in DNA and gene expression directly affect changes in the phenotype.

The purpose of this experiment was to utilize the methods that performed well in the previous experiment on real datasets that consist of SNP genotypes, gene expression levels and phenotype. In this case, we chose to focus on Bayesian Networks with multiple scoring systems and other software to analyze these datasets.

III. Methods

(a) *Bayesian Networks:*

This method uses an MCMC search method to search the space of all possible sets of SNPs associated with the phenotype[8]. It scores the sets of SNPs using one of several scoring criteria, all of which reward the set for how well it explains the phenotypic variation and penalizes the network for overfitting to the data ie. containing many SNPs[2,3].

(b) *Decision Trees*:

This method classifies the genomic factors and outputs a visualization of the classification as a tree. Each node represents a genomic factor and the branches represent the possible values

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

of the particular factor. The children or leaves of the node represent the phenotype that the method predicted.

(c) *Multifactor Dimensionality Reductions:* This method calculates the ratio of cases and controls and uses this measurement to create a threshold for the each of the combinations of genomic factors. If the ratio exceeds the threshold, it is labeled 'high-risk'; if the ratio is below the threshold, it is labeled 'low-risk'.

(d) *Other available software (Not included in research): iBMQ:* This analyzes SNP genotypes and gene expression data using an MCMC method. It calculates the probability that each SNP affects the expression level of each gene, accounting for interactions among SNPs. GENEVAR (by Wellcome Trust Sanger Institute) is an open source software that we can use to analyze and visualize SNP-gene associations in eQTL. ATHENA (by the Richie Lab at Penn State) is software that uses grammatical evolution neural networks to generate disease susceptibility models.

IV. Available Datasets:

- *Velez dataset:* This is a generated dataset was used for the two layer SNP-phenotype experiment, consisting of 20 SNPs, 200 samples.
- *1000 SNP epistatic dataset:* This is a generated dataset was used for the two layer SNP-phenotype experiment, consisting of 1000 SNPs, 200 samples, minor allele freq of 4. We used heritabilities of 0.01, 0.2 and 0.4.
- Maize dataset[4]: This is a real dataset that includes a total of 1 million SNPs, 29,000 gene expression levels, and numerous phenotypes (i.e. oil composition and amino acid content)
- *C. elegans dataset:* This is a real dataset that includes a total of ~1,400 SNPs, ~9,400 gene expression measurements, and 3 phenotypes (gonad reversal, tail rays, and distance between proliferation and differentiation).
- *Lymphoblastoid dataset:* This is a real dataset that includes a total amount of ~19 million SNPs and ~64,000 gene expression levels.
-*Leukemia dataset:* This is a real dataset that includes a total amount of 100,000 SNPs, 50,000 gene expression levels, and 1 phenotype (normal/diseased).

V. Results:

Two Layer SNP-Phenotype Model

In the results of this model, we noticed that if methods are more sensitive to low-effect SNPs demonstrated a high recall but low precision. Methods that avoid overfitting demonstrated a high precision but low recall. These inaccurate results may have occurred because of the low heritability that we used and because the model only incorporates SNP data.

(a) Bayesian Networks:

For the 20 SNP dataset, this method was able to detect the functional SNPs and achieve high recall (0.8) and low precision (0.15). The scoring system that had the highest recall values were BDeu alpha=30,54,162. For the 1000 SNP dataset, this method was also able to detect the functional SNPs, but resulted in about the same values for precision and recall (0.7). The scoring system that had the highest recall values for this experiment were BIC and BDeu alpha=1,5. This is surprising, because higher alpha values led to higher recall with the 20 SNP data set. Since this method produced acceptable results, we can use this method in the SNP-gene-phenotype experiment.

(b) Decision Trees: For the Velez 20 SNP dataset, decision trees had a moderate precision and recall at about 0.60. When looking at the tree visualization, the functional SNPs were embedded within the large tree. The disadvantage of this method is that it can only

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

handle one dataset at a time. As a result, we did not continue to use this method in the SNP-gene-phenotype experiment. However, the advantage of Weka[5] is that computational time is pretty fast and has a simple user interface.

(c) Multifactor Dimensionality Reduction (MDR): MDR was able to find the functional SNPs within the Velez 20 SNP dataset, but it also found some extra non-functional SNPs. Because of this, MDR had a very low precision and recall when averaging about 10 runs. When increasing the number of SNPs to 1000, MDR became too computationally complex to run. In order get one run a 1000 SNP dataset, the program took about 5 hours; also, the results of this run was inaccurate because it could not find the functional SNPs. As a result, we did not continue to use this method in the SNP-gene-phenotype experiment.

4. Conclusion and Future Directions

Our research done at UC Santa Barbara has demonstrated various techniques and experiments of machine learning. I have analyzed the intricate structure and process of decision trees, artificial neural networks, Bayesian networks, and multifactor dimensionality reduction, while applying them to real and synthetic datasets. Most of the methods were able to identify the causal SNPs, but only a few methods (Bayesian Networks and Decision Trees) had high accuracy. The other methods were able to find the causal SNPs; however, they would find extra SNPs which were not related to the disease. Also, Bayesian networks and decision trees were the only methods that were able to handle a large dataset. The computational time for the other methods were far too long. This research also explains how to evaluate the accuracy of a particular technique using classification accuracy, precision and recall. With the experience of this bioinformatics research, I plan to continue studying various machine learning techniques and applying them to data sets involving microarray data.

Figure 9: Two Layer SNP-Gene-Phenotype Model

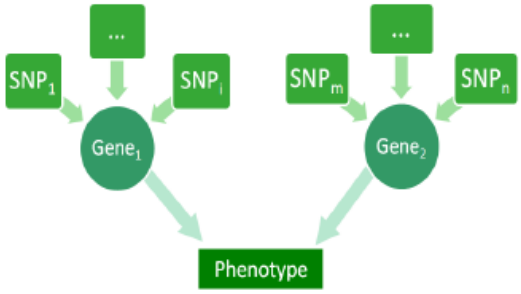


Figure 9: This model illustrates how a change in one nucleotide will affect gene expression and gene expression will overall affect the phenotype of the individual.

Acknowledgements

This research was conducted at UC Santa Barbara, under the supervision of PI Dr. Ambuj Singh. This work was supported by the National Science Foundation through grant IIS-1219254. I was involved in two summer internships at the UCSB campus. The graduate students that mentored my research were Nazli Dereli, Petko Bogdanov, and Nick Beck. I also collaborated with the Biology department faculty: Pradeep Joshi and Davon Callander.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

References

[1] Bisgaard, Søren, and Murat Kulahci. *Time Series Analysis and Forecasting by Example*. Hoboken, NJ: Wiley, 2011. *Wiley Online Library*. Web. July-Aug. 2012.

[2] Chang, Hsun-Hsien, and Michael McGeachie. "Europe PubMed Central." *Phenotype Prediction by Integrative Network Analysis of SNP and Gene Expression Microarrays*. Conf Proc IEEE Eng Med Biol Soc, 2011. Web. July-Aug. 2012.

[3] Han, Bing, Xue-wen Chen, Zohreh Talebizadeh, and Hua Xu. "Genetic Studies of Complex Human Diseases: Characterizing SNP-disease Associations Using Bayesian Networks." *ICIBM*. BMC Systems Biology, 17 Dec. 2012. Web. July-Aug. 2012. <http://www.biomedcentral.com/1752-0509/6/S3/S14>.

[4] Li, Hui, Zhiyu Peng, et al. "Genome-wide Association Study Dissects the Genetic Architecture of Oil Biosynthesis in Maize Kernels." *Nature.com*. Nature Publishing Group, 16 Dec. 2012. Web. July-Aug. 2012.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[6] Mitchell, Tom M. *Machine Learning*. 1st ed. New York: McGraw-Hill, 1997. Print.

[7] Stephen D. Bay and Dennis F. Kibler and Michael J. Pazzani and Padhraic Smyth. The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation. SIGKDD Explorations, 2. 2000.

[8] Pe'er, D. "Bayesian Network Analysis of Signaling Networks: A Primer." *Science's STKE*2005.281 (2005): Pl4. Print.

[9] Zhao, Yanchang. "RDataMining.com: R and Data Mining." *RDataMining.com: R and Data Mining*. N.p., 16 Mar. 2011. Web. July-Aug. 2012.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

A Histogramming Neural-Net Image Classifier of Some Triatomine Vectors of Chagas Disease

Jack K. Horner
PO Box 266
Los Alamos, New Mexico 87544 USA
jhorner@cybermesa.com

BIOCOMP 2014

Abstract

Chagas disease (CD) disease is a life-threatening tropical parasitic disease caused by the flagellate protozoan Trypanosoma cruzi. T. cruzi is typically transmitted to humans and other mammals by the bite of "kissing bugs" of the subfamily Triatominae (family Reduviidae), primarily by species belonging to the Triatoma, Rhodnius, and Panstrongylus genera. Rapidly identifying CD insect vectors in the field is crucial to effective control of the disease. Here I describe a histogramming neural-net classifier that supports rapid identification of adults of nine CD vector species.

Keywords: Chagas Disease, automated classification, neural net, image histogram

1.0 Introduction

Chagas disease (CD) disease is a life-threatening tropical parasitic disease caused by the flagellate protozoan *Trypanosoma cruzi*. *T. cruzi* is typically transmitted to humans and other mammals by the bite of "kissing bugs" of the subfamily *Triatominae* (family *Reduviidae*), primarily by species belonging to the *Triatoma*, *Rhodnius*, and *Panstrongylus* genera ([5]).

As many as 11 million people in Mexico, Central America, and South America have CD. Most of those infected do not know that they are. Large-scale population movements from rural to urban areas of Latin America and to other regions of the world have increased the geographic distribution of CD; the disease has been reported in several European countries ([5]).

Rapid field identification of CD vector species is essential to effective vector control. A tool that that could capture an image of a specimen and return its taxonomic identification would help to minimize the labor requirements for rapid field identification of the vectors.

Such a tool can be thought of as having two main subfunctions -- an *image-processing* function, which performs various digitized-image-analysis tasks and emits image information derivable solely from color-value distributions on pixel arrays (e.g., image histograms, maximum and minimum pixel intensity values, image differences, etc.; [15]), and an *automated classification* function, which accepts the output of the image processing and infers the taxonomic classification of the specimen.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2.0 Method

Images of adult specimens of nine triatomine species, (S), nominally regarded as the most common CD vector species in Brazil ([2]), were selected for classification support:

(S)

- *Triatoma infestans*
- *Triatoma dimidiata*
- *Triatoma brasiliensis*
- *Triatoma maculata*
- *Triatoma sordida*
- *Rhodnius prolixus*
- *Rhodnius neglectus*
- *Rhodnius pallescens*
- *Panstrongylus megistus*

Images of specimens of species in (S) were randomly selected from [1] and [16]-[18]. From this set I selected all images I judged to have "similar" focus, contrast, and specimen "quality", in the process excluding images I judged to have one or more of the following "pathologies":

- specimens with missing or occluded legs
- specimens showing significant anatomical, damage, or color anomalies
- in the case of [16]-[17], images evidently containing colors other than white or gray in their background
- specimens that had poor contrast in any region

Using the *Mathematica* ([9]) script shown in Figure 2, images of the selected specimens were imported from [16]-[18], "standardized" to remove various artifacts, grayscaled, Gaussian-filtered, and histogrammed (see Figure 2 for an example of the training set generator). 1000 of these synthetic histograms were generated per species. A linear perceptron with 50 input nodes, 20 hidden nodes, and 9 output nodes ([20], implemented in [19]) was developed ([21]) and trained on these histograms.

Ten images per species for each of the species in (S), disjoint from the set of training images, were obtained from [16]-[18] were similarly processed (but without Guaussian filtering) to create a set of "true positive" test histograms. Ten images per species for nine triatomine species not in (S), were obtained from [16]-[18] similarly processed (without Guaussian filtering) to create a set of "true negative" test histograms.

In total, 90 "true positive" (10 for each species in (S)) and 90 "true negative" test images (images of triatomines whose species were not in (S)) were randomly selected for testing.

In addition, 9000 "synthetic" test images, 1000 for each of (S), were created by Gaussian filtering (filtering radius normally distributed with mean = 30 pixels and standard deviations = 0.3 pixels). The basis images for each species were obtained by randomly selecting, for each species in (S), one of the 90 "true-positive" images.

The training-set generator script shown in Figure 2 was executed 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* operating environment.

3.0 Results

Example triatomine images are shown in Figure 1.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

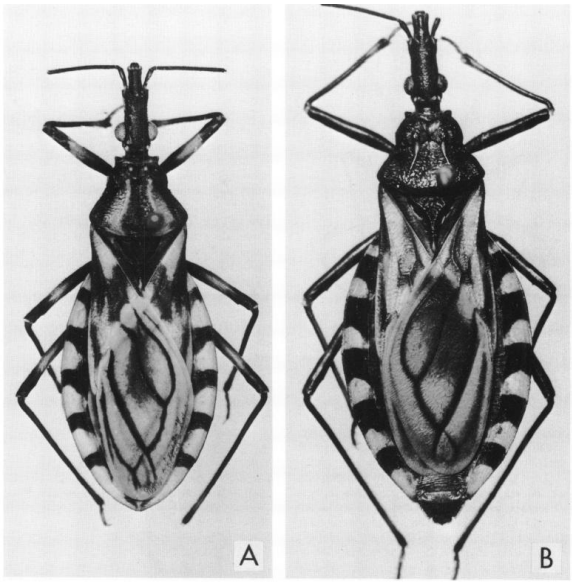


Figure 1. ***Triatoma brasiliensis*. A. Male Ceará, Brazil. B. Female, dark form, Bahia, Brazil. Adapted from Figure 43, [1].**

```
imax=1000;
numinputneurons=50;
imgwidth=300;
imgheight=600;
numreftypes=9;
binarizeThreshold=0.05
normalmean=20;
normalvar=0.2*normalmean;
 dirpath="C:\\Biodiversity_Institute\\Image_Generator\\Neural_net\\";
 normaldist=RandomVariate[NormalDistribution[normalmean,normalvar],imax];
 For [j=1,j≤numreftypes,j++,
  reffilename= "RefType"<> ToString[j]<>".JPG";
     orgimage=Import[dirpath<>reffilename, "JPG"];
  border=Binarize[orgimage,binarizeThreshold];
  mask=ColorNegate[border];
  maskedImg=ImageApply[1&,orgimage,Masking→mask];
     croppedimage=ImageResize[ImageCrop[maskedImg],{imgwidth,imgheight}];

     (*
imagelist=Table[Blur[croppedimage,RandomInteger[{blurradiusmin,blurradiusmax}]],{i, 1,
imax}]; *)
  imagelist=Table[Blur[croppedimage,Round[Abs[normaldist[[i]]]]],{i, 1, imax}];

graylist= Table[ImageAdjust[ColorConvert[imagelist[[i]],"Grayscale"]],{i,1,imax}];

simfilename="SimType"<> ToString[j] <>".txt";

outstream=OpenWrite[dirpath <>simfilename,BinaryFormat→True];

 For[k=1,k<=imax,k++,
```

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

```
            binning=ImageLevels[graylist[[k]],numinputneurons];
          Do[BinaryWrite[outstream,"
"];BinaryWrite[outstream,ToString[binning[[i]][[2]]],{i,1,numinputneurons}];
      BinaryWrite[outstream,"   "<>ToString[j]<>"\r\n"]];

     Close[outstream]];

oldlist=ReadList[dirpath<>"SimType1.txt", String];
 For[j=2,j≤numreftypes,j++,
  filename=dirpath<>"SimType"<>ToString[j]<>".txt";
  newlist=ReadList[filename,String];
  mylist=Join[oldlist,newlist];
  oldlist=mylist];
 trainsample=RandomSample[mylist,(imax*numreftypes)/2];
  trainfilename="Training.dat";
   outstream=OpenWrite[dirpath <>trainfilename,BinaryFormat→True];

    For[k=1,k≤ (imax*numreftypes)/2,k++,
        BinaryWrite[outstream,ToString[trainsample[[k]]]<>"    \r\n"]];


    Close[outstream];
```

Figure 2. A nominal *Mathematica* ([9]) script for generating the training set for the image neural net classifier.

The trained perceptron classified all 90 "true positive" test images correctly. It classified all 90 ("true negative") images of species not in (S) as "unknown".

The trained perceptron classified all 9000 synthetic test images correctly.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

4.0 Discussion and conclusions

The results of Section 3.0 motivate several observations:

1. An image-histogramming neural-net classifier can provide automated support for CD vector classification.

2. Image-histogramming renders the classification insensitive to in-focal-plane rotation and translation of the specimen of interest. However, it is subject to aliasing background and foreground elements.

3. Image-histogramming destroys all spatial information in an image. It yields only the distribution of a "headcount" of pixels by color-value (in this study, grayscaled).

4. Testing revealed that the performance of the classifier depends significantly on the "standardization" of images prior to classification.

5. The selection of training and test images based on a judgment of image "quality", as described in Section 2.0, would seem to involve expert judgment in some fundamental way. However, [1] and [16]-[18] strongly suggest that it is possible to define a photographic protocol that consistently produces images that allow the classilfer to perform as well as noted in Section 4.0. The explicit characterization of that protocol is an ongoing research project coordinated by the University of Kansas Biodiversity Institute.

6. The performance of the classifier on the 9000 synthetic test images strongly suggests that classifier is tolerant of plausible normal deviations from the training set.

5.0 Acknowledgements

R Gurgel ([16],[17]) and JMR Willoquet ([18]) provided the images used to train and test the classifier. This work benefited from discussions with Ed Komp ([10]), Town Peterson, and John Symons. For any infelicities that remain, I am solely responsible.

6.0 References

[1] Lent H and Wygodzinsky P. *Revision of the Triatominae (Hemiptera, Reduviidae), and Their Significance as Vectors of Chagas' Disease. Bulletin of the American Museum of Natural History* 163:3 (1979).

[2] Martinez D. Informal communication. 2013.

[5] US Centers for Disease Control. Parasites -- American Trypanosomiasis (also known as Chagas Disease): Detailed FAQs. http://www.cdc.gov/parasites/chagas/gen_info/detailed.html. 2012.

[6] US Centers for Disease Control. Parasites -- American Trypanosomiasis (also known as Chagas Disease): Antiparasitic Treatment. http://www.cdc.gov/parasites/chagas/health_professionals/tx.html. 2012.

[9] Wolfram Research. *Mathematica Home Edition v9.0.0*. http://www.wolfram.com/mathematica-home-edition/. 2013.

[10] Komp E. Informal communication. May 2013.

[15] Petrou M and Petrou C. *Image Processing: The Fundamentals*. Second Edition. Wiley. 2010.

[16] Gurgel R. *Rodrigo/ColRodrigo*, a triatomine image collection. Circa June 2013.

[17] Gurgel R. *RODRIGO2/triatominae UnB new photos*, a triatomine image collection. Circa June 2013.

[18] Willoquet JMR. *Will_Tdim _batch1/Photos_JMR*, a Triatoma dimdiata image collection.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Circa June 2013.

[19] University of Texas/Arlington. Image Processing and Neural Networks Laboratory. *NuClass V7.1.0*. http://www-ee.uta/eeweb/IP/. 1981-2004.

[20] Haykin S. *Neural Networks and Learning Machines*. 3rd Edition. Prentice Hall. 2008.

[21] Horner JK. *TriatomineNN*, a neural net classifier. 2013. Available from the author on request.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Model of Diagnosis Knowledge Base Ontology Framework based on Clinical Practice Guidelines

Hala Almutair[1] and Samir El-Masri[2]

[1]*Imam Mohammad bin Saud University*
[2]*College of Computer and Information Sciences King Saud University, Riyadh, Saudi Arabia*

Abstract. Most of caregivers started advising their physicians and medical staff to use clinical practice guidelines to help them in making the right diagnosis decision for the patient situation. However, the most available clinical practice guidelines are made for a certain disease or medical problem, and until now, there is no general framework for the clinical practice guidelines that can be used for any medical problem. This paper introduces a general knowledge base ontology framework for patient diagnosis based on clinical practice guidelines. This framework is a general base, which can be used with more specialization for quickly modelling a specific clinical practice guideline.

Keywords. Clinical Practice Guidelines; Ontology; Patient Diagnosis, Patient symptoms and signs, General knowledge base ontology, Diseases, Differential Diagnosis.

Introduction

Many medical errors occur due to the inaccurate diagnosis or treatment of the patient disease which could yields sometimes to the patient death. As a result, a new direction has been followed to benefit from the technology in the medicine field by using special medical information systems to support the physicians with their diagnosis and treatment decisions with the help of Clinical Practice Guidelines CPGs. Clinical practice guidelines are "systematically developed statements to assist practitioners and patient decisions about appropriate health care for specific circumstances" [1].

Many approaches have been developed for the computerization and execution of ontological clinical practice guidelines models. It has been clear that most of these researches and also the existing CPGs emphasize on the patient medical situation as a whole including diagnosis and treatment and apparently there is no such a framework for the patient diagnosis alone which is obviously the most important part of the whole medical situation since that if the physicians for example know exactly and for sure the patient disease they can easily decide what is the followed procedure for the treatment they need. Therefore, there is an urgent need to fill this huge gap and to build such CPG or more precisely a general framework that can fit at least 75% of all steps needed in diagnosis for most CPGs and this is what this paper introduces.

[2] Associate Professor ,College of Computer and Information Sciences, King Saud University
[1] Teacher Assistant, Imam Mohammad bin Saud University

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

1. The proposed general knowledge base ontology framework for patient diagnosis

This framework is a general base which can be used with more specialization for quickly modeling a specific clinical practice guideline and it consists of four steps. The first step of this methodology is to choose an appropriate clinical practice guidelines resource as the base of this research. National Guideline Clearinghouse NGC [2 has been chosen as a comprehensive, detailed information resource on clinical practice guidelines and to further their dissemination, implementation, and use. The second step and as the domain of the research, 30 different diseases has been chosen from different human organs and have been visualized in tables based on their symptoms, signs, and diagnosis procedures. The third step is capturing and modeling the common symptoms and signs among these 30 different diseases and with the help of the differential diagnosis that will go out at the end, the patient will be successfully diagnosed. The fourth and the last step in this methodology is the transformation of these models into a knowledge base ontology framework for patient diagnosis based on Clinical Practice Guidelines by using Protégé [3].

1.1 The selected 30 Diseases: symptoms, signs and Diagnosis

30 different diseases has been chosen for the study from NGC [2] (Table 1).

Table 1. The 30 selected diseases.

Disease no.	Disease Name	Disease no.	Disease Name
1	Celiac [4]	16	Hematological cancer
2	Gastroesophageal reflux [5]	17	Hypothyroidism
3	Gastroparesis [6]	18	Hyperthyroidism
4	Autoimmune Hepatitis AIH [7]	19	Tuberculosis
5	Acute Appendicitis [8]	20	Migraine
6	Dyspepsia [9]	21	peptic ulcer disease
7	Asthma	22	Lung cancer
8	Sore throat	23	Gastrointestinal food allergy
9	Pneumonia	24	upper respiratory food allergy
10	Acute rhinosinusitis	25	Lower respiratory food allergy
11	acute pancreatitis	26	ischemic stroke
12	chronic obstructive pulmonary	27	primary sclerosing cholangitis
13	type 2 diabetes mellitus	28	Hernia
14	Lower Urinary Tract Infection	29	lactose intolerance
15	Hepatitis B virus	30	Irritable bowel syndrome

The next step is to use NGC guidelines to summarize the symptoms, signs and diagnosis procedures of these 30 diseases. However, it is difficult here to explain all the information for this 30 diseases, instead the work has been shrink to the first 6 diseases from Table 1. Moreover, sample of the symptoms, signs and diagnosis procedures are presented (Table 2. and Table 3.).

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 2. Some symptoms and signs of the selected diseases.

Disease no.	Abdominal pain	Diarrhea	Weight Loss	Bloating	Heartburn	Nausea	Vomiting
1	Y		Y	Y			
2	Y			Y	Y	Y	Y
3	Y			Y		Y	Y
4	Y					Y	
5	Y					Y	Y
6	Y			Y	Y	Y	Y

Table 3. Some of the diagnosis procedures of the selected diseases.

Disease no.	erologic testing	IgA. TTG antibody testing	Endoscopy	Ultrasound	C urea breath test CUBT	Fecal antigen test
1	Y	Y	Y			
2			Y			
3						
4						
5				Y		
6					Y	Y

It is clearly from the above symptoms and signs table and diagnosis table that there are some common factors between two or more diseases. These factors are helpful in the differential diagnosis of the patient disease. Differential diagnosis means " the distinguishing between two or more diseases with similar symptoms by systematically comparing their signs and symptoms" [10].

1.2 Patient Symptoms and Signs Model

As a first step in any medical problem, the physician normally asks the patient about what he/she feels, and checks if there are any symptoms or signs in his/her patient and after that, the diagnosis process starts. In this research, the whole diagnosis process has been started from the initial 30 possible diseases and then with the available common symptoms and signs in every time, the diagnosis process will go out from a general picture to a more specific and precise picture of the possible patient disease. This process will be repeated at some levels until it reaches a reasonable set of possible patient diseases list. This process has been expressed clearly as a tree format. The rectangular nodes represent the possible diseases while the arrows represent the symptoms and signs of the patient and the rounded rectangular nodes represent the final and the exact diagnosis for the patient disease after dismissed all the excluded ones. (Figure 1.).

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

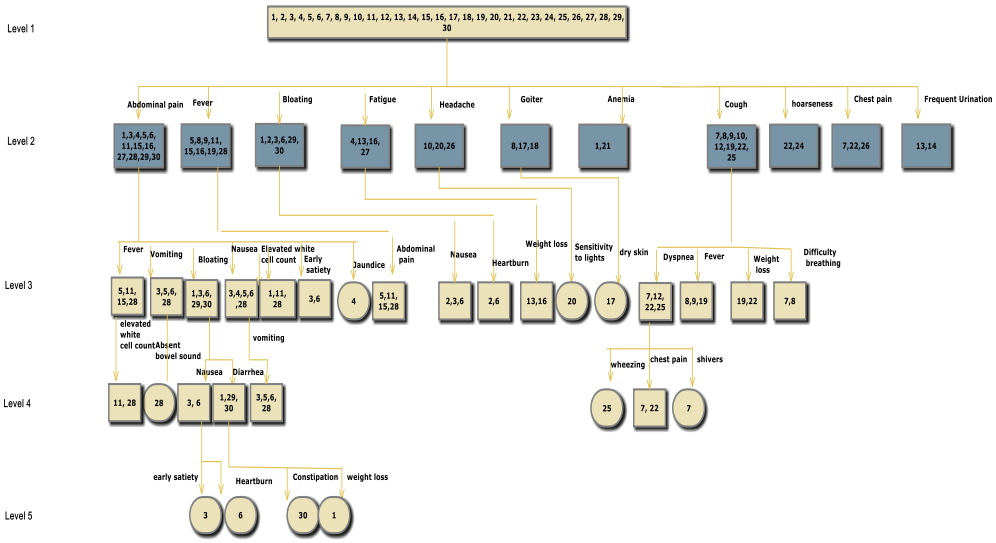


Figure 1. General patient symptoms and signs model

For instance, starting from the top with 30 possible diseases, if the patient complaint is about an abdominal pain, then the possible diseases from Table 1. that caused abdominal pain symptom (from Table 2.) are: 1, 3, 5, 6, 11, 15, 16, 27, 28, 29, and 30. Moreover, if the patient has fever, the above diseases list will be shortened to the diseases numbers :5, 11, 15, and 28 and so on.

1.3 Diseases Diagnostic Model

Next step is to list the different diagnosis tests and investigations to diagnose these 30 different diseases. In medicine, there are two types of investigations or diagnosis procedures; laboratory tests (Blood culture, Urine analysis and culture..) and Radiological tests (Ultrasound, MRI, ..etc). In this research, there are 45 different diagnosis procedures that can be used in the diagnosis process based on the physician needs and vision about the patient situation. The diseases Diagnostic Model is modeled and divided into three parts, one of them is presented in Figure 2.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

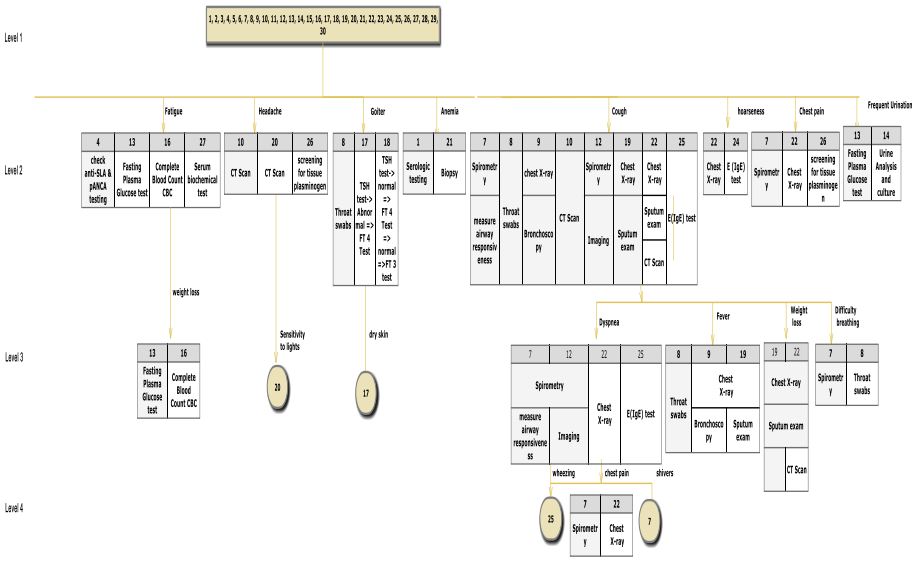


Figure 2. Diseases Diagnostic model

The physician starts thinking about the possible patient diseases based on their complaints and their symptoms and signs. At the end, he will have the exact diagnosis for the patient situation or at least he will reach to a differential diagnosis list and based on his vision and expectations the decision to have more investigations is decided. From Figure 3 for instance, if the patient complaints is about frequent urination, the differential diagnosis that comes to the physician mind is either the patient has Type 2 Diabetes Mellitus or has Lower Urinary Tract Infection and the exact diagnosis decision depends on the physician opinion to do Fasting Plasma Glucose test for Type 2 Diabetes Mellitus or Urine Analysis and culture for Unary Lower Urinary Tract Infection but not the both of them.

2. Discussion

It has been clear that the existing CPGs emphasize on one disease or one medical problem. Therefore, there is an urgent need to fill this huge gap and to build such CPG or more precisely a general framework that can fit at least 75% of all steps needed in diagnosis for most CPGs. This proposed knowledge base ontology framework for diagnosis based on CPGs will make building the clinical practice guidelines much more easier. The framework will be a general base for quickly modeling any specific clinical practice guideline.

3. Conclusion

This research proposed a new design for a knowledge base framework for a patient diagnosis based on CPGs. A full framework based on CPGs has been established to assist clinicians in the process of diagnosis. In order to build a specific diagnosis for a specific disease, the framework can save time and work. The next steps will be to complete this

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

research by building a general Diagnosis model based on the pervious selected 30 diseases side by side with patient symptoms and signs model.

References

[1] Clinical practice guidelines, The Open Clinical knowledge management for medical care, http://www.openclinical.org/guidelines.html .
[2] National Guideline Clearinghouse, http://www.guideline.gov/about/index.aspx
[3] what is protégé, http://protege.stanford.edu/overview
[4] ACG clinical guidelines: diagnosis and management of celiac disease: http://www.guideline.gov/content.aspx?id=45327
[5] Guidelines for the diagnosis and management of Gastroesophageal reflux disease, http://www.guideline.gov/content.aspx?id=43847
[6] Clinical guideline: management of Gastroparesis. http://www.guideline.gov/content.aspx?id=43612
[7] Diagnosis and management of autoimmune hepatitis. http://www.guideline.gov/content.aspx?id=23926
[8] Clinical policy: critical issues in the evaluation and management of emergency department patients with suspected appendicitis. http://www.guideline.gov/content.aspx?id=15598&search=Acute+appendicitis
[9] Dyspepsia. A national clinical guideline. http://www.guideline.gov/content.aspx?id=3723
[10] Differential Diagnosis, http://medical-dictionary.thefreedictionary.com/differential+diagnosis

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

An Image-Euclidean-Distance Classifier of Some Triatomine Vectors of Chagas Disease

Jack K. Horner
PO Box 266
Los Alamos, New Mexico 87544 USA
jhorner@cybermesa.com

BIOCOMP 2014

Abstract

Chagas disease (CD) disease is a life-threatening tropical parasitic disease caused by the flagellate protozoan Trypanosoma cruzi. T. cruzi is typically transmitted to humans and other mammals by the bite of "kissing bugs" of the subfamily Triatominae (family Reduviidae), primarily by species belonging to the Triatoma, Rhodnius, and Panstrongylus genera. Rapidly identifying CD insect vectors in the field is crucial to effective control of the disease. Here I describe an image-Euclidean-distance classifier that suppors rapid identification of adults of nine CD vector species.

Keywords: Chagas Disease, automated classification, image Euclidean distance

1.0 Introduction

Chagas disease (CD) disease is a life-threatening tropical parasitic disease caused by the flagellate protozoan *Trypanosoma cruzi*. *T. cruzi* is typically transmitted to humans and other mammals by the bite of "kissing bugs" of the subfamily *Triatominae* (family *Reduviidae*), primarily by species belonging to the *Triatoma*, *Rhodnius*, and *Panstrongylus* genera ([5]).

As many as 11 million people in Mexico, Central America, and South America have CD. Most of those infected do not know that they are. Large-scale population movements from rural to urban areas of Latin America and to other regions of the world have increased the geographic distribution of CD; the disease has been reported in several European countries ([5]).

Rapid field identification of CD vector species is essential to effective vector control. A tool that that could capture an image of a specimen and return its taxonomic identification would help to minimize the labor requirements for rapid field identification of the vectors.

Such a tool can be thought of as having two main subfunctions -- an *image-processing* function, which performs various image-analysis tasks and emits image information derivable solely from color-value distributions on pixel-arrays (e.g., image histograms, maximum and minimum pixel intensity values, image distances, etc.; [15]), and an *automated classification* function, which accepts the output of the image processing and infers the taxonomic classification of the specimen. This study is concerned only with automated classification.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2.0 Method

Images of adult specimens of nine triatomine species, (S), nominally regarded as the most common CD vector species in Brazil ([2]), were selected for classification support:

(S)

- *Triatoma infestans*
- *Triatoma dimidiata*
- *Triatoma brasiliensis*
- *Triatoma maculata*
- *Triatoma sordida*
- *Rhodnius prolixus*
- *Rhodnius neglectus*
- *Rhodnius pallescens*
- *Panstrongylus megistus*

Images representing the species in (S) were randomly selected from [1] and [16]-[18]. From this set, I selected images I judged to have "similar" focus, orientation, contrast, and specimen "quality", excluding images I judged to have one or more of the following "pathologies":

- specimens with missing or occluded legs
- specimens showing significant anatomical, damage, or color anomalies
- in the case of [2]-[3], images evidently containing colors other than white or gray in their background
- specimens that had poor contrast in any region

Using the *Mathematica* ([9]) script shown in Figure 2, digitized images of specimens of species in (S) were imported from [1] and [16]-[18], "standardized" to remove various artifacts, and the image Euclidean distance from each member of a set of equivalently standardized reference images from [1] and [16]-[18] that is disjoint from the set of test images was computed.

The classification of the reference image that has the smallest distance to a given test image is the predicted classification of that test image.

The script shown in Figure 2 was executed 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* operating environment.

3.0 Results

Example triatomine images are shown in Figure 1.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

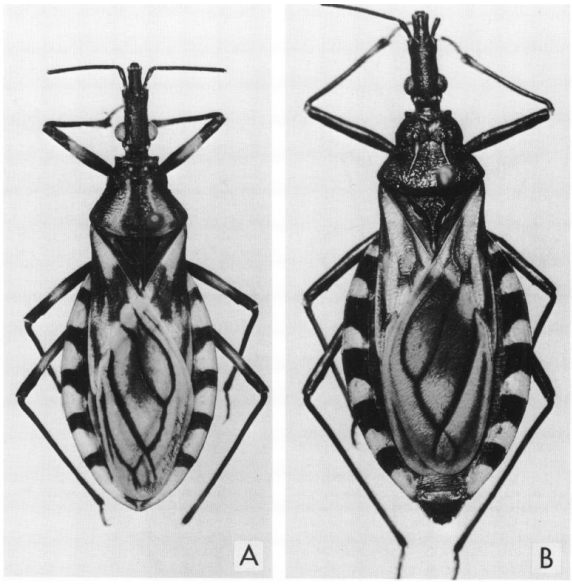


Figure 1. ***Triatoma brasiliensis*****. A. Male Ceará, Brazil. B. Female, dark form, Bahia, Brazil. Adapted from Figure 43, [1].**

```
**********************************************************************************************************

imgwidth=1000;
imgheight=1000;
maxdistimg = 10000.0;
filetype="JPG";
binarizeThreshold=0.05;

testfilelist={"C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae
UnB new photos\\Triatoma_sordida_Posse_GO_d (8)
f","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Triatoma_sordida_Posse_GO_d (10)
m","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Rhodnius
robustus\\Rhodnius_robustus_UnB_Marabai□-PA_d (2)
F","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Triatoma
costalimai\\costalimai_UnB_Mambai-GO-d-3-
M","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Triatoma
pallidipennis\\Triatoma_pallidipennis_UnB_d","C:\\Biodiversity_Institute\\Image_Generator\\Image_
Distance\\RODRIGO2\\triatominae UnB new photos\\Triatoma_brasiliensis_FeiradeSantana_BA_d (3)
f","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Triatoma
infestans\\Triatoma_infestans_UnB_d (8) M"
,"C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Rhodnius
neglectus\\RN_UnB_Curaca-
BA_2128_d","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae
UnB new photos\\Triatoma_sordida_Posse_GO_d (12) m",
"C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Panstrongylus_megistus_Brasília_DF_d (16)
m","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Rhodnius_prolixus_Salvador Colony _BA_d (2)
f","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Will_Tdim_batch1\\Photos_JMR\\Td
Manacal Alcohol\\Manacal Td 1 Alcohol\\IMG_0051"};

reffilelist={"C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\T
riatoma pallidipennis\\Triatoma_pallidipennis_UnB_d
(3)","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Triatoma
infestans\\Triatoma_infestans_UnB_d (4)
```

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

```
M","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Rhodnius
neglectus\\RN_UnB_NovoPlanalto-
GO_910_d","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB
new photos\\Triatoma_sordida_Posse_GO_d (11) m",
"C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Panstrongylus_megistus_Brasília_DF_d (14)
m","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Rhodnius_prolixus_Colony Salvador_BA_d (1)
f","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Will_Tdim_batch1\\Photos_JMR\\Td
Manacal Alcohol\\Manacal Td 3
Alcohol\\IMG_0117","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triato
minae UnB new photos\\Triatoma_brasiliensis_FeiradeSantana_BA_d (6)
f","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Triatoma
costalimai\\costalimai_UnB_Mambai-GO-d-5-
M","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\Rodrigo\\ColRodrigo\\Rhodnius
robustus\\Rhodnius_robustus_UnB_Marabaì-PA_d (4)
F","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Triatoma_sordida_Posse_GO_d (14)
m","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Triatoma_sordida_Posse_GO_d (3)
f","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Triatoma_infestans_Cepeim_ARGENTINA_d (1)
f","C:\\Biodiversity_Institute\\Image_Generator\\Image_Distance\\RODRIGO2\\triatominae UnB new
photos\\Rhodnius_neglectus_CampoAlegre_GO_d (18) m"};

(* loop over test files *)
For [i=1,i≤Length[testfilelist],i++,
 testfilename= testfilelist[[i]]<>"."<>filetype;
    orgimage=Import[testfilename,filetype];

 (* "standardize" the test image *)
    border=Binarize[orgimage,binarizeThreshold];
    mask=ColorNegate[border];
    maskedImg=ImageApply[1&,orgimage,Masking→mask];
    stdtestimg=ImageResize[ImageCrop[maskedImg],{imgwidth,imgheight}];


    oldimgdiff=maxdistimg;
    newimgdiff=maxdistimg;
    bestrefmatch="nothing";

 (* begin loop over reference files *)
 {For [j=1,j≤Length[reffilelist],j++,
   reffilename= reffilelist[[j]]<>"."<>filetype;
            orgimage=Import[reffilename, filetype];

   (* "standardize" the reference image *)
            border=Binarize[orgimage,binarizeThreshold];
            mask=ColorNegate[border];
            maskedImg=ImageApply[1&,orgimage,Masking→mask];
            stdrefimg=ImageResize[ImageCrop[maskedImg],{imgwidth,imgheight}];

   newimgdiff=ImageDistance[stdrefimg,stdtestimg];If[newimgdiff≤oldimgdiff,
     {oldimgdiff=newimgdiff; bestrefmatch=reffilename}, oldimgdiff=oldimgdiff]]};
              (*  end loop over reference files *)

  Print ["The best reference file match to test file ", testfilename,  " is ",
        bestrefmatch, " ; Euclidean distance = ", oldimgdiff]]  (*  end loop over test files *)
```

Figure 2. A nominal *Mathematica* ([9]) script for the image Euclidean distance classifier used in this study.

All 12 of the images in testfilelist (see top of Figure 2) were correctly classified by the script shown in Figure 2.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

On the platform described in Section 2.0, the classification took approximately 30 minutes per image. CPU usage peaked at ~50%. Memory usage peaked at ~7 GB.

4.0 Discussion and conclusions

The results of Section 3.0 motivate several observations:

1. An image-Euclidean-distance classifier can provide automated support for CD vector classification.

2. The test and reference file lists shown in Figure 2 can in principle be extended indefinitely. Those lists could of course be read from files instead of being implemented as string literals in the script.

3. In principle, the standardized reference images could be kept in memory, instead of being recreated for each test image as is done in the script shown in Figure 2. When I attempted to implement this approach, however, I encountered memory-management problems on the platform described in Section 2.0.

4. In principle, the image "standardization" shown in the script in Figure 2 could be abstracted to a *Mathematica* function. When I attempted to do this, however, I encountered memory-management problems on the platform described in Section 2.0.

5. The performance of the classifier depends significantly on the standardization of images prior to classification.

6. The selection of training and test images based on a judgment of image "quality", as described in Section 2.0, would seem to involve expert judgment in some fundamental way. However, [1] and [16]-[18] strongly suggest that it is possible to define a photographic protocol that consistently produces images that allow the classifier to perform as well as noted in Section 4.0. The explicit characterization of that protocol is an ongoing research project coordinated by the University of Kansas Biodiversity Institute.

5.0 Acknowledgements

R Gurgel ([16],[17]) and JMR Willoquet ([18]) provided the images used in this software. This work benefited extensively from discussions with Ed Komp ([10]) and Town Peterson. For any infelicities that remain, I am solely responsible.

6.0 References

[1] Lent H and Wygodzinsky P. *Revision of the Triatominae (Hemiptera, Reduviidae), and Their Significance as Vectors of Chagas' Disease. Bulletin of the American Museum of Natural History* 163:3 (1979).

[2] Martinez D. Informal communication. 2013.

[5] US Centers for Disease Control. Parasites -- American Trypanosomiasis (also known as Chagas Disease): Detailed FAQs. http://www.cdc.gov/parasites/chagas/gen_info/detailed.html. 2012.

[6] US Centers for Disease Control. Parasites -- American Trypanosomiasis (also known as Chagas Disease): Antiparasitic Treatment. http://www.cdc.gov/parasites/chagas/health_professionals/tx.html. 2012.

[9] Wolfram Research. *Mathematica Home Edition v9.0.0*. http://www.wolfram.com/mathematica-home-edition/. 2013.

[10] Komp E. Informal communication. May 2013.

[15] Petrou M and Petrou C. *Image Processing: The Fundamentals*. Second Edition. Wiley. 2010.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[16] Gurgel R. *Rodrigo/ColRodrigo*, a triatomine image collection. Circa June 2013.

[17] Gurgel R. *RODRIGO2/triatominae UnB new photos*, a triatomine image collection. Circa June 2013.

[18] Willoquet JMR. *Will_Tdim _batch1/Photos_JMR*, a Triatoma dimdiata image collection.
Circa June 2013.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Blood Vessel Detection in Images from Laser-Heated Skin

Alireza Kavianpour, Simin Shoari, Behdad Kavianpour
CEIS Dept. DeVry University, Pomona, CA 91768

Abstract

A computer method for recognizing blood vessels in an image constructed by infrared tomography is proposed. Blood vessels detection is important for efficient clinical treatment of a patient.
A blood vessel can be model as a block in an image and hence block detection algorithm is applied on an image constructed by infrared tomography. The approach is based on the parallel calculation on hypercube. Hypercube architecture of dimension n is fine-grain architecture with 2^n processors each holding a single pixel of the image. The hypercube operates in an SIMD mode. Two algorithms for computing blocks are explained.

Key Words: *Block detection, blood vessel, hypercube, image processing, and parallel processing.*

1 Introduction

Three dimensional reconstructions of the blood vessels in an image taken from part of skin by infrared tomography is an important clinical problem. For example efficient clinical results for laser therapy is based on the selection of proper pulse durations of laser [1,2,3,9] Milner et al. [7] have shown that for efficient treatment of Port of Wine Stain (PWS) (patients with birth's mark), or removing tattoos, laser pulse durations should be approximately equal to the thermal relaxation times of the targeted PWS blood vessels. In this paper we utilize an infrared detector to measure temperature changes induced in a skin exposed to laser radiation. Heat generated due to light absorption by skin diffuses to the surface and results in increased infrared radiation emission levels. By collecting and concentrating the emitted radiation on to an infrared detector, useful information regarding the tested skin will be derived and will be used for blood vessels detection. Vessel diameter varies greatly from patient to patient and even from site to site in one patient. The value of thermal relaxation times, *tau_r*, is directly proportional to the square of the diameter (*d*) of the tested vessel and inversely proportional to the thermal diffusivity of skin, $tau\_r = d^{\wedge 2}/16chi$ where *chi* is thermal diffusivity of skin ($0.9x10^{-3}$ cm^2/s).

The Beckman Laser Institute and Medical Clinic in Irvine, California has a laser that user can specify pulse duration over the range of 0.25-15 msec. With such a laser, proper selection of pulse duration of laser exposure for patients is important.

In this paper we model blood vessels as blocks and hence block detection algorithm will be used.

2 Experimental Results

All the experimental data used in this paper have been accomplished at The Beckman Laser Institute and Medical Clinic. In this center a 0.45 msec pulsed laser source (lambda = 585) with adjustable pulse durations is available. A high-speed Infrared Focal Plane (IR-FPA) camera system takes image of individual laser heated blood vessels. This camera acquires 217 infrared emission frames per second. The infrared signals collected by each detector element are digitized by 3.5 MHZ, 12-bit A/D converter and results are stored in computer. Figure 1 represents infrared tomography instruments.

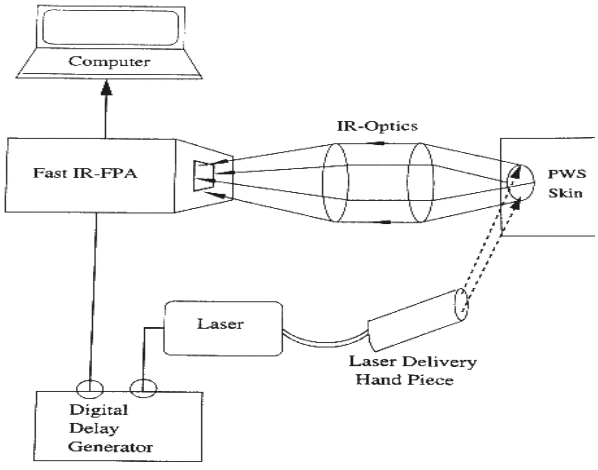


Figure 1 Infrared Radiometery instrumentation

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Each frame has 128x128 pixels. Infrared tomography uses a fast infrared focal plane array to detect temperature rises in a substrate induced by pulse radiation. In practice, a pulsed laser is used to produce transient heating of the object under study. The temperature rise, due to the optical absorption of the pulse laser light, creates an increase in infrared emission which is measured by a fast IR-FPA. If a pulse laser source is used to irradiate the skin, an immediate increase in infrared emission will occur due to optical absorption by hemoglobin contained within the blood vessels. An Infrared tomography record of a skin in response to pulsed laser exposure is composed of a sequence of infrared emission frames that indicate localized heating of blood vessels in tested skin. The value of each pixel is presented by *P(x,y,t),* where *x,y* are coordinates and *t* is the measured time sequence. Sample rate is 217 frames per second.

3 Block Detection Algorithm

In this section we present an algorithm for detecting connected black pixels. The input to this algorithm is *n* by *n* binary pixels.
Connectivity among pixels can be defined in terms of their adjacency. Two black pixels (x_1, y_1) and (x_2, y_2) are 8-neighbor if:

$$\text{Max}\{|x_1-x_2|,|y_1-y_2|\}<=1$$

and 4-neighbor if:

$$|x_1-x_2|+|y_1-y_2|<= 1$$

Two black pixels (x_1, y_1) and (x_k, y_k) are said to connected by 8-path (4-path) if there exists a sequence of black pixels (x_p, y_p), $2<= p<=k$, such that each pair of pixels (x_{p-1}, y_{p-1}) and (x_p, y_p) are 8-neighbors (4-neighbors). A maximal connected region of black pixels is called a connected block. We assume a block represents a blood vessel. In this paper we use 8-neighbors method.

3.1 Depth calculation of vessels

In this section depth of a vessel z will be calculated. Assume *P* is a set of pixels indicating vessel *V*. From 150 frames a curve indicating the relation between temperature change and time will be drawn. The time difference between initial jump and maximum point of this curve is called *delayed thermal peak* t_d .Using the equation $z=sqrt{4chi\ t_d}$, depth of a the vessel *V* will be calculated, where *chi* is thermal diffusivity of skin 0.9x10^-3 cm^2/s.

4 Hypercube Architecture

Hypercube multi-computer systems have become a subject of considerable interest to the system designers faced with challenging applications. An *n*-dimensional hypercube multi-computer system, or an *n*-cub for short, contains 2^n processors each of which is a self-contained computer with its own local memory. Each processor is assigned a unique *n*-bit address. Two processors are linked if and only if their addresses differ in exactly one bit position. Therefore, each processor has direct communication links to *n* other processors [4,5,6].
A hypercube computer with the dimension *n* is Single Instruction Multiple Data (SIMD) machine. All of the processing elements in the hypercube operate in a strict SIMD mode under the direct control of a single node. Each processing element has its own memory and all of the communication links are bidirectional. The hypercube topology has been proposed as architecture for high-speed image processing where its simple geometry adapts naturally to many types of problems.

5 Simulation Program

Application software called Application Visualization System (AVS) is used to process input images. Input images consist of 150 frames. The first few frames and several of last frames will not be used for processing since the early frames are totally black and last frames are white. In order to get better images with less blur and noise, longitudinal inversion algorithm is used. The new gray-level images are transforming into black and white (binary) images.

5.1 Network on AVS:

To implement algorithms on binary images a network on AVS with following blocks are defined:

File Description: This file reads 150 frames of input images each of size 128x128.

Orthogonal Slicer: By setting a parameter one of the 150 images is selected.

Crop: This file trims borders of selected image.

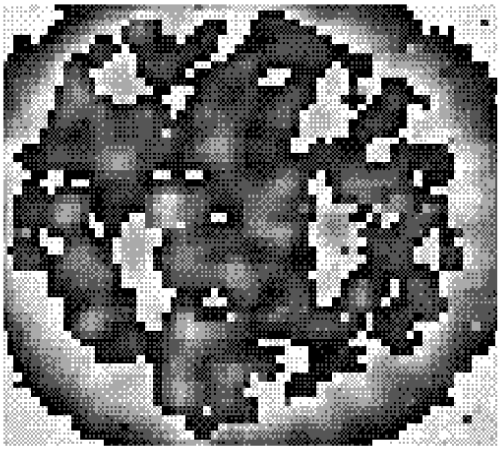
Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Clamp and Contrast: This file changes the gray-level image to black and white or binary images. The format of new image has been changed in order to be readable by processing program.

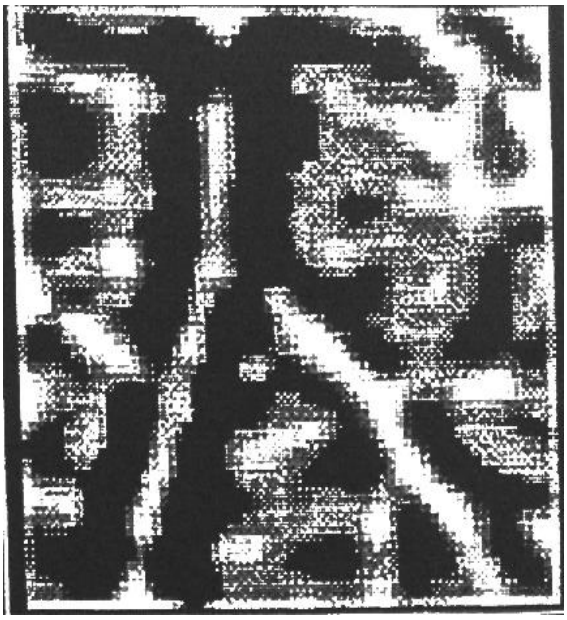
White Image: Stores the image file on the disk.

5.2 Processing Program:

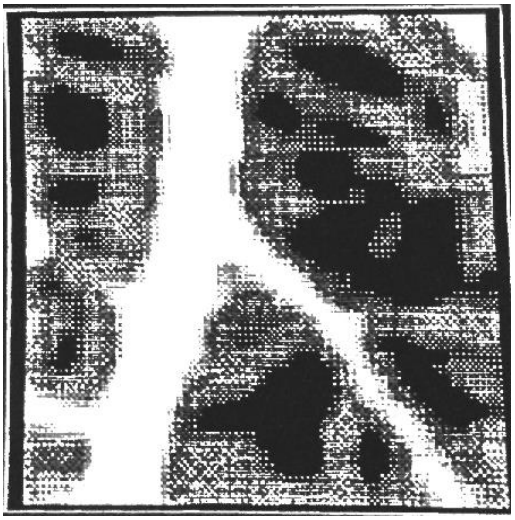
The simulation program for detecting blocks was written in C language. Figures 2 and 3 illustrate the result of block detection algorithm on a binary images.



(a) Raw image



(b) Frame 1



© Frame 4

Figure 2. Different frames from 3D tomographic CAM image

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Figure 3. Detected blocks in frame 1

Table 1 illustrates the computer simulation results for *128x128* size binary images using Block Detection Algorithm for frame number nine. Table 2 illustrates the computer simulation results for *128x128* size binary images using Block Detection Algorithm for block number B_5.

Table 3 illustrates the computer simulation results for *128x128* size binary images using Block Detection Algorithm for Different blocks.

Table 1 Results of simulation for block detection algorithm for frame #9

Block Number	*Number of Pixels*	*Diameter of Block*
B_1	5	2
B_2	6	4
B_3	5	10
B_4	8	4
B_5	22	7
B_6	9	4
B_7	15	5
B_8	25	7
B_9	2	2
B_{10}	7	3
B_{11}	6	5
B_{12}	4	2
B_{13}	23	5
B_{14}	12	3

Table 2 Results of simulation for block detection algorithm for block B_5

Frame Number	*Number of Pixels*	*Diameter of Block*
9	22	5
10	15	4
11	10	3
12	5	2
13	5	2
14	1	1
15	1	1
16	1	1

Table 3 Results of simulation for block detection algorithm for block B_i

Block Number *Number of Frames per Block*	*Maximum Number of Pixels*	*Maximum Diameter*
B_1 10	5	2
B_2 10	6	25
B_3 10	5	18
B_4 10	8	44
B_5 10	22	22
B_6 10	9	17
B_7 10	15	5
B_8 10	25	10
B_9 10	2	16
B_{10} 10	7	58
B_{11} 10	6	58
B_{12} 10	4	58
B_{13} 10	4	58
B_{14} 10	4	58

6 Summary

In this paper blood vessels are model as blocks. Parallel algorithms for detecting blood vessel on hypercube architecture are described. The result of simulation proves the usefulness of block detection algorithms in image processing and pattern recognition.

7 References

[1] Elisa R, Renzo P "Retianl Blood Vessel Segmentation Using Line Operators and Support Vector**"** *IEEE Trans Medical imaging***,** 26**:**1357-1365, 2007

[2] M. Bern, "Laser Surgery," Scientific American, pp. 84-90, June 1991

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[3] G. Bongiovanni, C. Guerra, and S. Levialdi, "Computing the Hough Transform on a Pyramid Architecture," Machine Vision and Applications, vol. 3(2), pp. 117-123, 1990

[4] A. Kavianpour, and N. Bagherzadeh,"Circle Detection in Black and White Images, "Patent pending, UC Case no. 91-195-1, 1991

[5] A. Kavianpour, and N. Bagherzadeh,"Finding Circular Shapes in an Image on a Pyramid Architecture, "Pattern Recognition Letters, vol. 13, no. 12, pp. 843-848, December 1992

[6] A. Kavianpour, and N. Bagherzadeh,"Parallel Line Detection for Image Processing on a Pyramid Architecture" Journal of Pattern Recognition and Artificial Intelligence} vol. 8, no. 1, pp. 37-349, 1994

[7] T. E. Milner, B. Anvari, S. Nelson, B. S. Tanenbaum, and L. O. Saasand "Imaging Laser-Heated Cutaneous Chromospheres," OSA Preceding on Advanced in Optical Imagining and Photon Migration vol. 21 pp. 269-271, 1994

[8] S. Kime, L. O. Svaasand, M. H. Wilson, M. J. Schell, T. E. Milner, S. Nelson, and M. W. Berns "Differential Vascular Response to Laser Photo thrombolysis, "The Journal of Investigative Dermatology vol. 103, no. 5, pp. 693-700, 1994

[9] Cemil Kirbas and Francis Quek "A Review of Vessel Extraction Techniques and Algorithms" Vision Interfaces and Systems Laboratory (VISLab), Department of Computer Science and Engineering ,Wright State University, Dayton, Ohio, January 2003

[10] Bankhead P, Scholfield CN, McGeown JG, Curtis TM "Fast Retinal Vessel Detection and Measurement Using Wavelets and Edge Location Refinement". PLoS ONE 7(3): e32435. doi:10.1371/journal.pone.0032435, 2012

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Proposal of Smart Blood Banks Central Distribution System in Saudi Arabia

Albatoul Althenyan[a] and Samir El-Masri[b]

[a]*Imam Muhammad Bin Saud University, College of Computer and Information Sciences, Department of Information Systems*

[b]*College of Computer and Information Sciences, King Saud University, Saudi Arabia*

Abstract. The goal of the Smart Blood Banks Central Distribution system is to be a substitute for the blood center in Saudi Arabia and to search for the required quantities of each blood type, save time, improve the process of exchange and distribute the blood units before expired. It will also promote a consolidation of hospitals in one blood bank and cooperation among them. These tasks need to develop an online intelligent distribution system that links all the hospital blood banks through one central system. The system should be able to process, store, distribute and exchange blood and blood components between blood banks under some conventions and regulations.

Keywords. Blood Bank, Blood Center, Blood Unit, Utilization, Smart System, Proposed System.

Introduction

The utilization of blood and blood components is increasingly becoming an important and indispensable component of clinical care especially in the surgically-related fields. In line with the hospitals emphasis on patient safety when facing shortages in blood and its components, they have to seek assistance from other blood banks. When we look to the blood bank system in Saudi Arabia, we don't see the concept of central blood banks. Meaning, there are more than 60 blood banks in the middle region works individually without any connection among them [1]. The lack connection led to significantly wastage for the most required and rarely source that annually may reach about 20% especially in medium and small banks. The presence of those gaps leads us to propose such a system [2].

Through the extensive literature review about systems and papers that discusses similar idea in the same field we found that they are rarely and probably not exists. The presence of blood centers in other countries helped to eliminate a lot of problems that we are trying to resolve in the proposed system. Most of the systems that we found are basically depend on the concept of management in the first place. Those systems are focus on managerial

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

functions such as reporting, inventory, accounting and monitored distribution of blood bags, and they are depending on connecting hospitals to one blood center, which means they always have the same source [3].

1. New Proposed System

Blood requests in Saudi Arabia are not in line with international stander and best parties; the concept of blood center is not exists yet like in other wealthy country that will defiantly reflect less complains of blood and its components. The main target of creating blood bank center in Saudi Arabia is to establish new policies / procedures in order to manage, control the process of requesting or replacement of blood and maintain the quality with high standard.

During side visits was done in Riyadh area and interviewing with Operations Analysts and Q.A. Coordinators of Blood Bank & Transfusion Services of Pathology & Laboratory Medicine in the largest blood banks, we realize all blood banks are struggling to get enough quantity to run their business. The idea of having large numbers of blood banks in same area without clear strategy, will negatively impact blood donators and patient life and lack the quantity of blood units in important hospitals which have much need, while other blood banks have a lot of not used blood units near to their expiry date.

After that, we recommend linking all blood banks in one smart distribution system. The system should be smart enough to meet the demands and achieve the balance between the m. Moreover, the system will control the distribution between different entities with consideration of priority, and when there are more than one request for a specific blood type, the system try to distribute the orders to several hospitals not to just one, taking into account a combination of factors. The effective way to build the system is make integration between the smart system and the blood banks systems to create connection among them. So, the system will play the main role as blood bank center which contain the individuals participant.

Trying to connect a large number of hospitals is not easy, for that we need a model or an algorithm to organize this correlation and facilitated at the same time. If we consider each hospital as a **Node(j)**and the connection between the system and hospitals as arcs with two directions, and the distance between hospitals on these arcs is **X(i)**, it is possible to say we are talking about network model that have multiple source and destinations (Hospitals) Figure 1 [4]. This model supported by Linear Digital Programming algorithm which helps us to apply the calculations for the supply and demand on the blood units [5].

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

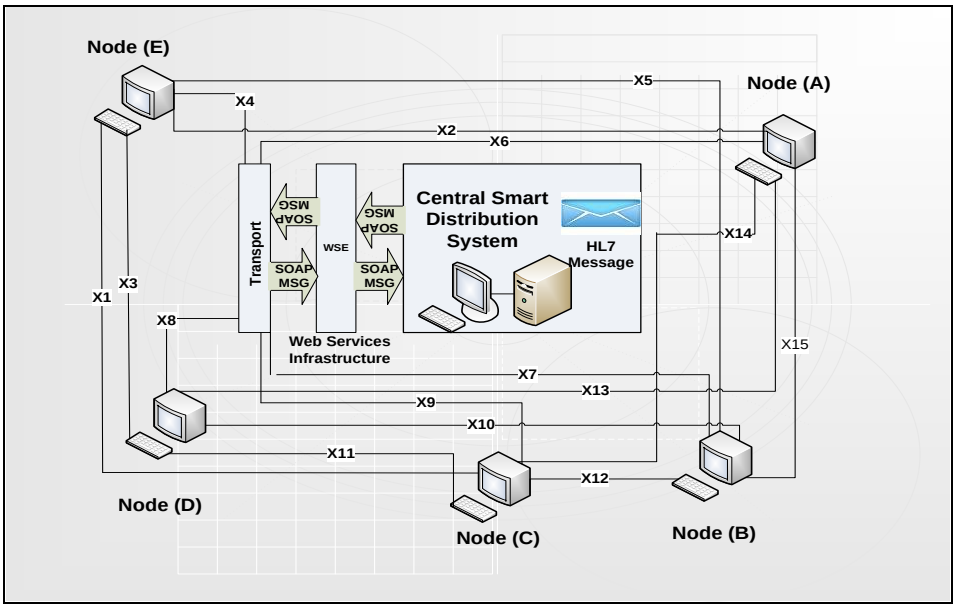


Figure 1. Proposed System Architecture

For each node (hospital) there are: constant minimum number of blood units, daily reserved blood units (reserved units for next day), updated minimum number and the available stock. By these numbers we can make equations to calculate then conclude the needy hospitals and the hospitals have surplus blood units as illustrated in the following case study.

1.1. Case Study

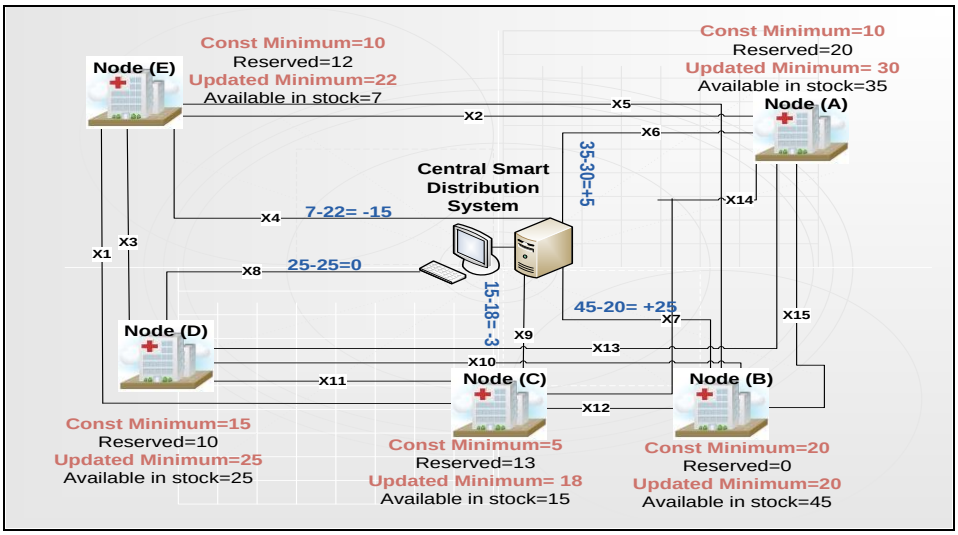


Figure 2. Ordering Representation Model

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

As represent in Figure 2 the ordering model for the smart system, we can see the nodes: A, B C, D and E are hospitals connecting to the central smart system. This case study was done for one blood type (e.g. O+) and the rest are in the same way. By the set of numbers available on each node we can calculate them to extract the status of each hospital by the following equation:

(Available Stock) - (Updated Minimum) = Status;

Note that the updated minimum is not a piece of information provided by the hospital, it is calculated by the system as follow:

(Constant Minimum) + (Daily Reserved) = Updated Minimum;

By the status sign and the number together we can conclude which hospitals are need blood and witch are supply it. For example in node **A** the result is (+5) the plus sign indicates that hospital **A** has extra five blood units that it can provide for other hospital. While in node **E** the result is (-15) the negative sign indicates that hospital **E** need for fifteen blood units. The system try to find this quantity from other hospital that have plus sign, it is not must be all from one hospital as it must achieve the balance between the taking units and the hospital size with the available blood stock and location. But in node **D** we find the result is equal to (0) which mean no need for more blood units and there are no supply units. Before the system start processing the orders, it must check the existence one of these cases:

- **Case#1:** Supply> Demand: The system will not facing any problem when the supply blood units are more than demands. In addition, there are still more blood units for supply in the system.
- **Case#2:** Supply< Demand: The problem will happen in such this case. In this situation the system starts to solve the highest priority orders. Priority is the value of negative sign number of result; the biggest value means the highest priority (e.g. -15 is high priority than -3). Thereafter, the smart system tries to provide the blood from available stock. When all available stock is consume and there are still high priority statuses, the system will enforce to withdraw the requested blood units from the constant minimum of appropriate hospital, and the needy hospital whose takes this blood units is responsible to return them as soon as they available in their stock.
- **Case#3:** Supply = Demand: In this case, all blood units provided from all hospital for this blood type were depleted, and no more blood units for providing.

The solution tries to make the best optimization for blood distribution across all linking hospitals by calculate the shortest distance between hospitals to increase the delivery speed. For example: when node C order blood units and node E will supply the units, the system must check the shortest path to node C, if node D need blood that can supply from node E the shortest path will (E-D-C= X3+X11), but if node D have status (0) which mean it doesn't has an order the shortest path will direct from E to C(X1). The system will distribute the blood twice time in a day for the best network exploiting. Before starting the calculations the system will receive all information from all hospitals through XML files, which refers to bigness of receiving data, and saving all in the system database.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2. Discussion

As of today, the existing blood banks system in Saudi Arabia is incomplete, the reason behind that, poor cooperation between different facilities in relation to data and blood utilization that may led to wasting large amounts of blood/components. However, the proposal will provide blood and blood components faster and easier in order to save patients life. We can said the system has currently benefits by supplying the needy hospital with sufficient quantity of blood units as soon as possible, and the beauty of proposed system is to streamline the blood bank industry as in:

- Maximum utilization of resource with less waist of blood units.
- High level of integration between different organization or health care provider,
- Generate individual statistics reports for each hospital and comparing between them to find out which hospitals constantly order amounts of blood and accountability to know where the defects in their work.
- Planning and estimation the future needs of blood bank activities, that is help the hospitals to intensification blood donation campaigns in case the expected were large quantities.

3. Conclusion

The proposed smart solution can totally replace the current practice of ordering via telephone calls. The system will facilitate and simplify the workflow between different blood banks as well as blood data / storage. It also eliminates the problem of blood shortage and make sure to use blood before it is expired. After reviewing such a case and studying the opportunities, the system architecture will be the optimum solution to achieve our expectations.

References

[1] NATIONAL BLOOD CENTRE MINISTRY OF HEALTH MALAYSIA ."Transfusion Practice Guidelines for Clinical and Laboratory Personnel". March 2008.

[2] "Do we need an organization for blood transfusion services?". http://www.alriyadh.com/2013/03/10/article816280.html. Last accessed Apr 2013.

[3] "Online blood bank management system Netbloodbank". http://www.netbloodbank.com. Last accessed Jun 2013.

[4] " Web Services Enablement for Healthcare HL7 Applications - Web Services Basic Profile Reference Implementation". http://msdn.microsoft.com. Last accessed Oct 2013.

[5] Lawrence, Jhon A. and Pasternack, Barry A. *Applied Management Science: Modeling, Spreadsheet Analysis, and Communication for Decision Making* Jhon Wiley &Sons, United States, 2002, pp.206-209.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Predicting the Co-receptors (CCR5 and CXCR4) of the Viruses R5, X4 and R5X4 that Cause AIDS (HIV-1) in Cells CD4 using The Bayes Classifier

Francisco Javier Luna Rosas[1], **Julio Cesar Martínez Romo**[1],
Carlos Alejandro de Luna Ortega[2], **Ricardo Mendoza Gonzalez**[1,] **Valentín López Rivas**[1]
Gricelda Medina Veloz[3]
e-mail: fcoluna2000@yahoo.com.mx

[1] Computer Science Department, Inst. Tec. Aguascalientes, México.
[2] Universidad Politécnica de Aguascalientes, México
[3]Universidad Tecnológica del Norte de Aguascalientes, México

Abstract – *The harming presence of retrovirus in CD4 cells of the immune system Known as AIDS (HIV-1), disease that is widespread throughout the planet. In particular, the R5 HIV-1 viruses use CCR5 as co-receptor for the virus entrance, the X4 virus HIV-1 use the CXCR4, while some strange viruses known as R5X4 or D-tropic, have the ability to use both co-receptors. A series of experiments were performed in this article to implement a Bayes Classifier that allows to asses different patterns that enable us to predict the co-receptor of the mutated R5X4.*

Keywords: Co-receptors, Viruses, AIDS, Bayes Classifier.

1 Introduction

The breast, lung, colon and prostate cancer are the most common today [1], but the frequency of cases of other types of cancer has increased, that is the case of cancer in blood, lymphatic system, skin, digestive system and the urinary system. At the same time, with the existence of several types of cancer, multiple factors exist that favor the appearance of cancer in different parts of the human body; so it is the case with blood cancer as the leukemia caused by the presence and action of a retrovirus (deltaretrovirus), like the syndrome of acquired immunodeficiency (AIDS) that is caused by the presence and action of another one retrovirus known as virus human immunodeficiency or HIV (Lentivirus), affecting present cells CD4 in the blood directly. This work focuses specifically in the study of these cells of the blood called CD4 that are a type of lymphocytes (leukocytes). These are an important part of the immune system in a human being. These cells are also called "T" cells and are attacked by retrovirus HIV.

2 Antecedents

In 1981, the International Committee on the Taxonomy of the Virus (ICTV) [2] proposed the following definition: "A species of the virus is a concept that will be represented normally by a group of chains of a variety of sources, or a population of chains of a particular source, that have in common a system correlated properties stable that differentiate a group from other groups of chains" [2]. Today, the ICTV recognizes more than 3,600 species of the virus. [2].
The animal viruses are classified in six classes: I, II III, IV, V, and VI.

2.1 Class VI Retroviridae

It is a group of virus of RNA that infects animals and human beings.

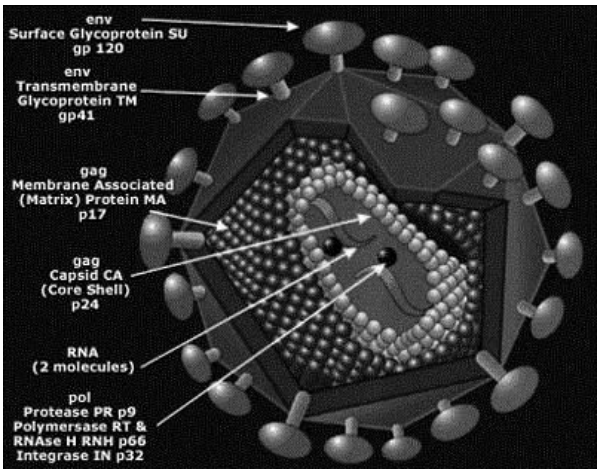


Fig. 1. Retrovirus Structure [3]

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

As it can be observed in Fig. 1, the general structure of retrovirus is mainly formed by a lipid-protein bilayer with two protein subunits that are codified by the gene env of the virus and their own lipid and protein components of the cellular membrane. Also called specific glycoprotein that cause the infection, which contains a spherical capsid formed by 3 protein subunits codified by the gene gag. Inside the necessary viral enzymes for the process of viral replication are found: protease (gene pro) and inverse transcriptase and integrase, codified by the gene pol. Some retrovirus cause cancer directly, integrating genes called oncogenes in the DNA of the cell guest, causing the malignant transformation of normal cells into cancer cells, these are called virus transforming acute. Others cause cancer indirectly activating proto-oncogen of the guest, these are called virus transforming not-acute. Another important characteristic is that some retrovirus is cytotoxic for certain cells, inflating them. Most remarkable it is the virus of the syndrome of the immunodeficiency in humans that destroys the lymphocytes CD4 T that they infect.

2.2 Classification of retrovirus

Actually, retroviruses are classified in 7 genera [2], as it can see in Fig. 2, where the different genera are observed from retrovirus according to the family which they belong. So it is the case of HIV pertaining to the family of Lentivirus.

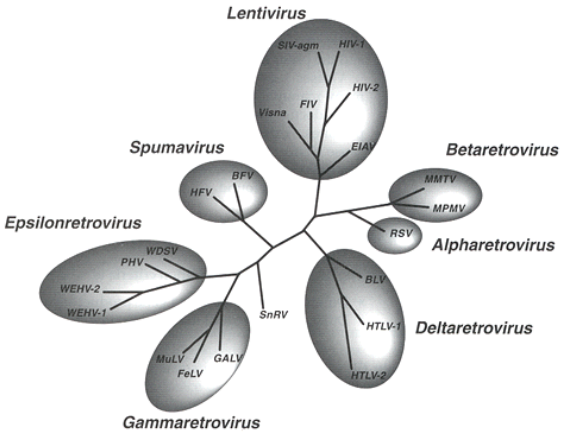


Fig. 2. Filogenetic analysis of retroviruses. (Quackenbush, S. and Casey, J.). Academic Press 2000 [2]

At the present time, the well-known and classified families by ICTV are:

a) Alpharetrovirus, Betaretrovirus, Epsilonretrovirus, Gammaretrovirus: Contain genomes simples.

b) Lentivirus, Spumavirus and Deltaretrovirus. They contain complex genomes.

Only retrovirus with endogenous simple genome and spumaviruses are living in their guests; in latent way. The Lentivirus is a citopatic retrovirus (retrovirus which damages the cell) that causes immunodeficiency syndromes fundamentally, neurological syndromes and autoimmune diseases of slow evolution.

2.3. Vital cycle of retrovirus

The vital cycle of retrovirus begins in the nucleus of an infected cell, (upper part of Fig. 3), formed by the genome of the virus (genes gag, pol, env), is contained into the membrane of the cell guest (envelope) [3].

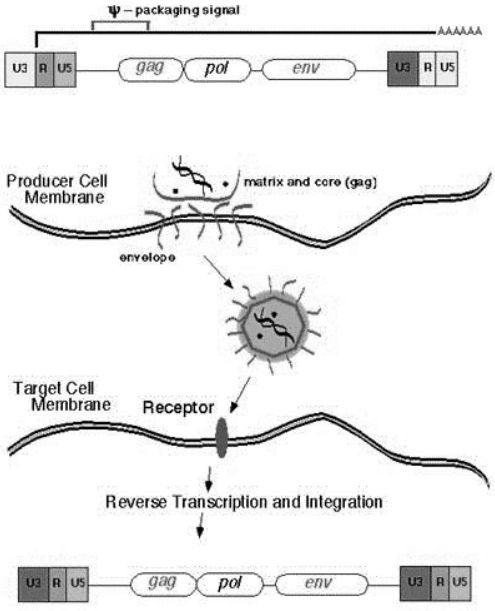


Fig. 3. Vital Cycle of retrovirus [3]

In this stage of the vital cycle the retroviral genome is an element of the DNA integrated in the cell guest. The genome of the virus is approximately 8-12 kilobases of the DNA (it depends on the retroviral species). The genome of the virus takes advantage of elements available in the cell guest to form the capsid that locks up the heart of the virus encapsulating genes and other elements (matrix and Core), immediately the virus leaves the cell guest as a free particle (central part of Fig. 3) and looks for other healthy cells to infect them, which is observed in the bottom part of Fig. 3 (Target cell membrane).

3 Retrovirus more common in humans

The human immunodeficiency virus (HIV) is a nontransforming retrovirus pertaining to the family of

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Lentivirus. In Fig. 4 the structure of retrovirus of the human immunodeficiency is appraised (HIV) [2].

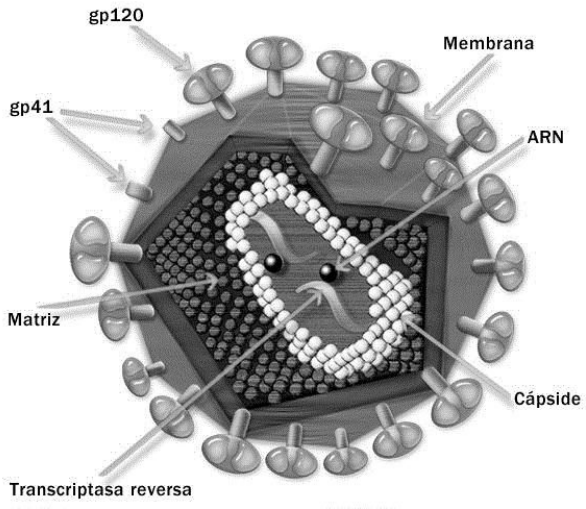


Fig. 4. HIV Structure [3]

3.1 Cellular damages causes by HIV

The human cells that the HIV more frequency infects are the CD4, and when they are multiplied to fight infections, they make more copies of the HIV involuntarily. In infections prolonged by HIV, the numbers of cells CD4 are diminished. This it is a sign of which the immune system has been debilitated. Lower counts of CD4 cells increase the possibilities that the individual will get sick [1],[2].

3.2 Detection of the HIV in humans

The analysis of the VIH means the tests that determine if an individual is or not infected with the human immunodeficiency virus (HIV) that causes the AIDS. Several types of analyses exists that are practiced generally in blood samples of individuals in study, also other corporal fluids samples are used, including a scraped cheek [1],[2].

Antibodies detection analysis. These analyses look for "antibodies" against the HIV in the blood and other corporal fluids. The antibodies are proteins produced by the immune system to fight a specific germ, in this case against HIV which delay two or three months to appear after the organism has become infected[4].

Viral count analysis. Analysis of viral load measures the amount of HIV in the blood. Different techniques exist. The technique Polymerase Chain Reaction (PCR) uses an enzyme to multiply the HIV in the blood sample. Soon a chemical reaction marks the virus, the markers are measured and the amount of virus is calculated [5], [6].

Analysis of CD4 cells. Consists making a count of CD4 and CD8 cells in the sample. The number of cells by cubical millimeter of blood is specified. An agreement does not exist about what is the normal average level of CD4 cells. The normal count of CD4 is between 500 and 1600 cells and CD8 are between 375 and 1100 cells. CD4 cells can diminish drastically in HIV+ people and in some cases they can reach zero. Because the count of CD4 cells varies much, also the percentage of CD4 cells is analyzed. This percentage talks about the totality of lymphocytes. If the analysis indicates that a 34% of CD4 exist, means that 34% of their lymphocytes are CD4 cells. The percentage is most stable than the number of CD4 cells. The normal rank is between 20% and 40%. A percentage underneath 14% indicates serious damage in the immune system. It is a signal of the AIDS in people infected with HIV [7].

4 The tropism of AIDS

The tropism of the Human Immune Deficiency Virus is defined as the highly attractive specification of the virus towards the host tissue, determined partly by the surface markers of its cells (for example CD4 cells). The Viruses develop a specific ability to attack cells in a selective way, such as the host organs and often certain cell populations found in organs of the host body.The patients with AIDS show decreasing of CD4+ lymphocytes as the disease progresses, thus in 1984 it was planned that it was precisely the CD4 molecule the specific receptor, so the VIH virus entered cell [8]. In 1986 it was shown that protein gp120, of the viral wrapping, grouped with the CD4 and both molecules co-acted as an immune complex, thus showing the gp120-CD4 union. The expression of the CD4 in the membrane is necessary but not sufficient so that the virus fuse with the cell, the search for the possible co-receptor extended for about 10 years. Actually, since 1996, the CCR5 and CXCR4 have been identified as the main co-receptors for the for VIH which has led to an understanding of the viral tropism and the pathogenesis in the molecular field [8]. In a few words it was observed from the beginning that all the isolated VIH-1 patients had a different tropism according to the cells that they infected. Some isolated patients infect macrophagus easily while others infect line cells of lymphocites T. Therefore, there are the so called isolated patients M-tropics (R5X4), (The M and the T for macrophages and limphocites respectively). We could say in a sense that the gp 120 of the protein of the virus is the key used by the VIH to enter the cell. The virus R5X4 are the mutated viruses of VIH, which have the ability to attatch to DNA with the co-receptor CCR5 or CXCR4, with the identification of patterns in the

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

normal viruses such as X4 and R5 and the use of a Bayes classifier, we will be able to determine the co-receptor that the R5X4 virus will use.

5 Materials and Methods

5.1 Data Collection

There are 149 HIV isolated sequences representing the three viral tropisms: 77 for M-tropic R5 (Table 3) , 41 for T-tropic X4 (Table 2) and 31 for dual-tropic R5X4 (Table 1), identified by Lamers in [9] and in the National Center for Biotechnology Information (NCBI. http://www.ncbi.nlm.nih.gov/ [10].

Table 1. Accession Numbers for Sequences of Different R5X4 Viruses

R5X4 Viruses			
AB014795	U08445	AF259019	AF112925
AF062029	AF355674	AF259025	M17451
AF062031	AF355647	AF259021	K02007
AF062033	AF355630	AF259041	U39362
AF107771	AF355690	AF258970	AF069140
U08680	M91819	AF258978	AF458235
U08682	AF035532	AF021607	AF005494
U08444	AF035533	AF204137	

Table 2. Accession Numbers for Sequences of Different X4 Viruses

X4 Viruses			
AB014785	X01762	AF258981	U27408
AB014791	L31963	AF259003	AF411966
AB014796	U08447	AF021618	U27399
AB014810	AF355660	AF128989	U08822
U48267	AF355748	M17449	U08738
U08666	AF355742	AF075720	U08740
AF069692	AF355706	U48207	U08193
AF355319	AF180915	U72495	AF355330
AF355336	AF180903	AY189526	
M14100	AF035534	AF034375	
A04321	AF259050	AF034376	

The DNA and the amino acids that make up the gp 120 protein of each virus were obtained from the data base (NCBI http: http://www.ncbi.nlm.nih.gov/) [10]. Table 4 shows the virus category, the access number and the amino acids of several selected viruses from each of the groups that represent the three viral tropisms (R5, X4 and R5X4).

5.2 Generation of Characteristics

Design of their own software in MatLab 2013,WEKA [11] and DNAStar [12] were used to calculate 16 general statistics per position, the first nine were figured out from the properties of the amino acids according to Table 5 with their own software (for example, amino acid type, charge, volume (A3), mass (Daltons), HP Scale, Surface Area, Alpha Helix, B-strand and Turn) and the remaining properties were extracted from DNAStar (molecular weight, strongly basic (+), strongly acidic (-), the hydrophobic, polar amino acids, iso-electric point, charge at PH 7.0, etc.). The properties of amino acids were grouped later in Tables (Data Sets) and loaded to the data mining software WEKA [56]. In In WEKA we figured out some features such as: minimum and maximum global average and standard deviation of all the viruses that belong to each category (R5, X4 and R5X4), for example if we take the charge at PH 7.0, the standard deviation is 0.77 and the average is 9.698 for the viruses R5.

Table 3. Accession Numbers for Sequences of Different R5 Viruses

R5 Viruses			
AF062012	AY010852	M38429	U08453
L03698	U08670	U27443	AF307755
AF231045	U08798	U79719	AF307750
AY669778	AY669715	U04909	AY043176
U08810	U08710	U04918	AY158534
U51296	U16217	U40908	AX455917
AF407161	M26727	U08450	AY043173
AB253421	AJ418532	AF112542	AF307757
U08645	AJ418479	M63929	U08803
U08647	AJ418495	U66221	U88824
AB253429	AJ418514	AF491737	U69657
AY288084	AJ418521	U08779	AF355326
AF307753	U23487	L22084	U88826
AF411964	U04900	U27413	U08368
U08823	AF022258	AF005495	U27426
AF411965	AF258957	U52953	AJ006022
U92051	AF021477	AF321523	U08795
AF355318	U08716	L22940	
AY010759	U39259	U45485	
AY010804	AF204137	AB023804	

Table 4. Protein gp120 of Different Viruses (R5,X4, R5X4)

VIRUSES	ACCESS NUMBER	PROTEIN (GP120)
R5X4	AB014795	VSTQLLLNGSLAEEEIIIRSENLTNNVKNIIVHLNRSVEIN CTRPSNNTRTRVTLGPGRVWYRTGEIIGDIRKAYCEINGT KWNKVLTKVTEKLKGHFNKTVIFQQP
	AF062029	SFDPIPIHYCTPAGYAILKCNDKNFNGTGPCKNVSSVQCT HGIKPVVSTQLLLNGSLAEEEIIIRSENLTNNAKTIIVHLN KSVEINCTRPSNNTRTSLKIGPGQVFYRTGDIIGNIRAAY CEINGTKWNKVLKQVTGKLEEHFKNKTIIFQPPSGGDLEI TM
X4	AB014785	VSTQLLLNGSLAEEEIIIRSENLTNNVKNIVHLNRSVEINC TRPSNPTRTRITMGPGRVWYRTGEITGSIRAYCEINGTK WNKVLKQVTEKLKKHFNKTVIFQQP
	AB014791	VSTQLLLNGSLAEEEIIIRSENLTNNVKNIIVHLNRVEINC TRPSNPTRTRITMGPGRVWYRTGEITGSIKKAYCEINGTK WNKVLKQVTEKLKEHFNKTVIFQQP
R5	AF062012	AGYAILKCNDKNFNGTGPCKNVSSVQCTHGIKPVVSTQ LLLNGSLAEEEIIIRSENLTNNAKTIIVHLNKSVEINCTRPS NNTRTSITMGPGQVFYRTGDIIGDIRKAYCEINGTKWNE
	AY669778	VQARQLLSGIVQQQSNLLRAIEXQQHMLQLTVWGIKQL QARVLAVERYLKDQKFLGLWGCSGKIICTTAVPWNSTW SNKSFEEIWNNMTWTEWEREISNYTNQIYEILTESQNQQ DRNEK DLLELDK

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 5. Amino Acid Properties [52]

Amino acid residues	Charge	Volume(A3)	Masa(Daltons)	HP Scale	Surface Area	2D structure propensity		
						Alpha Helix	B-strand	Turn
Alanine(A)	0	67	71.09	1.8	0.74	1.41	0.72	0.82
Arginine(I)	+1	148	156.19	-4.5	0.64	1.21	0.84	0.90
Aspargine(N)	0	96	114.11	-3.5	0.63	0.76	0.48	1.34
Aspartic Acid(D)	-1	91	115.09	-3.5	0.62	0.99	0.39	1.24
Cystine(C)	0	86	103.15	2.5	0.91	0.66	1.40	0.54
Glutamine(Q)	0	114	128.14	-3.5	0.62	1.27	0.98	0.84
Glutamic Acid(E)	-1	109	129.12	-3.5	0.62	1.59	0.52	1.01
Glycine(G)	0	48	57.05	-0.4	0.72	0.43	0.58	1.77
Histidine(H)	0	118	137.14	-3.2	0.78	1.05	0.8	0.81
Isoleucine(I)	0	124	113.16	4.5	0.88	1.09	1.67	0.47
Leucine(L)	0	124	113.16	3.8	0.85	1.34	1.22	0.57
Lysine(K)	+1	135	128.17	-3.9	0.52	1.23	0.69	1.07
Methionine(M)	0	124	131.19	1.9	0.85	1.30	1.14	0.52
Phenylalanine(F)	0	135	147.18	2.8	0.88	1.16	1.33	0.59
Proline(P)	0	90	97.12	-1.6	0.64	0.34	0.31	1.32
Serine(S)	0	73	87.08	-0.8	0.66	0.57	0.96	1.22
Threonine(T)	0	93	101.11	-0.7	0.7	0.76	1.17	0.90
Tryptophane(W)	0	163	186.21	-0.9	0.85	1.02	1.35	0.65
Tyrosine(Y)	0	141	163.18	-1.3	0.76	0.74	1.45	0.76
Valine(V)	0	105	99.14	4.2	0.86	0.90	1.87	0.41

5.3 Results

We will initially focus on the two-class case. Let ω_1, ω_2 be the two classes in which our patterns belong. In the sequel, we assume that the a priori probabilities $p(\omega_1)$, $p(\omega_2)$ are known. This is a very reasonable assumption, because even if they are not known, they can easily be estimated from the available training feature vectors. Indeed, if N is the total number of available training patterns, and N_1, N_2 of them belong to ω_1 and ω_2, respectively, then $p(w_1) \approx N_1 / N$ and $p(w_2) \approx N_2 / N$. Describing the Bayes rule we have [13][14]:

$$p(w_i \mid x) = \frac{p(x \mid w_i) P(w_i)}{p(x)} \quad (1)$$

Where each of the elements of this equation has the following meaning.

$p(\omega_i / x)$: The probability that a feature vector x, belongs to the class ω_i.

$p(x / \omega_i)$: The probability that given the class ω_i, the value of the random variable is, precisely x. In other words is the pdf of the class ω_i as a random variable.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

$p(w_i)$: The a priori probability of occurrence of an element of the class ω_i.

$p(x)$: The a priori probability of occurrence of an object to classify a feature vector equal to x (considered as a specific numerical vector), this element can be neglected because it has the same value for all classes.

In this way the maximum search is now focused on the values of the evaluating pdf of x [13][14].

The Bayes classification rule can now be declared as:

$$\text{If } P(\omega_1|x) > P(\omega_2|x) \;\; x \to \omega_1 \;\; x \text{ is classified to } \omega_1$$

$$\text{If } P(\omega_2|x) > P(\omega_1|x) \;\; x \to \omega_2 \;\; x \text{ is classified to } \omega_2 \tag{2}$$

Fig. 5 presents an example of two equiprobable clases and shows the variations of $P(x \mid \omega_i), i = 1,2$, as functions of x for the simple case of a single feature $(l = 1)$. The dotted line at x_0 is a threshold partitioning the feature space into two regions, R_1 and R_2. According to the Bayes decision rule, for all values of x in R_1 the classifier decides ω_1 and for all values in R_2 it decides ω_2.

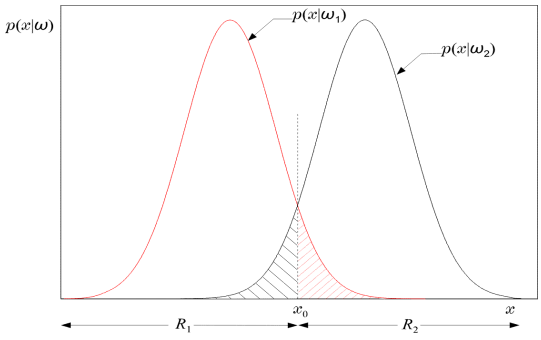


Fig. 5. Example of the two regions R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes[14]

However, it is obvious from the figure that decision errors are unavoidable. Indeed, there is a finite probability for an x to lie in the R_2 region at the same time to belong in class ω_1. Then our decision is in error. The same is true for points originating from class ω_2. It does not take much thought to see that the total probability of committing a decision error is given by [14]:

$$p_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x/\omega_2)dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x/\omega_1)dx \tag{3}$$

Since we have seen that, this classifier is based on managing of the second member of the Bayes theorem:

$$p(w_i/x)\cdot p(w_i); i = 1,2...N \tag{4}$$

What is the probability density function (pdf) of the different classes. $w_1, w_2 ... w_n$. First, we will restrict our study to normal or Gaussian distributions, which occurs in the most practical cases. We begin with the case of one-dimensional distribution for example, when we working with a single feature. Expression of the pdf is as follow:

$$p(x/w_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp^{-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2}} \tag{5}$$

Where μ_i y σ_i are the average and standard deviation typical.

Table 6. Measuring Classification Error Probability

	Viruses R5 (%)	Viruses X4 (%)	Error R5	Error X4	Optimal Threshold
Molecular Weight	99.01%	99.12%	0.981	0.871	7.85
Hydrophobic Amino Acids	99.997%	99.998%	0.003	0.002	6.53
Isolectric Point	99.83%	99.84%	0.162	0.153	7.43
Charge at PH 7.0	99.97%	99.99%	.03	0.01	6.953

We consider general statistics of molecular weight, iso-electric point and charge at PH for the two categories of viruses R5 and X4, and we first got the PDF of each one of the characteristics based on its μ_i and σ_i and for both types of ω_1 and ω_2 later, we calculated the classification error (3) in the zone in which PDFs are trans lapped (see Fig. 5). 100,000 data that fallow a normal behavior were considered to value the stoutness of the classifier. Table 6 shows the classification average, reason of error, and optimums threshold classification for the features previously mentioned.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

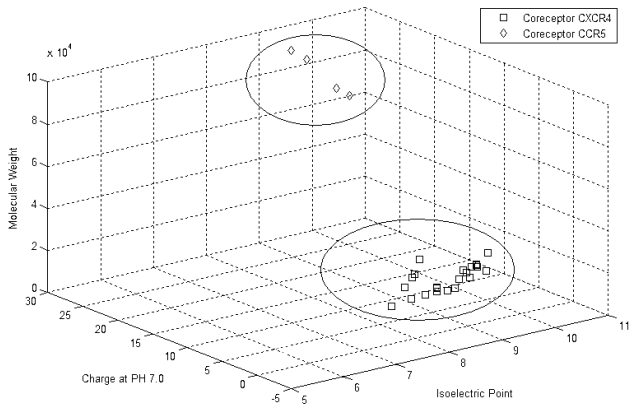


Fig. 6. Predicting the co-receptors CCR5 or CXCR4 from the mutated virus (R5X4), according to its characteristics of Molecular Weight, Charge at PH and Iso-Electric Point

Once the classifier was trained with the features previously mentioned (molecular weight, iso-electric point and charge at PH), the next step was to predict the co-receptor of the mutated virus (R5X4), the distribution of both co-receptors is shown in Fig. 6, as we can see in Fig. 6, the mutated viruses are in one of the distinctive groups, the left and right circles are separated according to the type of co-receptor. The circle on the left top represents the amount of mutated viruses that enter exclusively in the co-receptor CCR5 (in this case from the 29 mutated viruses, only 4 have the characteristics that allow them to enter the co-receptor CCR5). The circle on the right bottom represents mutated viruses that have characteristics that allow them to enter to co-receptor CXCR4 (in this case, from the 29 mutated viruses 25 have this type of characteristics).

6 Conclusions

In this article we observe that the viruses R5 HIV-1 use CCR5 as a co-receptor for the viral entrance, the X4 HIV-1 viruses use the CXCR4, while some strange viruses known as R5X4 or D-tropic, have the ability to use both co-receptors. We performed a series of experiments to implement a Bayan Classifier that allows us to asses different patterns that enables us to predict the co-receiver of the mutated virus R5X4.

7 Referencias

[1] Jawetz, Melnick, & Adelberg's Medical Microbiology. Twenty-Sixth Edition, McGraw-Hill Companies, Inc 2013. ISBN 978-0-07-179031-4.
[2] International Comittee on Taxonomy of Viruses (ICTV-2014): http://ictvonline.org/index.asp .
[3] Stanford University School of Medicine. http://www.stanford.edu/group/nolan/tutorials/ret_6_gpedesc.html (USA-2014).
[4] Denis, F., Leonard, G., Sangare, A., et al. Comparison of 10 Enzyme Immunoassays for Detection of Antibody to Human Immunodeficiency Virus Type 2 in West Africa Sera. J Clin Microbiol 1988; 26(5):1000-4.
[5] Viral Count Analysis (Roche Labs 2014). http://www.roche.com/index.htm.
[6] Viral Count Analysis (Biomerieux Labs 2014). http://www.biomerieux.com/
[7] **Analysis of CD4 cells (I**nfonet AIDS 2014). http://aidsinfonet.org/ .
[8] Lara Villegas Humberto H, Ixtepan Liliana del C., Rodríguez Padilla Cristina. El Tropismo y su Identificación. Reporte Técnico, Laboratorio de Bioseguridad Nivel III. Departamento de Inmunología y Virología. Facultad de Ciencias Biológicas. Universidad Autónoma de Nuevo León, México.
[9] Lamers Susanna L., Salemi Marco, McGrath Michael S. and Fogel Gary B. Prediction of R5, X4, and R5X4 HIV-1 Coreceptor Usage with Evolved Neural Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 5. No.2 April-June 2008.
[10] The National Center for Biotechnology Information Advances Science and Health by Providing Access to Biomedical and Genomic Information. Data Base On-Line(NCBI-2014). http://www.ncbi.nlm.nih.gov/.
[11] Weka 3.7 Data Mining Software in Java On_Line(2014). http://www.cs.waikato.ac.nz/ml/weka/
[12] Software for Life Scientists On-Line(2014). http://www.dnastar.com/
[13] Bishop, C.M. Pattern Recognition and Machine Learning; Springer-Verlag: N. York, USA, 2006.
[14] Theodoridis, S.; Koutroumbas, K. Pattern Recognition, 3rd ed.; Academic Press: California, USA, 2006.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Biomechanical Evaluation for Fibular Allograft Combined with Impaction Bone Grafting in Treating the Femoral Head Necrosis: Thorough Debridement or not?

Guangquan Zhou[1,4], Xiumin Chen [2], Zhihui Pang[1], Yujing Xu[3], Wei He[1*], Liao Shaoyi Stephen[4], Qinqun Chen[1], Honglai Zhang[1]

[1]Laboratory of National Key Discipline Orthopaedics and Traumatology of Chinese Medicine & School of medical information engineering, Guangzhou University of Chinese Medicine; The first affiliated hospital, Guangzhou University of Chinese Medicine, Guangzhou, China.

[2]Department of rheumatology, Guangdong Provincial Hospital of Chinese Medicine; China & Postdoctoral Mobile Research Station, Guangzhou university of Chinese Medicine, Guangzhou, China.

[3]University of Science and Technology of China–City University of Hong Kong Joint Advanced Research Center, China

[4]Department of Information systems, City University of Hong Kong, Hong Kong, China

Guangquan Zhou, Xiumin Chen and Zhihui Pang contributed equally to this work.

[*]corresponding Author: Wei He/Tel:(86)13450277305/Email: 491127323@qq.com

Abstract - Fibular allograft combined with impaction bone grafting is an effective hip preserving method that be employed for avoiding the total hip replacement in the early stage of the femoral head necrosis. However, whether thorough debridement should be used with FAIBG is controversial. Thorough debridement can better protect the anterolateral column and reduces the necrosis area of stress concentration phenomenon compared to partial debridement theoretically, but it will bring bigger trauma region and higher incidence of complications, and require longer surgery recovery time. In most cases, the choice is made based on the experience and preference of different surgeons. This study first proposes employing computational biomechanical technology to explore the different mechanical performance of FAIBG with or without thorough debridement, which provides biomechanical basis for choosing the proper treatment in clinic.

Keywords: Computational biomechanics, thorough debridement, fibular allograft combined with impaction bone grafting, anterolateral, stress transfer path

1 Introduction

There is a rapid increase in the incidence of the femoral head necrosis (FHN) all over the world， which is caused by the widespread of steroid [1-2] and alcohol [3-6]. FHN is associated with high morbidity and disability. Patients with FHN often have high risk to have collapse of femoral head, arthritis or dearticulation, and finally result in hip replacement (HR). The normal lifespan of a hip implant is about 12 years, so for the young age of the patients with HR, it is going to be several surgical treatments [7]. Hence, femoral head preserving is a better alternative treatment for patients with FHN in the early stage.

Fibular allograft combined with impaction bone grafting (FAIBG) is an effective head preserving method for avoiding total hip replacement (THR) in the early stage of FHN. FAIBG provides both repaired materials and biomechanical structural support during the healing of the necrosis region [8-11]. However, whether thorough debridement should be used with FAIBG is controversial. "With thorough debridement" means that the necrotic bone should be clean up completely; while "Without thorough debridement" means that the necrotic bone should be partial debridement. Thorough debridement can better protect the anterolateral column and reduces the necrosis area of stress concentration phenomenon compared to partial debridement theoretically, but it will bring bigger trauma region and higher incidence of complications, and require longer surgery recovery time. In most cases, the choice is based on the experience and preference of different surgeons. At the same time, the studies about comparing the risk of collapse of postoperative femoral head accompanied with thorough debridement and without thorough debridement are relatively rare.

To provide scientific biomechanical basis for FAIBG, this study presents two subject-specific FHN cases without collapse of the femoral head to compare the mechanical performance between FAIBG with thorough debridement and without thorough debridement.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2 Materials and Methods.

2.1 JIC classification

In 2001, the Japanese Investigation Committee (JIC) [12] revised diagnostic criteria to clarify the definition of osteonecrosis of the femoral head (ONFH). According to the JIC classification criteria, FHN are classified into subtypes A, B, C1 and C2, based on the location of the lesion in the weight-bearing area. Type A lesions occupy the medial one-third or less of the weight-bearing portion. Type B lesions occupy the medial two-thirds or less of the weight-bearing portion. Type C1 lesions occupy more than the medial two thirds of the weight-bearing portion but don't extend laterally to the acetabular edge. Type C2 lesions occupy more than the medial two-thirds of the weight-bearing portion and extend laterally to the acebtabular edge.

Recent studies have shown that patients who conform to the JIC C criteria are suitable for FAIBG. But these conclusions are mainly based on clinical observation experience and require to be proved in both theory and practice. Whether employing thorough debridement with FAIBG or not is still controversial. Hence, we reconstructed two subject-specific models (including JIC C1 and C2, Figure 1) to provide biomechanical basis for FAIBG to explore the performance of different debridement region in treating FHN.

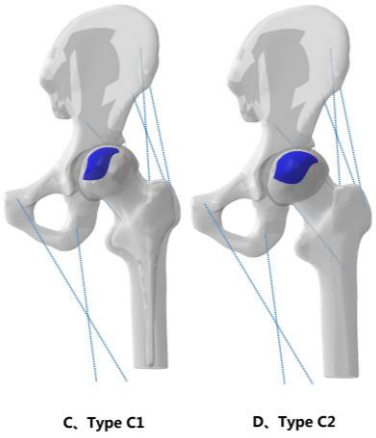


Figure 1 3D subtype models of JIC Classification

2.2 Generation of intact finite element models

A patient (P1) with weight of 70 kg who was diagnosed with JIC C1 FHN and a patient (P2) with weight of 60 kg who was diagnosed with JIC C2 FHN were selected for biomechanical evaluation on the proximal femur. Computed tomography datasets (0.5 mm thickness; Toshiba aquilion 64, Japan) of each case were employed to reconstruct solid models with grey level processing of the software MIMICS 15.1 based on the function of 'Thresholding', 'Edit Masks', and 'Calculate 3D'. The solid models in the format of STL were input to 3-matic pre-processor, in which the surface-fitting could be performed. Based on the function of 'Wrap', 'Patching', we could find the fit hip to generate NURBS models. The interface between the ilium and femoral head was used to identify cartilage geometry. All NURBS models in the format of igs were input to ABAQUS V6.13 (SIMULIA co., France) to generate nonlinear elastic finite element models. Based on the initial hip geometry, we simulated the physiological and pathological models by different materials.

Then, all these models were input to ABAQUS V6.13 to generate isotropic 10-node tetrahedral elements. The mesh size was 4 mm. The initial models were consisted of elements (146879 of P1; 156471 of P2) and nodes (213970 of P1; 230541 of P2). In these models, single-legged stance was considered as a representative body position and a ground reaction force equivalent to body weight was performed on a rigid plate that was tied to the distal part of the femur in Figure 2. Constraints were applied on public symphysis and sacroiliac joint. All the six degrees of freedom were constrained to zero. Seven muscles were modeled as axial connectors, muscle forces were set according to the literature [13]: the adductor longus = 560 N, adductor magnus = 600 N, gluteal maximus = 550 N, gluteal medius = 700 N, gluteal minumus = 300 N, piriformis = 500N and tensorfascia latae = 300 N. The models consist of cortical, trabeculae, cartilage and lesion bone. The material properties used in biomechanical experiment were obtained from the literature [14-16]: $E_{cortical}$=15100 MPa, $E_{trabeculae}$=445 MPa, $E_{cartilage}$= 10.5 MPa, E_{lesion}=124.6 MPa, $\nu_{cortical}$=0.3, $\nu_{trabecular}$=0.22, $\nu_{cartilage}$=0.45 and ν_{lesion}=0.152.

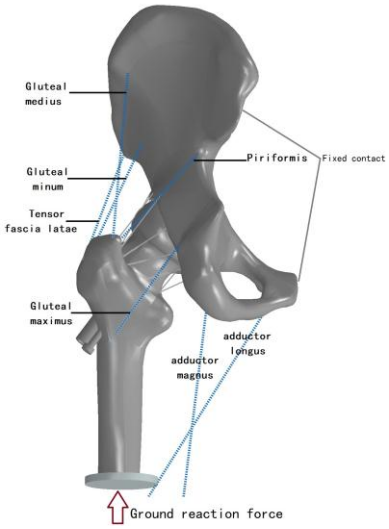


Figure 2 Load and constraint conditions

Parametric analysis was designed to explore the effects of debridement extent of necrotic bone on cases that need surgery. The maximum debridement radius was defined as r and the debridement extent variants were depicted

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

schematically in Figure 3. We assumed the anterolateral cortical stress corresponding to the debridement extent of necrotic lesion had an increased radius R (R=1/4r, 3/8r, 1/2r, 5/8r, 3/4r, 7/8r and r), where R=1/4r referred to the least debridement and R= r denote the thorough debridement. For the simulation of the allogeneic fibular implant, dimensions (80 mm in length and 6 mm in radius) were obtained from the manufacturer. The axial direction of fibula was defined by entry point and the lesion centroid. The entry point was located in the trochanteric lateral cortex of the femur. The distance of the cortical bone from the apex of the fibula was 5 mm. The rest of voids were occupied by impaction cancellous bone after debridement.

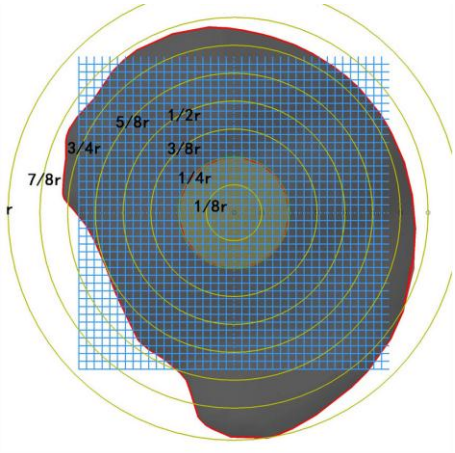


Figure 3 Debridement size of necrotic lesion.

3 Results

3.1 Stress transfer path

The principal stress transfer characteristics are the most important biomechanical index in evaluation of performance of FHN. In all femoral heads, the principal stress transfer patterns are computed when a mid-stance of gait occurs. Figure 4A and 4C show that the principal stress distributions in the healthy conditions are from the top of the femoral head to the femoral calcar. In Figure 4B and 4D, the stress transfer paths are broken off and the areas bearing principal stress are less than approximately 50% of the healthy simulations. The principal stress transfer efficiency reduces obviously.

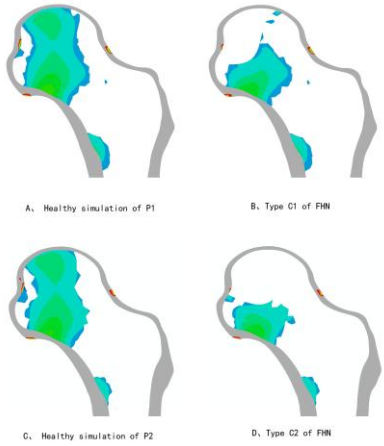


Figure 4 the principal stress distributions in the femoral head

3.2 Stress distribution of anteraolateral column

FAIBG has a considerably small risk of structure collapse compared with untreated situation. Figure 5A and 6A show the healthy stress distribution of anterolateral cortical bone and the maximum stress value are 23.95 MPa of P1 and 25.99 MPa of P2. Figure 5B shows that JIC C1 stress of 30.31 MPa raises about 26.56%, which is higher than the healthy condition (P1). Figure 6B shows that the JIC C2 stress of 34.58 MPa raises about 33.05%, which is higher than the healthy condition (P2). Figure 5C and 6C show that the postoperative stress is 23.52 MPa in P1 and 25.31 MPa in P2, which is approximately 22.4% lower than the JIC C1 condition (P1) and 26.81% lower than the JIC C2 condition (P2) after FAIBG procedure. The peak stresses of the two postoperative cases return to near-healthy levels. It is obviously that the stress concentration regions in JIC C1 and C2 are the areas that the red arrows point to. After FAIBG procedure, the stress concentration regions are disappeared. Figure 5D-I and 6D-I show that stress has no significant changes as the debridement radius increasing.

3.3 The peak stress of the residual necrotic bone

Figure 7 displays that the debridement size will affect the stress gradient in the residual necrotic bone. Seven different necrotic debridement sizes, ranging from 1/4r to r, are chosen to study the effect of debridement radius on the residual necrotic bone. The relation between debridement size and the stress of the residual necrotic bone is shown in Figure 7. When the debridement radius is 1/4r, there is a 3762% increase in the peak stress compared to JIC C1 condition and a 1217% increase in the peak stress compared to JIC C2 condition. When the debridement is not less than 3/8r, the

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

peak stress in the residual lesion is rapidly falling and returns to the physiological level.

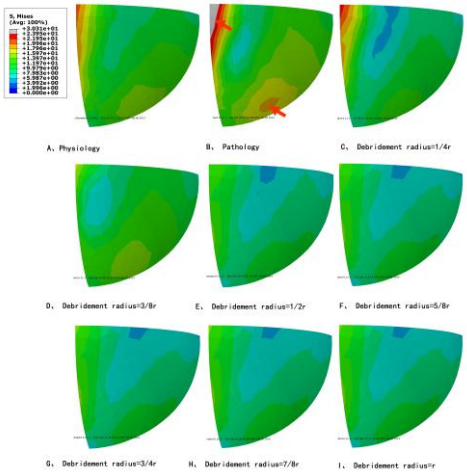


Figure 5 Anterolateral stress distribution of P1

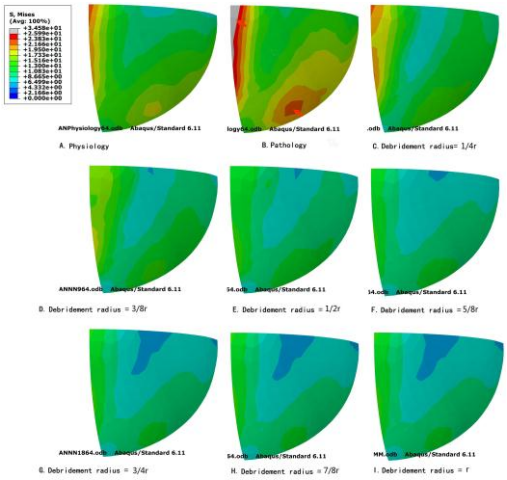


Figure 6 Anterolateral stress distribution of P2

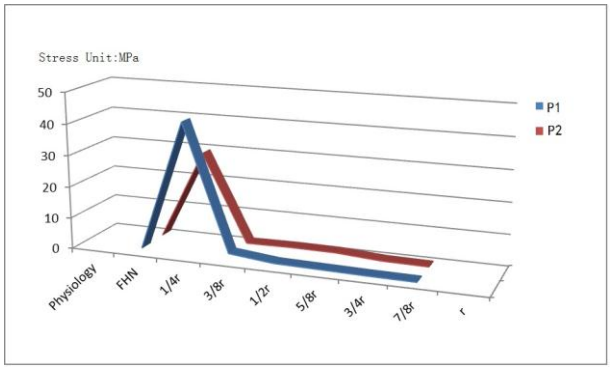


Figure 7 the peak stress of the residual necrotic bone

3.4 Model validation

The principle compressive trabecula loads principle compressive stress of the femoral head (Figure 8C/D), which correlates well with the bone density distribution (Figure 8B) [17]. The shape and location of the biomechanical transfer path for both load cases are consistent with the trabecular features in the cross-sections of the cadaver bone (Figure 8A) [18, 19]. It is clearly that trabeculae in the corresponding areas are thinner. In the same time, there are strong similarities of stress patterns between the simulation results of our study and the previous studies results in the literature [13]. Hence, we think that the FE results could mirror the physical phenomenon of the hip and evaluate the results.

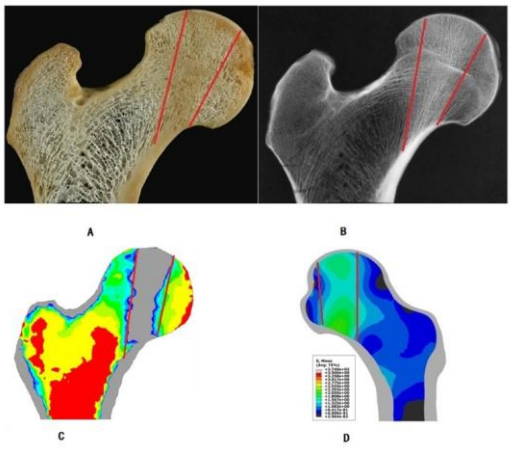


Figure 8 Photograph (A), radiograph (B), the previous simulation results (C) and the computational results (D) of our study. The shape and location of the biomechanical transfer path (D) are consistent with the trabecular features in the cross-sections of the cadaver bone (A), BMD distribution (B) in radiograph and stress distribution in the previous simulation results.

4 Discussion

FAIBG represents a proven technique for maintaining the shape of the femoral head and reducing the risk of collapse of FHN in its early stages. Rosenwasser [20] first described, thorough debridement and bone grafting for treating FHN in 1994. This technique is an effective method for young patients with FHN in early stage, which will delay the progression of osteoarthrosis and subsequent THA. Tao W [21] reported a 80% clinical success rate with a mean follow-up time of 24 months among fifteen patients who had surgical therapy by thorough debridement with bone grafting. However, these procedures may cause serious artificial damage and complication by capsulotomies or the destruction of the cortical bone of the femur neck fundus, and require relatively high surgical devices and technique. In 2008, Shi FL [22] reported of 67 hips, with the treatment of internal bracket implanting with partial debridement for FHN. They came up with a 64.2% (43/67) success rate with an average

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

follow-up of 23 months. In 2013, Shi FL [23] comments their results by treating 25 of 40 patients using allograft fibula with partial debridement for FHN. They reported the satisfactory results in 18/25 (72%) patients with 24 months follow-up. These minimally invasive procedures could reduce the artificial damage and complication but got a poorer clinical outcome, since these procedures couldn't provide both repaired materials and biomechanical structural support during the healing of the necrosis region. FAIBG with proper debridement is an effective head preserving method and we have achieved an average clinical success rate of 90.3 % with a mean follow up time of 37.5 months [24]. All these views are based on the clinical observation experience and lack of biomechanical basis. Hence, both "thorough debridement" and "partial debridement" are not accepted universally because these is no compelling evidence indicate that which method could be better in reducing the collapse risk of femoral head. It encourages us to introduce our experiences on computational biomechanical analysis of debridement extent to reduce collapse risk of FHN.

In our study, we adopted a subject-specific computational approach to consider changes in stress distribution of anterolateral cortical bone and the residual necrotic bone. Figure 4 shows that the stress transfer path in both JIC C1 and C2 are completely broken off, which indicate that surgical intervention should be involved. The effect of debridement size with FAIBG on the collapse risk is clearly demonstrated in Figure 5-6. After FAIBG, the stress of anterolateral cortical bone in all conditions could return to physiological level and the decrement/increment of stress almost less than 0.1 percent as the debridement radius increasing in two cases; hence, the collapse risk of femoral head can be reduced effectively using allo-fibula support to bear the load. When debridement size is not less than 3/8r, the von Mises stress of the residual bone also return to pathological level, which denotes that the progression of necrosis wouldn't deteriorate after surgical intervention. Our results provide specific biomechanical evidence to support the viewpoints that FAIBG can resist the collapse of FHN and FAIBG with thorough debridement has lower risk for developing collapse risk compared to partial debridement.

Thorough debridement has been reported by previous study [20, 21, 25-28]. However, this procedure is difficult and time-consuming, which is associated with serious artificial damage. FAIBG with partial debridement can not only reduce the anterolateral cortical stress but also ensure not increase the stress of the residual bone. This technique has distinct biomechanical basis, and it is time-saving and requires relatively lower surgical devices. It also brings low risk of artificial damage. Hence, FAIBG without thorough debridement seems to be superior compared to FAIBG with thorough debridement.

5 Conclusions

In this paper, we propose employing computational biomechanical technology to explore the different mechanical performance of FAIBG with or without thorough debridement, which provides biomechanical basis for choosing the proper treatment in clinic. Eighteen computational models were constructed and used to simulate two subtypes of FHN with seven debridement radius of FAIBG procedure. The simulation results provide specific biomechanical evidence to support that FAIBG procedure can resist the collapse of FHN. Furthermore, FAIBG without thorough debridement, which not only requires relatively low surgical devices but also reduces artificial damage, seems to be a better method in resisting the collapse of JIC C1 and JIC C2 FHN. This paper is a preliminary approach to investigate FAIBG procedure with thorough debridement, more detailed analysis will be reported in the near future.

6 Declaration of conflicting interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

7 Funding

This study is supported by National Science Foundation of China under Grant (81173284) and Joint Guangdong Science and Technology Department and Guangdong Traditional Chinese Medicine Academy Research and Speciality Fund (2012A032500005) and Natural Science Foundation of Guangdong Province (S2013010011992, S2012010008123). There are no financial and personal relationships with other people or organizations that could inappropriately influence our work for all authors of this paper.

8 Acknowledgments

The authors would like to thank Ms. Yang Xiaolu and Mr.Song Junjun for data analysis support.

9 References

[1] KL Chan, CC Mok. Glucocorticoid-Induced avascular bone necrosis: Diagnosis and management, The open orthopaedics journal, 6,449-57, Oct 2012.

[2] Weinstein RS. Glucocorticoid-induced osteonecrosis, Endocrine, 41, 2,183-190, Apr 2012.

[3] Matuso K, Hirohata T, Sugioka Y, et al. Influence of alcohol intake, cigarette smoking, and occupational status on idiopathie osteonecrosis of the femoral head, Clin Orthopm, 234,115-123, Sep 1988.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[4] Hirota Y, Hirohata T, Fukuda K, et al. Association of alcohol intake, cigarette smoking and occupational status with the risk of idiopathie osteonecrosis of the femoral head, Am J Epidemiol, 137,5,530-8, Mar 1993.

[5] Wang Y, Li Y, Mao K, et al. Alcohol-induced adipogenesis in bone and marrow: a possible mechanism for osteonecrosis, Clin Orthop Relat Res, 410,213-224, May 2003.

[6] Shigemura T, Nakamura J, Kishida S, et al. The incidence of alcohol-associated osteonecrosis of the knee is lower than the incidence of steroid-associated osteonecrosis of the knee: an MRI study, Rheumatology, 51,4,701-706, Apr 2012.

[7] Ditri L., Montanari M., Melamed Y., et al. Femoral Head Necrosis. In D. Mathieu (Ed.), Handbook on Hyperbaric Medicine, (Springer Netherlands), 547-552, 2006.

[8] Brannon JK. Influence of acetabular coverage on hip survival after free vascularized fibular grafting for femoral head osteonecrosis, J Bone Joint Surg Am, 89,448-449, Feb 2007.

[9] Katz MA, Urbaniak JR. Free vascularized fibular grafting of the femoral head for the treatment of osteonecrosis, Techniques in Orthopaedics, 16,44-60,Mar 2001.

[10] Malizos KN, Soucacos PN, Beris AE. Osteonecrosis of the femoral head. Hip salvaging with implantation of a vascularized fibular graft, CLin Orthop Relat Res, 314, 67-75. May 1995.

[11] Urbaniak JR, Coogan PG, Gunneson EB, Nunley JA. Treatment of osteonecrosis of the femoral head with free vascularized fibular grafting. A long-term follow-up study of one hundred and three hips, J Bone Joint Surg Am, 77,681-694, 1995.

[12] Sugano N, Atsumi T, Ohzono K,et al. The 2001 revised criteria for diagnosis, classification, and stagine of idiopathic osteonecrosis of the femoral head, J Orthop Sci, 7,601-605, 2002.

[13] Nina S.Sverdlova,Ulrich Witzel, Principles of determination and verification of muscle forces in the human musculoskeletal system: Muscle force to minimise bending stress, Journal of Biomechanics, 43,387-396, Feb 2010.

[14] Brown TD, Hild GL, Pre-collapse stress redistributions in femoral head osteonecrosis--a three-dimensional finite element analysis, J Biomech Eng, 105,171-176, May 1983.

[15] Brown TD, Way ME, Ferguson AB Jr. Mechanical characteristics of bone in femoral capital aseptic necrosis, Clin Orthop Relat Res, 156, 240-247, May 1981.

[16] Grecu D, Pucalev I, Negru M, Tarnita DN, Ionovici N, Dita R. Numerical simulations of the 3D virtual model of the human hip joint, using finite element method, Rom J Morphol Embryol, 51,151-155, 2010.

[17] Jang In Gwun, Kim Il Yong. Computational study of Wolff's law with trabecular architecture in the human proximal femur using topology optimization, Journal of Biomechanics, 41:2353-1361, Aug 2008.

[18] Christopher Boyle, Il Yong Kim. Three-dimensional micro-level computational study of Wolff's law via trabecular bone remodeling in the human proximal femur using design space topology optimization, Journal of Biomechanics, 44,935-942, Mar 2011.

[19] Jang In Gwun, Kim Il Yong. Computation simulation of trabecular adaptation progress in human proximal femur during growth, Journal of Biomechanics, 42,573-580, Mar 2009.

[20] Rosenwasser MP, Garino JP, Kiernan HA, Michelsen CB. Long-term follow up of thorough debridement and cancellous bone grafting for osteonecrosis of the femoral head, Clin Orthop,306,17-27, Sep 1994.

[21] Tao W., Wei W., Zong SY. Treatment of osteonecrosis of the femoral head with thorough debridement, bone grafting and bone-marrow mononuclear cells implantation, Eur J Orthop Surg Traumatol, 24,197-202, Feb 2014.

[22] Shi FL, Lu FX, Li XH, et al. Clinical observation on internal bracket implanting for treatment of adult necrosis of femoral head and finite element analysis, Chinese journal of bone and joint injury, 23,3,186-188, Mar 2008.

[23] Shi FL, Chen J, Li XH, et al. Fan-shaped decompression and allograft fibula supporting internal fixation for treatment of early femoral head necrosis in adults, Chinese journal of tissue engineering research, 17,44, 7758-7763, Oct 2013.

[24] He W, Li Yong, Zhang QW, et al. Primary outcome of impacting bone graft and fibular autograft or allograft in treating osteonecrosis of femoral head, Chinses Journal of Reparative and Reconstructive Surgery, 23, 5,530-533, Nov 2009.

[25] Meyers MH, Jones RE, Bucholz RW, et al. Fresh autogenous grafts and ostaochondral altografts for the treatment of segmental collapse in osteonecrosis of the hip, Clin Orthop, 174, 10-12, Apr 1983.

[26] Ko JY, Meyers MH, Wanger DR. "Trapdoors" Procedure for osteonecrosis with segmental collapse of the femoral head in teenagers, J Padiatt Orthop, 15, 7-15, Jan-Feb 1995.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[27] Meyers MH, Convery FR. Grafting procedure in osteonecrosis of the hip, Sem in Arthroplasty, 2,189-197, 1991.

[28] Gardeniers J, Yanmano K, Buma P, Sloff. Impaction grafting in the femoral head, proceedings of the second afor symposium on osteonecrosis of the femoral head and hip around fracture, 28-30, 1999.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Keratoconus Disease and Three-Dimensional Simulation of the Cornea throughout the Process of Cross-Linking Treatment

Kaya H.[1], Çavuşoğlu A.[2], Çakmak HB.[3], Şen B.[4], Çalık E.[5]
[1] *Ministry of National Education, Department of Information Technologies, Ankara, Turkey*
[2] *The Scientific and Technological Research Council of Turkey, Ankara, Turkey*
[3] *Yıldırım Beyazıt University, Atatürk Education and Research Hospital, Ankara, Turkey*
[4] *Yıldırım Beyazıt University, Department of Computer Engineering, Ankara, Turkey*
[5] *Karabük University, School of Health Sciences, Karabük, Turkey*
hilalkaya@meb.gov.tr,abdullah.cavusoglu@tubitak.gov.tr, hbcakmak@ybu.edu.tr, bsen@ybu.edu.tr, elifcalik@karabuk.edu.tr

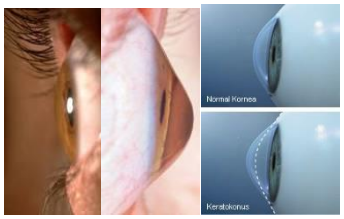
***Abstract*—Keratoconus is the corneal disease that comes out by the progressive thinning and tapering of the cornea. Vision gradually decreases as the sphere-shaped cornea becomes more tapered and conical. With Corneal Cross-Linking treatment, as increasing the number of cross-links that are existing in the connective tissues of corneal layer, cornea hardens and becomes more resistant.**

Purpose of this study is monitoring the changes in the cornea between the processes before and after the treatment by three-dimensional simulation techniques in case of Cross-Linking Treatment is preferred, after creating a dataset by preparing two-dimensional cornea images with data mining methods. With this application, it can be possible to follow-up the healing process after the treatment and also monitor whether the treatment has achieved the desired results or not. This system is intended to be developed in order to support eye specialists on disease diagnosis, treatment and follow-upstages. By the Ethics Committee approval report, dated 15th April 2013 and numbered 43, 749 digital image data was provided and this study was carried out. In this study, it's seen that follow-up process of the disease by analyzing two-dimensional cornea images can be improved by using three-dimensional images.

***Index Terms*—Cross-linking, Keratoconus, Medical diagnostic imaging, Medical simulation.**

I. Introduction

Keratoconus can be defined as the forward extension of the cornea (the transparent, breaker layer such a watch glass in front of the eye) as tapering conically (Fig. 1). Disease is more common among the women. Changing the refractive power of the cornea, it causes moderate or severe degree of irregular astigmatism and blurred vision. In the final stages of Keratoconus corneal swelling and blanching can be seen [1].



(a) (b) (c)

Fig. 1. a)Healthy Eye [2], b) Eye with Keratoconus [2], c) Comparison of Healthy Cornea and Cornea with Keratoconus [3]

Light enters the eye through the cornea and because it provides clear vision by breaking or focusing the rays, cornea is a very important part of the eye. In Keratoconus, that the shape of the cornea would be changed and the vision would be distorted, problems may arise in some activities such as using car or computer, watching television and reading etc. Especially seen in young people, it may lead to some negativity in their education and employment lives and in case of not being treated, it causes serious vision problems. Therefore, early diagnosis and selecting the appropriate treatment is of great importance. If the patient is in the stated age group (15-35), Keratoconus is progressive, patient does not feel comfortable longer with the lenses and the cornea is thicker from 400 micrometers, it's suitable to apply to the Cross-Linking methods. Purpose of this treatment method is to stop the advancing Keratoconus, to improve the vision quality by reducing the refractive defect and to eliminate the need for corneal transplantation (Keratoplasty). According to various studies that have been conducted, if the Cross-Linking treatment is preferred, the progression for the disease can be stopped in the rate of 90-98%.

Due to the prevalence map of the disease over the world, our country is located in the region where the disease is most prevalent (Fig. 2). Depending on the high rate of the young population in our country and that existing in the sunny and pollen-rich climate zone, disease can be seen from 2000 to 2500 per person. Due to this high rate, a system is needed to facilitate making decisions at the point of diagnosis and treatment steps.

[1]*Ministry of National Education, Department of Information Technologies, hilalkaya@meb.gov.tr*

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Fig. 2. Prevalenceof the Disease in the Region [4]

Fig. 3. Scheimpflug Camera and Placido Disc Combination [13]

II. KERATOCONUS

The disease has a genetic trait and it can be seen as high astigmatism or Keratoconus in high probability. In patients, irregular astigmatism or myopia is constantly increasing and bilateral involvement would be seen in time. The majority of patients complain of frequent changes in glasses, but soon after, these glasses will be inadequate and visual impairment will continue. Keratoconus can be associated with corneal injury, some specific eye diseases and systemic diseases. After making excimer laser surgery to an unsuitable eye, due to the weakening of the vitreous of the eye, it may be occurred [1]. In all cases that Keratoconus is suspected, doing corneal topography before diagnosis is of great importance [5].

There is not a precise classification method in Keratoconus which everyone compromise. So far, various classification methods are used considering such parameters as conus morphology, clinical findings, visual acuity, disease progression, keratometry, topography derived values, corneal aberrations alone or their particular combinations. First classification based on the progression of the disease was made by Amsler and then similar reclassifications have been made [6, 7]. Amsler evaluated Keratoconus in 4 stages. For diagnosis and classification of Keratoconus, various classification methods obtained from corneal topography systems have been formed. Rabinowitz and Mc Donnell, Rabinowitz and Rasheed, Mc Mahon et al. and Mahmoud et al. made classification using corneal topography results [8-11].

Recently, possibility of the diagnosis using anterior segment parameters thanks to the devices using Scheimpflug camera system was defended (Fig. 3). Scheimpflug camera system is a next generation system that can record three-dimensional images as making rotation to the axis of the eye with its rotating camera. In this system, besides topography maps, Keratoconus can be diagnosed and the severity of the disease can be evaluated by using parameters as corneal volume, anterior chamber angle, anterior chamber volume and anterior chamber depth [12]. These devices can record three-dimensional images but offers in two-dimensional form to the user. Images that made more understandable by modeling in three-dimensional form in our study are taken from Scheimpflug camera system.

A. Cross-Linking Treatment

Importance of correctly identifying and classifying Keratoconus is increasing in time because nowadays variety of treatment options is developed and these options are very effective on the treatment. Hereafter treatments such as Collagen Cross-Linking used in progressive Keratoconus cases can make possible to stop the progression of the disease [5].

Collagen Cross-Linking method is a kind of treatment that is used for Keratoconus disease in last years and is applied to the corneas thicker than 400 micro-. After treatment, thin cornea hardens and becomes more resistant. In this way, sharpness of the cornea and the disease progression can be halted/very decelerated. Among half of the patients, cornea is flattened in an amount of approximately 2-3 sizes [14].

Raiskup-Wolf et al. followed the results of Cross-Linking treatment on Keratoconus cases in 6 years period. In their paper of this study, at the end of 3rd year, a decrease of 4.84D in corneal slope and an accompanying increase in best corrected visual acuity was reported to be worth [15]. Various studies were conducted on the subjects of how to use the treatment, current status and monitoring the results. Utine et al., Gündüz, Utine, Raiskup and Spoerl, Zhang, Caporossi et al., Tahzib et al. shared the effects of the treatment on the disease and the postoperative corneal changes [16-22].

B. Three-Dimensional Imaging in Medicine

Using the computerized devices benefitting from 3-D technologies provide decision support to the field experts on making right decisions on the diagnosis in a short time, determining most effective and result-oriented treatment, realizing and monitoring the treatment and operations.

With the help of 3-D devices especially used in cancer treatment, as clearly determining the tumor localization, healthy cells are prevented from damaging. As an example for this type of treatment, "Varian Linear Accelerator" named 3-D imaging device, started to be used at Gazi University Faculty of Medicine in order to support target oriented radiation treatment, brought many contributions to the patient comfort and the treatment. In this treatment method, with the computerized planning device that collects

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

the tumor-related information as the location, size, precision as making computerized simulation, it can be possible to mark the tumor and surrounding delicate tissue and determine the treatment doses [23]. Displaying the face in 3-D form is recognized as one of the advances in physical modeling techniques using the engineering methods in medicine. Kumar and Vijai performed 3-D modeling of the face in a different 3-D imaging approach [24].

In this study, providing decision support to the doctor is intended on the course of the Keratoconus disease as displaying the cornea in 3-D form. Therefore, visibility of the corneal region with Keratoconus will be increased.

III. METHODOLOGY

A. Data

Study was conducted on 749 digital image data of totally 122 patients recorded by Scheimpflug Camera and Placido Disc Combination between the dates 24th January 2009 and 24th January 2012. Data was provided from Yıldırım Beyazıt University Atatürk Education and Research Hospital by the Ethics Committee approval report dated 15th April 2013 and numbered 43.

B. Application

Firstly, the dataset that will be used in the study was obtained through 749 images by using data mining steps from the 2-D images of Scheimpflug Camera and Placido Disc Combination. After this step, 144 images had to be neglected because of the scanning problems especially closed eyelids and the final data set has 605 original 2-D images. Images were grouped using Multilayer Perceptron and Logistic Regression methods using the thickness values that were obtained from our application reading from 2-D images. With the help of 3-D imaging application developed using the grouped 2-D image data obtained in the previous step; more easily interpretable 3-D maps were obtained.

Steps of our study are detailed in Fig. 4:

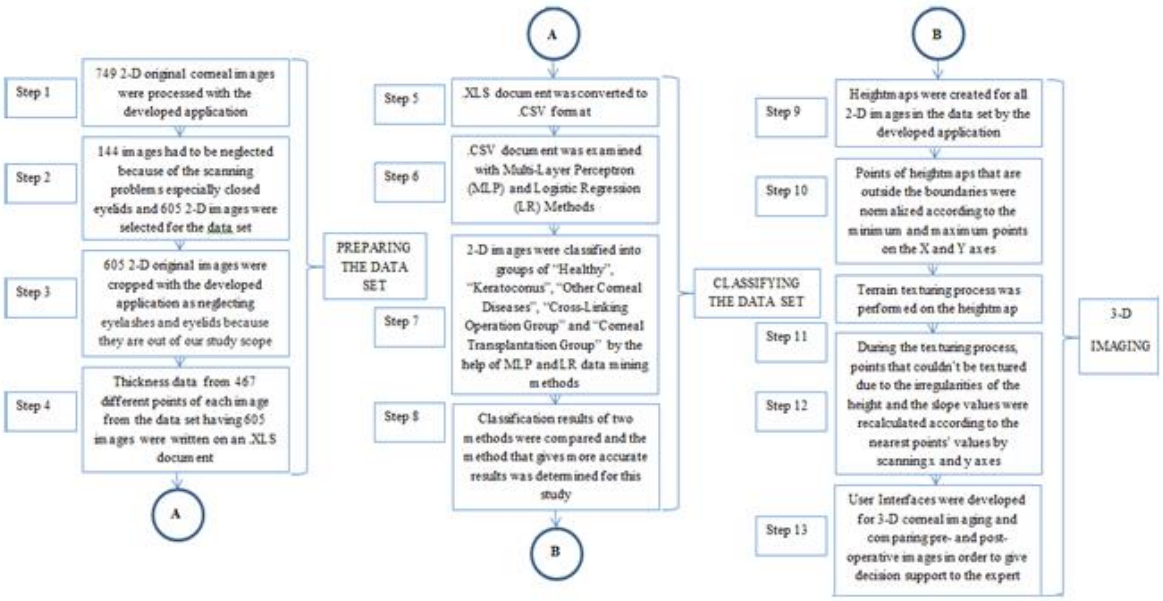


Fig. 4. Application Steps of the Study

Application architecture of this study is also seen in Fig. 5:

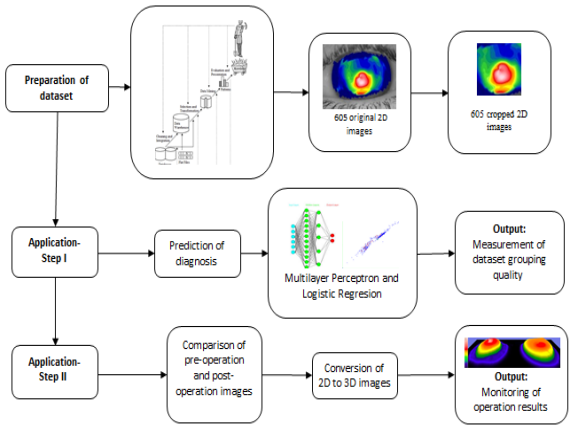


Fig. 5. Application Architecture

Multilayer Perceptron is used in this study as one step of data mining process. A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one (Fig. 6). Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network [25]. Mathematical expression of each perceptron's computation can be expressed as (1)

$$y = \varphi\left(\sum_{i=1}^{n} \omega_i x_i + b\right) = \varphi(\mathbf{w}^T\mathbf{x} + b) \qquad (1)$$

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

where **w** is the vector of weights, **x** is the vector of inputs, $\boldsymbol{b}$ is the bias and $\boldsymbol{\varphi}$ is the activation function.

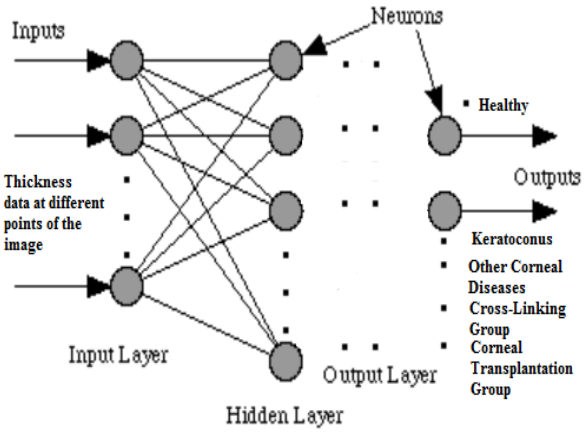


Fig. 6. Structure of a Multilayer Perceptron

MLP is a feed forward network of interconnected neurons (Fig. 6) usually trained using the error backpropogation algorithm. This popular algorithm works by iteratively changing a network's interconnecting weights (in proportion to a 'training rate' set by the Artificial Neural Network (ANN) engineer) such that the overall error (i.e., between observed values and modeled network outputs) is reduced [26].

The other step of our study's data mining part is Logistic Regression. Purpose of using Logistic regression (LR) is to establish a model using least variable to be optimum fitting that can define relationship between dependent and independent variables and that is biologically acceptable [27]. The logistic regression model can be expressed as (2)

$$\ln[p/(1-p)] = \boldsymbol{a} + \boldsymbol{B}\mathrm{X} + \mathrm{e} \qquad (2)$$

where **ln** is the natural logarithm, $\log_{exp}$, where exp=2.71828..; **p** is the probability that the event Y occurs, p(Y=1); **p/(1-p)** is the "odds ratio"; $\mathbf{ln}[\boldsymbol{p}/(\mathbf{1}-\boldsymbol{p})]$ is the log odds ratio, or "logit"; $\boldsymbol{a}$ is the coefficient on the constant term; $\boldsymbol{B}$ is the coefficient(s) on the independent variable(s); **X** is the independent variable(s) and **e** is the error term.

In our study, thickness data obtained from 467 points of 605 2-D imaging data each was analyzed with unsupervised MLP and LR methods. 70% of this data is used for training the system and 30% of the data is used for testing and deciding which classification group that the image belongs to.

In order to use three-dimensional modeling techniques in the application developing process, XNA Framework DLLs were added on Microsoft. NET C# platform. Microsoft XNA Framework is a tool that enables developing games for software developers using Visual Studio C# language on Windows and Xbox 360 platforms. Standard game development procedures require a lot of code and time; XNA Framework is intended to facilitate this process. To realize this idea, it's presented that the most important thing the programmers should take care of is the code. XNA Framework takes the items on itself that processing and designing period takes time as graphics card, resolution, image processing. Also creates a game window for developers and provides shaping in the situations as changing window resolution and window resizing [28]. With the help of these DLLs, software codes were developed that supports the transactions as follows;

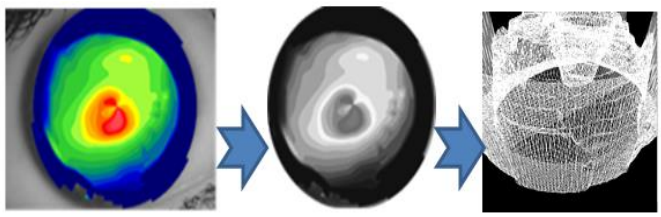
- Creating a height map (terrain) from the RGB or grayscale corneal image,
- Cleaning the image parts outside the normal range (normalize the image according to the minimum and the maximum points),
- Terrain texturing,
- During the texturing process, if some points cannot be textured due to the irregularities of the height and the slope values, recalculating these values according to the nearest points' values by scanning x and y axes respectively.

Stages of the normalization process involve finding the points that have minimum and maximum height values respectively for X and Y axes and displaying after recalculating all points according to these values (3). Steps of normalization formula [29];

$$\begin{aligned} I&: \{X \subseteq \mathbb{R}^n\} \rightarrow \{\mathrm{Min}, .., \mathrm{Max}\} \\ I_N&: \{X \subseteq \mathbb{R}^n\} \rightarrow \{\mathrm{newMin}, .., \mathrm{newMax}\} \end{aligned} \qquad (3)$$

$$I_N = (I - Min)\frac{\mathrm{newMax} - \mathrm{newMin}}{\mathrm{Max} - \mathrm{Min}} + \mathrm{newMin}$$

After normalization and texturing processes on the height map seen in Fig.7-c, three-dimensional corneal image in Fig.8 was obtained.



(a) (b) (c)

Fig. 7. a) RGB Image, b) Grayscale Image, c) Height Map

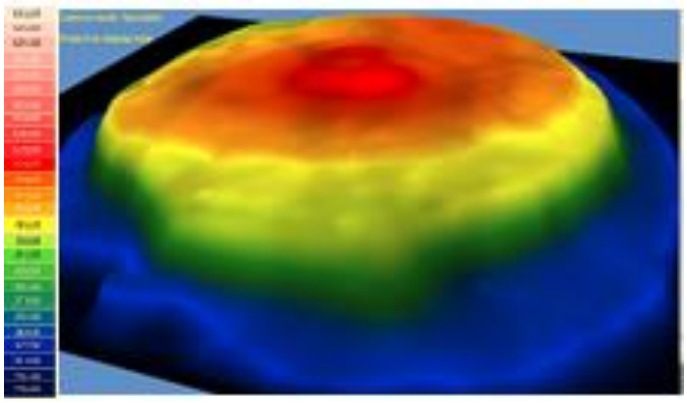
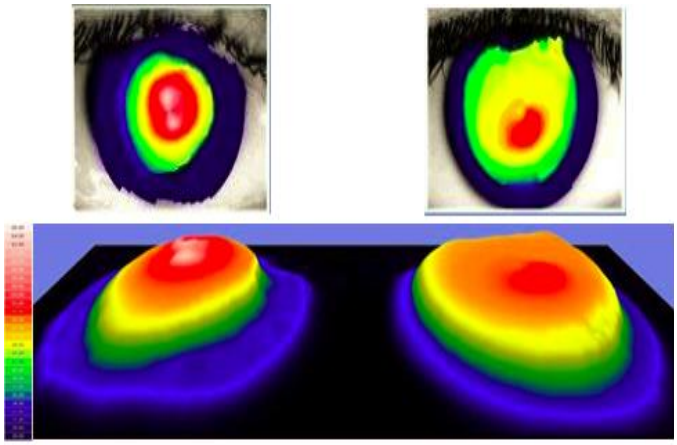


Fig. 8. Textured Three-Dimensional View

In our study, it's aimed to provide monitoring and evaluation processes can be made by the experts even not experienced on Keratoconus, not overlooking any detail in the diagnosis process of the disease with the help of easily readable and interpretable maps. Thanks to the system that was developed, recorded images for pre-operation and post-operation in 2-D form were transformed to 3-D images (Fig. 9).

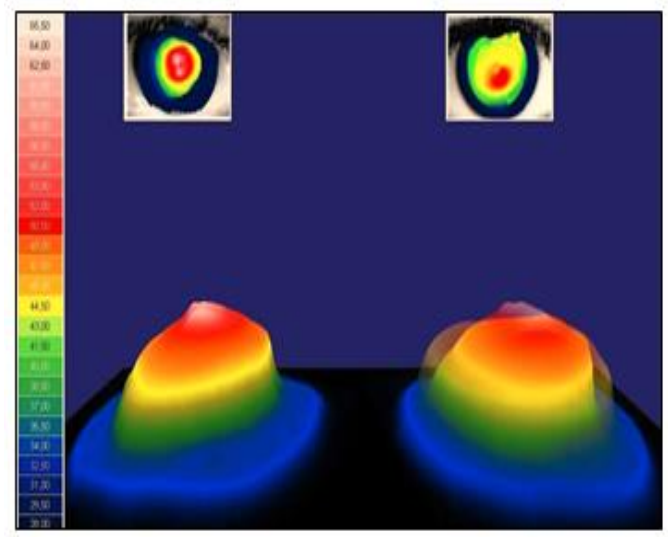
Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America



(a) (b)

Fig. 9. a) 2-D and 3-D Pre-operation Images, b) 2-D and 3-D Post-operation Images

In the next step, to facilitate monitoring the effects of the treatment, images were displayed overlapping (Fig. 10). In the Fig. 10-b, it can be seen that when the cross-links that are existing in corneal layer hardened and became more resistant after the treatment, prolonged tissues were withdrawn.



(a) (b)

Fig. 10. a) Pre-operation images, b) Comparison of the Overlapping Pre-operation and Post-operation Images

IV. Conclusions and Recommendations

Early diagnosis and suitable treatment planning at the right time is extremely important for success in treating Keratoconus disease. Because the disease can be treated if diagnosed early, need for corneal transplantation can be prevented or slowed down for many patients by the help of early diagnosis. Along with diagnosing 11-12 year-old-children with amblyopia that have astigmatism in one eye and that their vision can not be increased in their past life, often times Keratoconus subsequently arises in many of these patients. Because that the disease cannot be identified by routine examination in early stages, special topographic devices are needed for the diagnosis of the disease. To notice the tapering of the patient's cornea with the naked eye can be available in advanced stages of the disease which require corneal transplantation and large number of patients live without being aware of their disease because diagnosing the disease needs special tests.

In this study, in case of Cross-Linking treatment is preferred, it's aimed to simulate the changes occurred in the cornea by using 3-D modeling techniques between pre-operation and post-operation periods as selecting from the cornea images of good quality in recording by using data mining methods. With the help of the application developed, it is possible to monitor the treatment if it's successful in achieving the desired results as well as monitoring the healing process in the post-treatment period. It's intended to support the eye specialists in the diagnosis, treatment and follow-up phases by the established system.

In our study, 605 images were selected from 749 images by data cleaning steps in data mining process. Thickness data obtained from 467 different points of 605 images each were analyzed with MLP and LR methods. 424 rows (70%) of data were used to train the system and 181 (30%) rows of data were used to test and decide the classification groups and members.

TABLE I. Results of Classification Algorithms

Algorithm	Multilayer Perceptron	Logistic Regression
The number of input variables	**Total number of instances:**605 **Number of training data:** 424 (70% of total 605 instances) (This data is not used for classification, only for training the system) **Number of test and classification data:** 181 (30% of total 605 instances) (This data is used for classification of the images)	**Total number of instances:**605 **Number of training data:** 424 (70% of total 605 instances) (This data is not used for classification, only for training the system) **Number of test and classification data:** 181 (30% of total 605 instances) (This data is used for classification of the images)
The number of correctly classified instances	**176 instances** of total number of 181 test and classification instances **97.2376 %** of 181 test and classification instances	**159 instances** of total number of 181 test and classification instances **87.8453%** of 181 test and classification instances
The number of incorrectly classified instances	**5 instances** of total number of 181 test and classification instances **2.7624%**	**22 instances** of total number of 181 test and classification instances **12.1547%**
Classification time	1011.16 s	6.7 s
True classification	Very Good	Good

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Results of the classification methods used in the study are shown on Table 1. Multilayer Perceptron with the accuracy rate of 97.2376 % is more successful than Logistic Regression with the accuracy rate of 87.8453%. Accuracy rates of the data mining methods can be calculated as using the equation (4).

$$\textbf{\textit{Accuracy rate of the method}}\ (\%) = (\textbf{\textit{Number of the instances that are classified in the accurate group/Number of all instances}}) * \mathbf{100} \quad (4)$$

2-D data set grouped with MLP (because it has produced more correctly classified instances) was modeled in 3-D form on the .NET C# platform with the help of XNA Framework DLLs. With the help of 3-D images, readability of the cornea was increased and we had the chance to compare the classification results acquired with data mining methods and 3-D imaging processes with each other. This study showed that displaying of the disease and the healing process can be improved by using more interpretable 3-D images.

In our future project, we plan to segment the affected corneal region of actual two-dimensional corneal images by the help of image processing and image analysis methods of image registration and image segmentation. Then original images will be viewed in 3-D format and also segmented damaged parts will be viewed in 3-D form. These 3-D images can also be compared and the post-operation form of the cornea can be predicted before the operation by the help of these comparison interfaces. With this study, we plan to provide decision support to the field experts on deciding the right treatment and estimating the recovery rate of the disease prior to the treatment. Thus, compliance status of the operation with the patient's cornea will be known before the operation. These two studies can be determined as innovations in this study field according to the literature reviews. Because there is a lack in the existing studies on making predictions and decision support about corneal diseases and the results of the operations using 3-D modeling of the cornea. These studies will set light to the future studies on this field.

ACKNOWLEDGEMENT

This study is supported by Republic of Turkey-Ministry of Science, Industry and Technology in the scope of SAN-TEZ Project numbered 0477.STZ.2013-2. Also we would like to thank to Yıldırım Beyazıt University Ataturk Education and Research Hospital board for their permission to use the digital image data with Ethics Committee approval.

REFERENCES

[1] Internet:http://www.igh.com.tr/keratokonus (Accessing Date:04.03.2013)

[2] Internet: http://www.omerfarukyilmaz.com/keratokonuslu-goz.html (Accessing Date: 30.01.2013)

[3] Internet: http://www.dunyagoz.com/bultenler/bulten2/ (Accessing Date: 30.01.2013)

[4] Internet: http://www.keratokonus.com.tr/keratokonus_nedir.html (Accessing Date:20.06.2013)

[5] Kocamış S.İ. Keratokonus tanısında CLMI'nın etkinliğinin güncel ölçümlerle karşılaştırılarak değerlendirilmesi, Uzmanlık Tezi, Ankara Atatürk Eğitim ve Araştırma Hastanesi 1.Göz Kliniği (Tez Danışmanı: Doç.Dr. Hasan Basri ÇAKMAK), Ankara, 2011.

[6] Amsler M. Le keratocone fruste au javal. Ophtalmologica 1938; 96: 77-83.

[7] Hom M., Bruce A.S. Manual of contact lens prescribing and fitting. London: Butterworth-Heinemann; 2006, 503-544.

[8] Rabinowitz Y.S., Mc Donnell P.J. Computer-assisted corneal topography in keratoconus. Refract Corneal Surg. 1989; 5: 400-408.

[9] Rabinowitz Y.S., Rasheed K. KISA% index: a quantitative videokeratography algorithm embodying minimal topographic citeria for diagnosing keratoconus. J. Cataract Refract Surg. 1999; 25:1327-1335.

[10] MCMahon T.T., Szczotka-Flynn L., Barr J.T., Anderson R.J., Slaughter M.E., Lass J.H., Iyengar S.K; CLEK Study Group. A new method for grading the severity of keratoconus: the Keratoconus Severity Score (KSS). Cornes 2006; 25(7): 794-800.

[11] Mahmoud A.M., Roberts C.J., Lembach R.G:, Twa M.D., Herderick E.E, McMahon T.T. CLMI: the cone location and magnitude index. Cornea 2008; 27: 480-487.

[12] Emre S., Doğanay S., Yoloğlu S. Evaluation of anterior segment parameters in keratoconic eyes measured with Pentacam system. J. Cataract Refract Surg. 2007; 33: 1708-1712.

[13] Internet:http://www.vsy.com.tr/Urunler/Refraktif-Cerrahi/Korneal-Wavefront/Amaris-Corneal-Wavefront-Analyser/Korneal-Wavefront-Analizi/Schwind-Sirius.aspx (Accessing Date: 30.01.2013)

[14] Internet:http://www.banucosar.net/ic_sayfa.aspx?id=137 (Accessing Date: 04.03.2013)

[15] Raiskup-Wolf F., Hoyer A., Spoerl E., Pillunat L.E. Collagen cross-linking with riboflavin and ultraviolet-A light in keratoconus: Long-term results. J.Cataract Refract Surg. 2008; 34: 796-801.

[16] Utine C.A., Çakır H., Altunsoy M. Korneanın ektatik hastalıklarının tedavisinde kollajen çapraz bağlama", T. Oft. Gaz. (39) 2009, 153-160,

[17] Uçakhan Gündüz Ö.Ö. Keratokonusta alternatif tedavi yöntemleri: İntrastromal halka segmentler ve kollajen çapraz bağlama: Güncel durum", MN Oftalmoloji 2009, 3: 39-44.

[18] Utine C.A. Hafif ve orta derecedeki keratokonusun tedavisinde radyal keratotomi uygulaması, Uzmanlık Tezi, İstanbul, 2005.

[19] Raiskup F., Spoerl E. Corneal cross-linking with hypo-osmolar riboflavin solution in thin keratoconic corneas", American Journal of Ophthalmology 2011, 152(1): 28-32.

[20] Zhang Z.Y. Corneal thickness change in eyes undergoing corneal cross-linking", American Journal of Ophthalmology 2012, 153(2): 383.

[21] Caporossi A., Mazzotta C., Baiocchi S., Caporossi T. Long-term results of riboflavin ultraviolet A corneal collagen cross-linking for keratoconus in Italy: The Siena Eye Cross Study, American Journal of Ophthalmology 2010, 149(4): 585-593.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[22] Tahzib N.G., Soeters N., Lelij A.V. Pachymetry during cross-linking, Ophthalmology 2010, 117(10): 2041-2041.
[23] Internet:http://www.sagliktagundem.com/haber/kansere_3_boyutlu_tedavi.htm (Accessing Date: 27.01.2013)
[24] Kumar T.S., Vijai A. 3D reconstruction of face from 2D CT scan images, Procedia Engineering 2012, 30, 970-977.
[25] Internet:http://en.wikipedia.org/wiki/Multilayer_perceptron (Accessing Date: 25.01.2014)
[26] Dawson C.W., Wilby R.L., Harpham C., Brown M.R., Cranston E., Darby E.J. Modelling ranunculus presence in the rivers test and itchen using artificial neural networks, GeoComputation 2000.
[27] Bircan, H. Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama", Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2004(2): 186,189,197(2004).
[28] Internet: http://www.sdtslmn.com/kodmod/xna-framework-1-xna-nedir/ (Accessing Date: 25.01.2014)
[29] Internet: http://en.wikipedia.org/wiki/Normalization(image_processing) (Accessing Date: 25.01.2014)

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

A Bayesian Network Classifier of Some Triatomine Vectors of Chagas Disease

Jack K. Horner
PO Box 266
Los Alamos, New Mexico 87544 USA
jhorner@cybermesa.com

BIOCOMP 2014

Abstract

Chagas disease (CD) disease is a life-threatening tropical parasitic disease caused by the flagellate protozoan Trypanosoma cruzi. T. cruzi is typically transmitted to humans and other mammals by the bite of "kissing bugs" of the subfamily Triatominae (family Reduviidae), primarily by species belonging to the Triatoma, Rhodnius, and Panstrongylus genera. Rapidly identifying CD insect vectors in the field is crucial to effective control of the disease. Here I describe a Bayesian network triatomine classifier that supports rapid identification of adults of nine CD vector species.

Keywords: Chagas Disease, automated classification, Bayesian network

1.0 Introduction

Chagas disease (CD) disease is a life-threatening tropical parasitic disease caused by the flagellate protozoan *Trypanosoma cruzi*. *T. cruzi* is typically transmitted to humans and other mammals by the bite of "kissing bugs" of the subfamily *Triatominae* (family *Reduviidae*), primarily by species belonging to the *Triatoma*, *Rhodnius*, and *Panstrongylus* genera ([5]).

As many as 11 million people in Mexico, Central America, and South America have CD. Most of those infected do not know that they are. Large-scale population movements from rural to urban areas of Latin America and to other regions of the world have increased the geographic distribution of CD; the disease has been reported in several European countries ([5]).

Rapid field identification of CD vector species is essential to effective vector control. A tool that that could capture an image of a specimen and return its taxonomic identification would help to minimize the labor requirements for rapid field identification of the vectors.

Such a tool can be thought of as having two main subfunctions -- an *image-processing* function, which performs various image-analysis tasks and emits image information (e.g., image histograms, maximum and minimum pixel intensity values, image differences, morphometrics, etc.; [12]), and an *automated classification* function, which accepts the output of the image processing and infers the taxonomic classification of the specimen.

2.0 Method

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Images of adult specimens of nine triatomine species, (S), nominally regarded as the most common CD vector species in Brazil ([2]), were selected for classification support:

(S)

- *Triatoma infestans*
- *Triatoma dimidiata*
- *Triatoma brasiliensis*
- *Triatoma maculata*
- *Triatoma sordida*
- *Rhodnius prolixus*
- *Rhodnius neglectus*
- *Rhodnius pallescens*
- *Panstrongylus megistus*

Five morphological measurement-ratios of adult triatomines (requiring a maximum of nine length measurements), (D), were selected from [1] for classification support. (These properties were selected because a preliminary assessment appeared to be adequate to discriminate among the members of (S). Testing subsequently confirmed that the first four members of (D) are collectively sufficient to discriminate among all members of (S), but also showed that *Triatoma brasiliensis* was not as well resolved as the rest of (S), thus motivating the addition of the fifth discriminator, First/Third Antenna-Segment Length Ratio):

(D)

- Head Length/Width Ratio (all classifiers)
- Head/Pronotum Length Ratio (all classifiers)
- Ante-/Post-Ocular Length Ratio (all classifiers)
- First/Third Rostral Segment Length Ratio (all classifiers)
- First/Third Antenna-Segment Length Ratio

The morphological features required by these metrics are illustrated in Figure 1.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

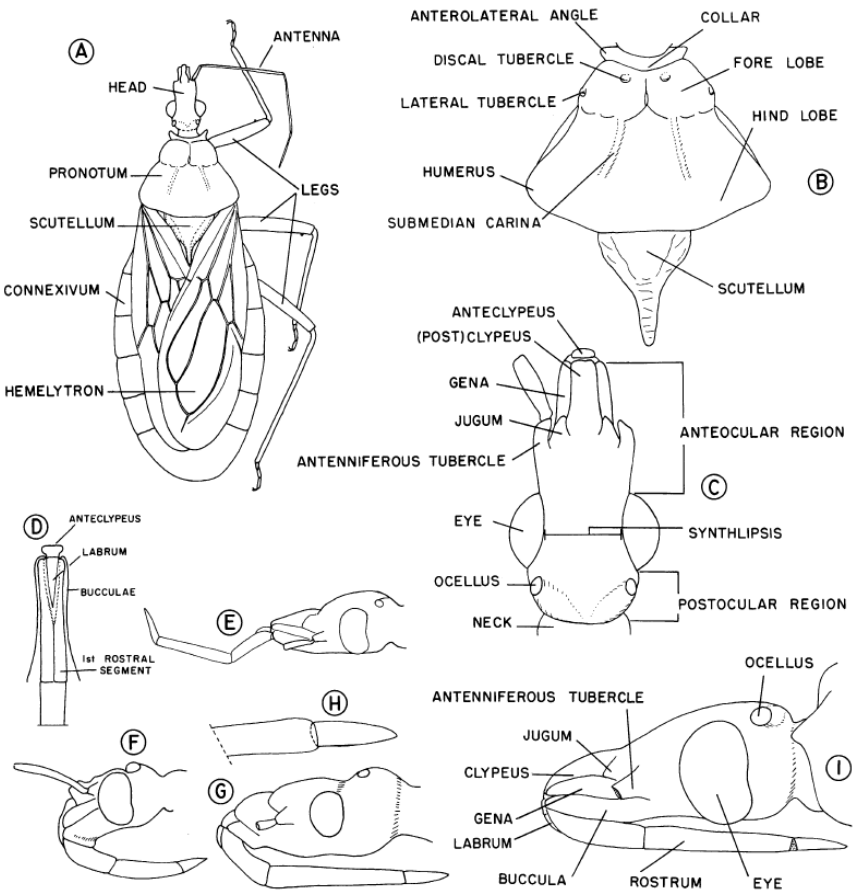


Figure 1. Some general morphology of triatomines. Adapted from [1], p. 141.

A semi-automated, experimental morphometrics-extraction tool, *MorphEx* ([13]; based on mouse-picking coordinates of extrema of the members of (D)) was implemented in *Mathematica* ([9]).

Members of (D) for 10 randomly selected specimen images for each of the species in (S) were obtained using *MorphEx*. Similarly, relevant morphometrics of 10 specimen images for nine triatomine species not in (S) were obtained.

In addition, 1000 "synthetic" test images of each species in (S) were created by Gaussian filtering of images obtained from [1]. "Reference" image features required by (D) (an image of a pronotum from *T. infestans*, an image of a head from *T. brasiliensis*, an image of an antenna segment from *T. dimidiata*) were extracted by manual graphics editing. A tool, *MorphAlign* ([15]), based on the *SIFT* method ([16]) was implemented and used to automatically extract the correlates of these reference image features from the synthetic images. The morphometrics in (D) were automatically computed from these correlates.

A Bayesian network (BN, [4]), *BTC_Adult_V6*, that models the relationships between (S) and (D) was developed in the Windows *Netica* ([3]) Bayesian network development and runtime framework.

A BN is a system of conditional probabilities C ([7], Section 9.1) mapped onto a directed acyclic graph ([8], pp. 12, 23) G. In this mapping, the nodes of G represent random (“state”) variables ([7], Section 3.1) of C. A value of a random variable V in C is called a "state" of V. An

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

edge E from a node (random variable) A to a node (random variable) B signifies that the values (states) of B are a conditional probability function of the values (states) of A. If the composition of probabilities along each path in G is a conditional probability, then G is a Bayesian network.

A BN requires a posit of prior probabilities. For all classifiers, a uniform prior probability distribution was assigned to the members of (S).

In *BTC_Adult_V6*, the nominal ranges of values of members of (D) were partitioned as shown in Figure 3. For each member *d* of (D) and each member *s* of (S), the prior probability was modeled as a normal distribution. The mean of the distribution was assumed to be the midpoint of the range *r* of *d* for *s* as documented in [1]. *r* was assumed (by fiat) to have 90% of the total probability. Given these assumptions, the standard deviation of the distribution for each (*s*,*d*) pair was computed using a *Mathematica v9* ([9]) script ([14]).

In *BTC_Adult_V6*, there is a distinct normal distribution for each (s,d) pair, for a total of 45 distribution-definitions.

For any member *d* of (D), the prior-probability definitions, together with their associated interactive user-interface constraints, prohibit assertion of measurement values that lie outside the union (across all species) of values of *d* that are documented in [1].

3.0 Results

Example triatomine images are shown in Figure 2.

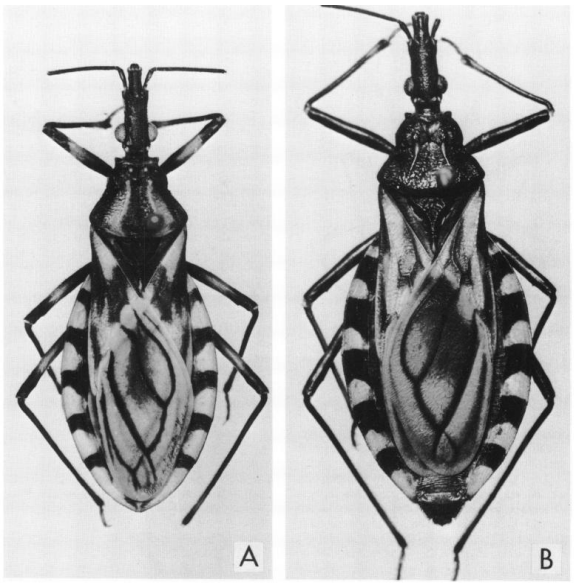


Figure 2. *Triatoma brasiliensis*. A. Male Ceará, Brazil. B. Female, dark form, Bahia, Brazil. Adapted from Figure 43, [1].

A nominal user view of *BTC_Adult_V6* is shown in Figure 3.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

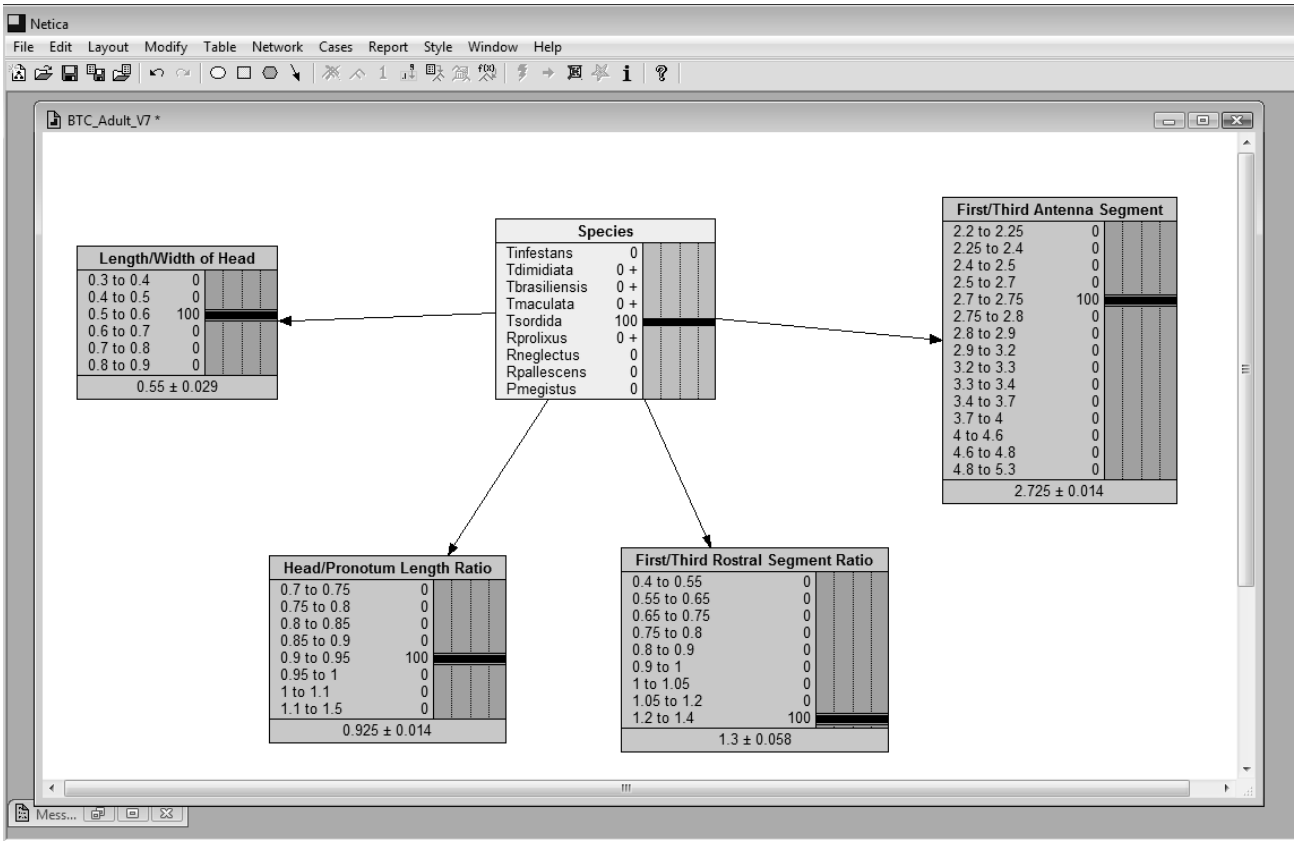


Figure 3. A nominal user view of *BTC_Adult_Vx*.

On the screen, there is one box for (S) and one box for each member of (D). Arrows correspond to the conditional probability of the diagnostic variables (D) on Species (S).

Each box in Figure 3 has two or three regions, delimited by horizontal borders.

The top region of a box contains the name of a (random) variable of interest, e.g., "Length/Width of Head".

The region immediately below the topmost region of a box consists of three elements (read horizontally):

i. a value-range for the variable named in the top region of the box

ii. to the right of (i), a numerical literal (expressed as a percentage) indicating the probability that the variable of interest has a value lying in the value-range

iii. to the right of (ii) a (segment of a) a histogram representation of the (percentage) probability that the variable of interest has a value lying in the value-range denoted by (ii). Taken as a whole, the histogram spanning the middle region of the box represents the probability distribution for the variable named in (i), conditional on the variables at the tails of the arrows whose heads touch the box. For

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

example, in Figure 3 the box in the lower left is associated with the variable "Length/Width of Head", given a "Species".

If a box has three regions, the lowest region reports the mean±standard_deviation of the probabilities of the random variable associated with that box.

A box with a grey background means the variable corresponding to that box is intended as an "input" (also called an "asserted-value" or "finding") variable. Input variables represent information that is posited as given. For example, in Figure 3, "Head/Pronotum Length Ratio" is an "input" with an asserted value of "0.95 to 1".

A box with a pink background means the variable corresponding to that box is intended as an "output" (also called a "calculated") variable. In typical operation, Box (S) has a grey background; each of (D), a pink background. For example, in Figure 3, "Species" is an "output"/"calculated" value which indicates that it is "100.0%" probable that the species of interest is *Triatoma infestans*, given the inputs asserted in the grey boxes in Figure 3.

The basic operation of the classifiers is simple. In interactive mode the user places the mouse pointer over a value of a member of (D), and clicks once. The resulting species prediction (a probability) for each member in each of (S) will appear in the "Species" box.

BTC_Adult_V6 also supports a non-interactive (in the *Netica* vocabulary, a "case-file") mode for automaticallly processing sets of morphometric descriptors. The relevant morphometrics of ten specimen images for each of the species in (S) that were obtained using *MorphEx* were submitted to *BTC_Adult_V6* via a *Netica* "Case File". All specimens were correctly classified.

Similarly, relevant morphometrics of ten specimen images for nine triatomine species not in (S) were obtained and submitted to *BTC_Adult_V6*, which determined those input-sets to be of "unknown" species.

BTC_Adult_V6 correctly classified all 9000 synthetic test images based on the automated extraction of image features by ImageAlign.

4.0 Discussion and conclusions

The results described in Section 3.0 motivate at least two observations:

1. A BN can provide automated support for CD vector classification. It is especially useful for providing probability-constrained classifications when we have only partial information about morphological features relevant to those classifications.

2. The assignment of prior probabilities in a BN classifier in general is not unique, nor does it need to be. All that is required is that the classifier, as a whole, is to correctly resolve the species of interest.

3. It is rare to find triatomine specimens whose rostral segments are extended, thus limiting the use of this metric in classification. Similarly, triatomine antennae are quite fragile, and it is not uncommon to find specimens in which one or more the antenna segments missing. The absence of both extended rostral segments and antenna strongly limits the power of the classifier.

4. How well the BN technique described here can be extended to include accommodate additional species and morphometrics is an open question. There is research in progress at the University of Kansas Biodiversity Institute addressing this question.

5.0 References

[1] Lent H and Wygodzinsky P. *Revision of the Triatominae (Hemiptera, Reduviidae), and Their Significance as Vectors of Chagas' Disease. Bulletin*

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

of the American Museum of Natural History 163:3 (1979).

[2] Martinez D. Informal communication. 2013.

[3] Norsys Software Corporation. *Netica*. http://www.norsys.com. 2011.

[4] Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Second Revised Printing. Morgan Kaufmann. 1988.

[5] US Centers for Disease Control. Parasites -- American Trypanosomiasis (also known as Chagas Disease): Detailed FAQs. http://www.cdc.gov/parasites/chagas/gen_info/detailed.html. 2012.

[6] US Centers for Disease Control. Parasites -- American Trypanosomiasis (also known as Chagas Disease): Antiparasitic Treatment. http://www.cdc.gov/parasites/chagas/health_professionals/tx.html. 2012.

[7] Chung KL. *A Course in Probability Theory.* Third Edition. Academic Press. 2001.

[8] Diestel R. *Graph Theory*. Springer. 1997.

[9] Wolfram Research. *Mathematica Home Edition v9.0.0*. http://www.wolfram.com/mathematica-home-edition/. 2013.

[10] Komp E. Informal communication. May 2013.

[11] Jensen FV. *Bayesian Networks and Decision Graphs.* Springer. 2001.

[12] Petrou M and Petrou C. *Image Processing: The Fundamentals*. Second Edition. Wiley. 2010.

[13] Horner JK. *Morphex*, a semi-automated tool for extracting morphometrics from digitized triatomine images. Available on request from the author.

[14] Available from the author on request.

[15] Horner JK. *MorphAlign*, a SIFT-based tool for automatically extracting morphometrics from images. 2013. Available from the author on request.

[16] Lowe DG. Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision* 2 (1999), 1150–1157.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Treatment Ontology Framework of Clinical Practice Guidelines

Eman Alqaissi and Samir El-Masri

College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia

eman.alqaissi@gmail.com
selmasri@gmail.com

Abstract

The goal of computerizing Clinical Practice Guidelines (CPGs) is to facilitate their use in practice. Many models are presented and used in different Decision Support Systems (DSSs). These models concern about representing CPGs in different ways, but none of them is used as a framework that unifies CPGs development. The problem exists when medical experts whose job focuses on the development of CPGs, try to develop them by using different templates. This paper proposes such an ontology framework; that is especially in treatment recommendations. The framework unifies the representation of CPGs in a machine-readable format by adopting the use of SNOMED CT terminology for all instances. In addition, it meets the reusable, comprehensive, efficiency, flexibility, accuracy, and consistency benefits.

Key words: clinical practice guidelines, activity diagram, class diagram, SNOMED CT, ontology, knowledge base, framework, treatment

1. Introduction

A Clinical Practice Guideline (CPG) has been defined by the Institute of Medicine (IOM) as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific circumstances." [1]. The benefits of CPGs are numerous, they can be used to improve the process and outcomes of health care, to make efficient use of resources, and to improve the quality of clinical decisions; which will improve the patient care quality, reduce medication errors, and minimize the cost of patient treatment. CPGs allow the use of knowledge at the appropriate point of patient care, and also the reuse of knowledge when it is applied to different situations [2]. CPGs are prepared by panels of expertise and they are evidence-based. Many countries concerned about the development and implementation of CPGs such as Australia, Canada, England, United State and Japan [3].

The ontology concept is used for representing CPGs, in literature; many definitions for ontology were introduced. Gruber [4] proposed the most popular definition of ontology as "*…a formal, explicit specification of shared conceptualization*".

SNOMED CT is the most comprehensive reference clinical terminology constantly updated, to support the effective coding, retrieving, and analyzing of clinical data, with the aim of improving patient care. SNOMED CT is used in over 50 countries around the world and is a key terminology standard recommended by *Infoway* for use in health information and communication technologies in Canada [5]. In addition, the American academy of ophthalmology adopts the use of SNOMED CT [6].

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2. Related Work

M. Peleg et al. [7] mentioned in their comparison, many Computer Interpretable Guideline (CIG) representation languages which were developed to represent clinical knowledge contained in CPGs. Their common goal is to computerize CPGs, in order to help clinicians by facilitating the process of accessing information contained in clinical guidelines. A lot of standardization works for representing CPGs have been proposed. The most common formats are Arden Syntax [8], PROforma [9], EON [10], GLIF [11], PRODIGY [12], Asbru [13] and Guide [7]. These models represent complete CPGs aspects, in this paper the focus is on their representations of treatment part in CPGs.

S. Raza Abidi and S. Shayegani [14] discussed a knowledge modeling approach for the form and function of CPGs by developing ontological model to computerize CPGs. They suggest a complete general model for CPGs that could be used as a template for authoring CPGs by health professionals. The main problem exists in their representation is their focus on a small number of CPGs for modeling. In addition, they use inductive reasoning approach for deriving conclusion which may not be true until deductive reasoning approach is applied in some examples.

3. The Proposed Treatment Ontological Framework

This framework stems from the analysis of CPGs samples selected from authoritative resource which is National Guideline Clearinghouse (NGC) [15]. The resulted treatment class diagram from this analysis is converted to ontological model. Moreover, the framework is merged with SNOMED CT, which is the most comprehensive standard terminology. The use of SNOMED CT is ideal because this supports interoperability with any clinical system. All classes in treatment ontology should be related directly to *Concept* class in SNOMED CT ontology. By following merging wizard from Refactor tab in protégé, the two ontologies are merged in one place. Then, all treatment ontology classes are considered as subclasses for the *Concept* class in SNOMED CT ontology, by using generalization and many-to-one merging technique. Figure1 is the OntoGraf modeling that help in visualizing the classes and their relationships as a network with arcs, where each arc represents a relationship. In fact, in this case it shows subclass relationships.

As the inductive reasoning approach is used to build the proposed framework, the deductive reasoning approach is also applied on five cases which are selected randomly from CPGs. These cases include *Stable coronary artery disease with or without angina*, *Acute bacterial sinusitis*, *Rhinitis*, *Renal cell carcinoma*, and *Chronic and recurrent gout*. In this paper, one of these CPGs is explained in detail, to prove the truth of the proposed CPGs treatment ontology framework. The reasoner HermiT 1.3.6 which is built in protégé software is used, to check the inconsistencies in the classes, their relationships, compute the inferred super classes, and many other inference services.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

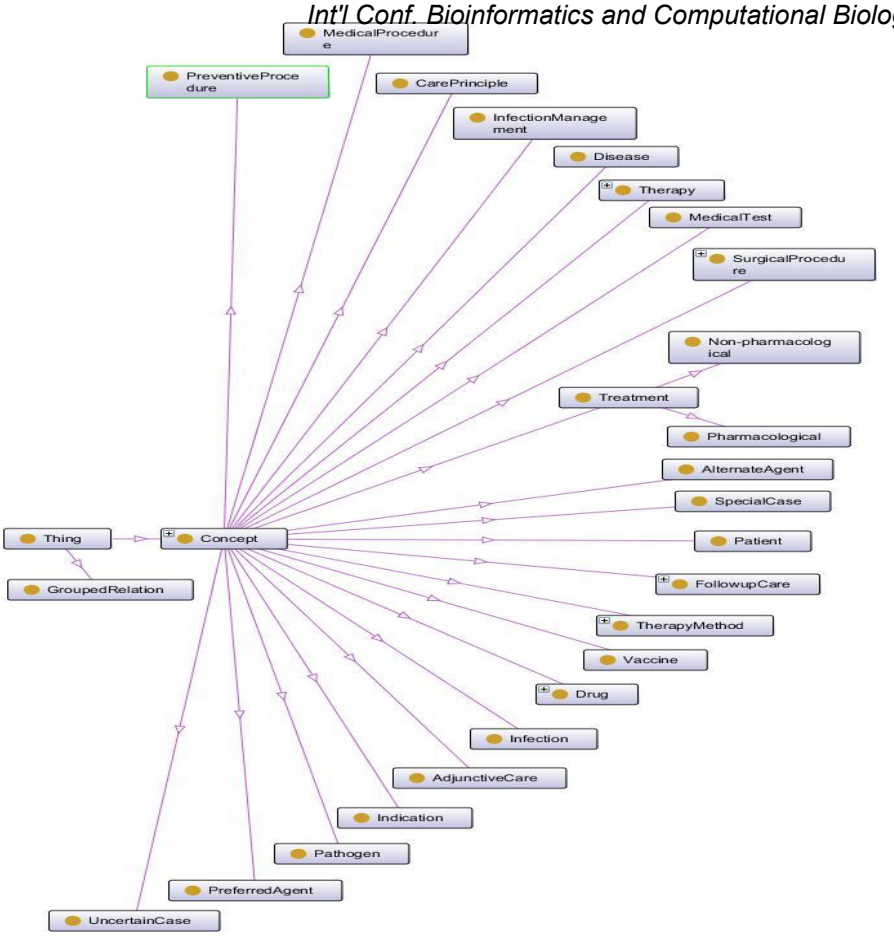


Figure 1: OntoGraf plugin for the proposed treatment ontology

The acute Bacterial Sinusitis in Children CPG has the following classes and their matched classes in the proposed treatment framework.

Table1. Instances and their matched classes in acute bacterial sinusitis treatment

CPG Text Based Treatment Information (Instances)	**Matched Classes**
Acute bacterial sinusitis	Disease
Outpatient, Child	Patient
Outpatient observation for 3 days	Care Principle
Antibiotic therapy Amoxicillin with or without clavulanate High-dose amoxicillin-clavulanate Clindamycin and cefixime OR linezolid and cefixime OR levofloxacin	Therapy and Drug
Reassess initial management if worsening or failure to improve reported.	Follow up Care

These instances are created by using individual tab in protégé. The following figure shows acute bacterial sinusitis treatment creation. The proposed treatment framework covers about 99% of the treatment contained in this guide because the instance called "reassess initial management" cannot be represented in SNOMED CT. Each instance is represented by SNOMED CT *conceptID*. These

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

concept IDs must be all in *Concept* class, in this example acute bacterial sinusitis disease is represented by STC_ 75498004 which is taken from Snomobile application that works as a database for all SNOMED CT concepts and available at apple store. After running the reasoner HermiT 1.3.6, the following figure provides all inferences. In this sample, there is no grouped relation, but the original relationships between classes are defined and applied between instances in a very simple manner.

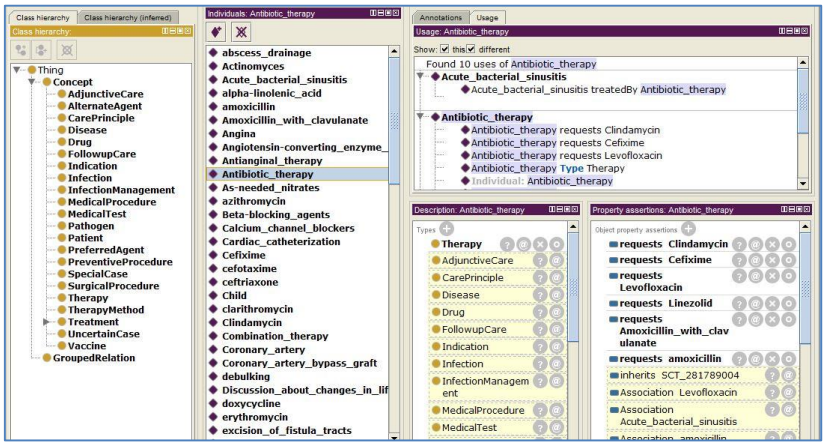


Figure 2. Acute bacterial sinusitis treatment after running HermiT reasoner

The description view of acute antibiotic therapy shows that it belongs to Therapy class, and used to treat acute bacterial sinusitis. After applying reasoner, the antibiotic therapy belongs to many other classes either directly or indirectly such as adjunctive care, care principle, follow-up care, etc. because relationships to these classes exist in the treatment framework. Furthermore, in the property view relations to antibiotic therapy are mentioned along with its concept ID.

4. Conclusion

The framework of treatment ontological model, proposed in this paper, is specialized in treatment procedures contained in clinical practice guidelines that are evidence-based. Many representation languages for representing CPGs take their place in practice and adopted in many systems. The proposed framework is one contribution in the area which tries to override as much as disadvantages to facilitate this representation with a unified framework that use SNOMED CT standard terminology. The framework provided a template that facilitates CPGs creation and development process. In addition, it helps in building treatment knowledge base, which can be used in Decision Support System (DSS) to benefit and guide clinicians.

5. References

[1] A. Latoszek-Berendsen, H. Tange, H. J. van den Herik, and A. Hasman, "From clinical practice guidelines to computer-interpretable guidelines. A literature overview," Methods Inf Med, vol. 49, pp. 550-70, 2010.

[2] D. Isern and A. Moreno, "Computer-based execution of clinical guidelines: a review," *Int J Med Inform*, vol. 77, pp. 787-808, Dec 2008.

[3] H. S. Ahn and H. J. Kim, "Development and implementation of clinical practice guidelines: current status in Korea," *J Korean Med Sci*, vol. 27 Suppl, pp. S55-60, May 2012.

[4] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, pp. 199-220, 1993.

[5] Infoway website. [Online]. Available: https://www.infoway inforoute.ca/ ,last accessed on October 2013.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[6] H. D. Hoskins, P. L. Hildebrand, and F. Lum, "The American Academy of Ophthalmology adopts SNOMED CT as its official clinical terminology," *Ophthalmology*, vol. 115, pp. 225-6, Feb 2008.

[7] M. Peleg, S. Tu, J. Bury, P. Ciccarese, J. Fox, R. A. Greenes, *et al.*, "Comparing computer-interpretable guideline models: a case-study approach," *J Am Med Inform Assoc*, vol. 10, pp. 52-68, 2003 Jan-Feb 2003.

[8] M. Samwald, K. Fehre, J. de Bruin, and K. P. Adlassnig, "The Arden Syntax standard for clinical decision support: experiences and directions," *J Biomed Inform*, vol. 45, pp. 711-8, Aug 2012.

[9] J. Fox, N. Johns, and A. Rahmanzadeh, "Disseminating medical knowledge: the PROforma approach," *Artif Intell Med*,vol. 14, pp. 157-81, 1998 Sep-Oct 1998.

[10] F. Ongenae, F. De Backere, K. Steurbaut, K. Colpaert, W. Kerckhove, J. Decruyenaere, *et al.*, "Towards computerizing intensive care sedation guidelines: design of a rule-based architecture for automated execution of clinical guidelines," *BMC Med Inform Decis Mak*, vol. 10, p. 3, 2010.

[11] L. Ohno-Machado, J. H. Gennari, S. N. Murphy, N. L. Jain, S. W. Tu, D. E. Oliver, *et al.*, "The guideline interchange format: a model for representing guidelines," *J Am Med Inform Assoc*, vol. 5, pp. 357-72, 1998 Jul-Aug 1998.

[12] K. Zheng, R. Padman, M. P. Johnson, and S. Hasan, "Guideline Representation Ontologies for Evidence-Based Medicine Practice," in Handbook of Research on Advances in Health Informatics and Electronic Healthcare Applications: Global Adoption and Impact of Information Communication Technologies, ed: IGI Global, 2010, pp. 234-254.

[13] S. Miksch, Y. Shahar, and P. D. Johnson, "Asbru: A Task-Specific, Intention-Based, and Time-Oriented Language for Representing Skeletal Plans," presented at the 7th Workshop on Knowledge Engineering: Methods & Languages (KEML-97), 1997.

[14] S. Abidi and S. Shayegani, "Modeling the Form and Function of Clinical Practice Guidelines: An Ontological Model to Computerize Clinical Practice Guidelines," in *Knowledge Management for Health Care Procedures*. vol. 5626, D. Riaño, Ed., ed: Springer Berlin Heidelberg, 2009, pp. 81-91.

[15] National Guideline Clearinghouse (NGC). [Online]. Available: http://www.guideline.gov/about/index.aspx, last accessed on October 2013.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

SESSION

PROTEIN CLASSIFICATION AND STRUCTURE PREDICTION, FOLDING, AND COMPUTATIONAL STRUCTURAL BIOLOGY + DRUG DESIGN

Chair(s)

TBA

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

An Efficient Algorithm for Protein-Protein Interaction Network Analysis to Discover Overlapping Functional Modules

Ying Liu

[1]Department of Computer Science, Mathematics and Science, College of Professional Studies, St. John's University, Queens, NY 11439

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. Previously we proposed an efficient, novel, and fast overlapping clustering method is proposed based on purifying and filtering the coupling matrix (PFC). In this paper, we further improved PFC. The experimental results show that the improved 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

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

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).

Liu and Foroushani [16] proposed a PFC filtering by simple, local criteria. In this paper, we propose a new PFC approach, filtering by corroboration.

3.1 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 compromise 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 constrains 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.

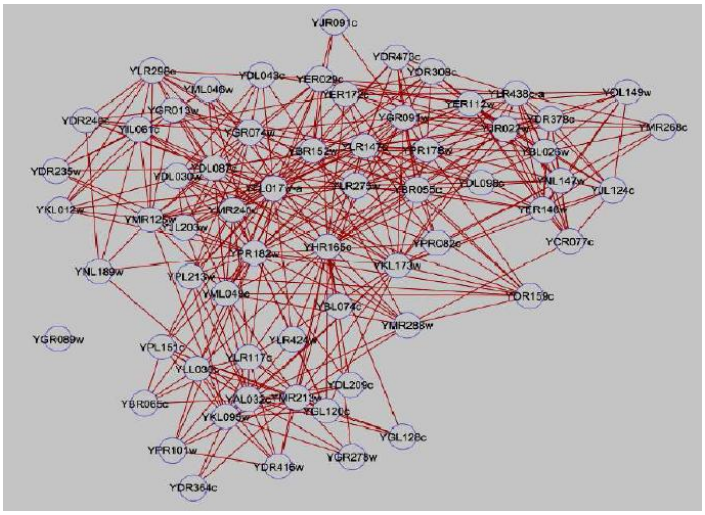


Figure 1 This Figure shows the community for the row labeled "YKL173w" in the purified coupling matrix of yeat PPI network. It is one of the clustered selected by PFC1. Out of the 63 proteins in this community, 58 belong to MIPS Funcat 11.04.03.01**.**

3.2 Filtering by Corroboration

Filtering by local criteria gives impressing results but it does not guarantee that a few of the remaining clusters do not overlap in majority of their elements. Although PFC is an overlapping clustering algorithm, very large overlaps between clusters are bound to indicate presence of redundant clusters. At the same time, repeated concurrence of large groups of proteins in different rows does reinforce the hypothesis that these groups are indeed closely related, and that the corresponding rows represent a high quality cluster. These observations can be used to construct an alternative filter that removes both low quality and redundant clusters from the coupling matrix. The main idea is that a line A is corroborated

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

by a Line B if the majority of nonzero elements in A are also nonzero in B. The following summarizes this filter:

Given the sparse nature of the involved matrices, this Corroboration based filter can be implemented very efficiently in Matlab. It discards by design redundant clusters (out-degree>0 in the confirmation graph indicates that there is a similar cluster with a higher rank) and retains only high quality clusters (clusters with a high in-degree in the confirmation graph have been confirmed by presence of many other clusters with similar structure). The ranking by row sum helps consolidate and summarize relevant parts of smaller clusters into larger ones. Figure 2 gives two examples of clusters selected by this approach on Yeast-PPI network.

Given the Binary version of the Purified Coupling Matrix B
Calculate Overlap Matrix $O = B * B$
Normalize $O(i,j)$ by Size of Module j
Calculate Corroboration Matrix $C = \lfloor O ./ \alpha \rfloor$
Where: $0.5 < \alpha \leq 1$; and "./" is the Matlab cellwise division.
Calculate Common Corroborator Matrix $C_{Com} = C * C'$
Rank the rows of C_{com} by the sum of their entries
Interpret C_{com} as description of a directed Confirmation graph between clusters, where the direction of confirmation is from lower ranked to higher ranked rows.
Select clusters whose in-degree in the confirmation graph is higher than a threshold and whose out degree is 0.

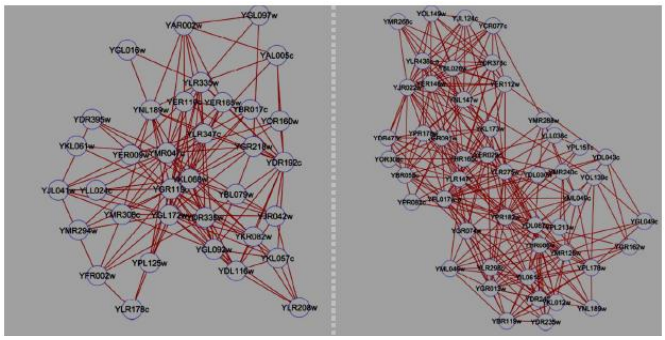


Figure 1: Two of the clusters selected by PFC2. The left Figure shows the selected community for the row labeled "YDR335w" in the purified coupling matrix. Out of the 35 proteins in this community, 29 belong to MIPS Funcat 20.09.01(nuclear transport). The right Figure shows the selected community for the row labeled "YKL173w" in the purified coupling matrix. 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(Splicing).

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 and Figure 3 are 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
PFC1	183	44.76	-19.35	0.31	55/183
PFC1	40	25.4	-19.40	0.17	36/40

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

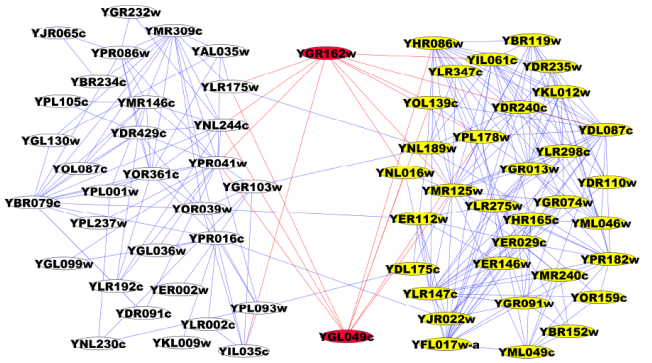


Figure 2. result #1: The dominant function for the left module is translation initiation (10 out of 31) for the right module, it is nuclear mRNA splicing (27 out of 33); both overlap nodes are involved in translation initiation and Protein-RNA complex assembly.

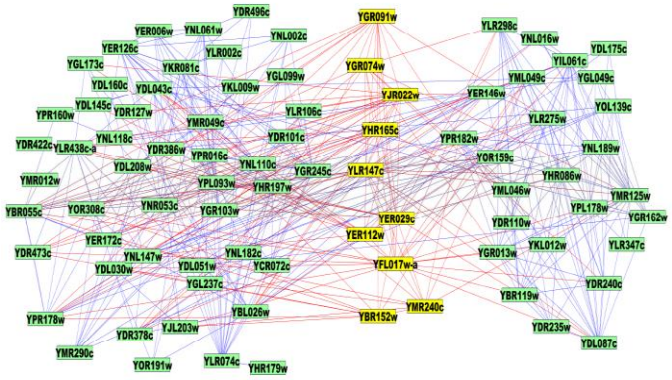


Figure 3. result #2. There is a relatively large overlap (yellow nodes). All 10 overlap nodes are involved in "nuclear mRNA splicing, via spliceosome-A". The same is true for ca.25% (12 out of 45) of the green nodes to the left and 68% (17 out of 25) of the green nodes to the right of the overlap. Furthermore, 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.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[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/funcat>.

[16] Liu, Y, and Foroushani, A. An Efficient Soft Graph Clustering Method for PPI Networks based on Purifying and Filtering the Coupling Matrix. BioComp 2011.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Dynamic protein-protein interaction networks and the detection of protein complexes: an overview

Eileen Marie Hanna and Nazar Zaki
College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi, UAE

Abstract - *Developing computational approaches for the detection of protein complexes in protein-protein interaction networks continues to be an evolving area of research. These approaches seek to complement the experimental methods which are usually expensive in terms of time and cost. A protein-protein interaction dataset is typically modeled as a static network whose vertices and edges respectively represent all the proteins and their interconnections. Despite the agreeable accuracies attained by various computational methods when applied on such networks, their additional improvements seem to face some limitations. It is believed that the more enrichment with biological information is added to the interaction networks and complex-detection algorithms, the better will be the overall quality of the results. In this paper, we stress on the importance of reflecting the dynamic nature of protein interaction networks as a primary enhancement phase and we highlight possible aspects by which it could be acquired.*

Keywords: protein-protein interactions, protein complex, dynamic protein-protein interaction network.

1 Introduction

From metabolism to signal transduction, transport, cellular organization and ultimately all biological processes, proteins are the key players. Their interconnections shape interaction networks which define highly-organized cellular systems [1]. Biological functions are often acquired through collaborations of interacting protein groups referred to as protein complexes [2]. The progress in identifying protein complexes, involved in normal molecular events as well as phenotypes associated with diseases, allows the progressive development of effective cures. Accordingly, various experimental methods were designed to identify complexes given protein-protein interaction (PPI) data. However, in addition to their high computational cost, they are also susceptible to high error rates [3]. Therefore, several computational approaches came into the picture to complement the experimental activities. For instance, protein complexes detected by computational algorithms with suitable accuracy and quality could guide the experimental examinations and expectantly reduce the necessary biological explorations.

In a computational setting, a PPI dataset is usually modeled as a graph whose vertices and edges represent all the proteins and their interactions respectively. In this context, the majority of the computational approaches are based on the concept by which protein complexes correspond to dense subgraphs. These methods include, but are not limited to, Markov Clustering (MCL) [4] which uses random walks in protein interaction networks; the molecular complex detection (MCODE) algorithm [5] which identifies complexes as dense regions grown from highly-weighted vertices; the clustering based on maximal cliques (CMC) method [6]; the Affinity Propagation (AP) algorithm [7]; ClusterONE [8] which identifies protein complexes through clustering with overlapping neighborhood expansion; the restricted neighborhood search (RNSC) algorithm [9,10]; the RRW algorithm which generates complexes by using repeated random walks [11]; and CFinder [12] which is based on the clique percolation method. Other approaches which are not based on the density notion include ProRank [13,14] which mainly uses a protein ranking algorithm to identify essential proteins in a PPI network; ProRank+ [15] which is an improved version of ProRank, it reflects the fact that proteins can be multifunctional and thus could belong to multiple complexes and it applies a merging procedure to improve the detected complexes; and finally PEWCC [16,17] which assesses the reliability of PPI data based on the weighted clustering coefficient notion prior to detecting protein complexes. When evaluated based on reference sets of biologically- identified protein complexes, these algorithms were on the right track. Nevertheless, their improvements towards reducing false positive and false negative outcomes seem to be bounded by the way in which PPI data is originally utilized and by the false positive and false negative interactions as well. The traditional experimental approaches used to study PPIs, such as yeast two-hybrid (Y2H) [18] and TAP-MS [19], do not provide temporal, spatial or contextual information across which a PPI occurs. In contrast, recent methodological advances, such as ChIP-chip [20] and ChIP-seq [21] can make such informative data available. Consequently, advances in the computational approaches developed to analyze PPI networks, including those designed to detect protein complexes, ought to relate to such diversity of information that is currently presented. PPI networks are dynamic in nature [22]. Accordingly, modeling the dynamicity of PPI networks is a necessary shift in the way such networks are viewed and studied [23]. It is actually essential and allows us to expand our

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

knowledge about how cellular processes occur. In this paper, we highlight the advantages, potential approaches and possible bottlenecks of this emerging construal of PPI networks.

2 The advantages of shifting to dynamic PPI networks

2.1 Enhancing the replication of real biological events

The shift to dynamic PPI networks in computational approaches of systems biology comes as a natural response to advances in experimental methods by which novel types and increased amounts of biological data are generated. As an interdisciplinary area of research, the more representative are its building models and methods, the better is its aptitude. Moreover, when cellular interactions are reproduced in a more realistic manner, the accountability and accuracy of the results produced by computational methods will certainly augment.

2.2 Potentially uncovering previously unknown biological facts

A PPI dataset is conventionally represented as a comprehensive graph which includes the proteins along with all their interactions. However, not all the interconnections happen at the same time. In fact, the occurrence of a PPI is subject to various temporal, spatial and contextual conditions. Obviously, encompassing such conditionality parameters elucidates the dynamics of PPIs. In view of that, by combining biological information, we would reach a computational visualization level of protein interaction events that could verify or even contradict biological concepts. Furthermore, previously unknown facts may be learned, such as the characterization of hub proteins in [24] as "party hubs" which interact with their partners at the same time or "date hubs" which connect to their partners at different times and locations.

2.3 Possibly overcoming data limitations

The biological methods used to identify protein interactions are very sensitive to experimental settings. Therefore, the PPI datasets that they generate are always liable to high error rates. Many algorithms were developed to filter protein interactions according to their reliability levels. For example, some of these methods use weighting schemes based on the number of common neighbors of interacting proteins such as CDdistance [25], FSWeight [26] and AdjustCD [27]. Similarly, the PE-measure introduced in [16,17] reduces the level of noise in protein interaction networks by looking for subgraphs that are closest to maximal cliques based on the weighted clustering coefficient measures. In addition, possible enrichment data that can be used to model the dynamicity of PPI networks, such as gene expression profiles [28] and gene ontology [29], suffer from low gene coverage in contrast with most PPI datasets, in which the number of interacting proteins is typically very high [30]. The recurrence of information and/or inferences that are drawn from different types of biological data can be seen as a confidence indicator. In view of that, combining various datasets, although not fully-credible, in the direction of modeling PPI dynamics could potentially reduce data limitations such as the effect of false positives and false negative rates, as well as low coverage issues.

2.4 Increasing the ability to categorize the information deduced from PPI networks

Dynamic PPI networks, once modeled, can provide a closer view of their corresponding cellular events. Accordingly, in contrast with static PPI networks, the information revealed by dynamic networks is at a higher level of details. For instance, in the problem of identifying protein complexes in protein interaction networks, most of the presented algorithms do not differentiate between functional modules and protein complexes. That is mainly due to the absence of embedded information in the networks that could guide the search. In fact, complexes are formed by proteins which interconnect at the same time and place, whereas the members of functional modules may interact at different times and places [31]. Accordingly, when PPIs are bounded by spatiotemporal conditions inferred by gene expression and gene ontology datasets for example, the detected components could more likely be categorized as protein complexes or functional modules. Likewise, dynamic PPI modeling may highly contribute to the detection of protein subcomplexes in PPI networks. Various approaches were developed to solve this important research problem, but all based on static networks [32]. As dynamic modeling could reveal the mechanisms of protein-complex formation and could yield better complex-detection approaches, it could also provide the same for the detection of subcomplexes.

2.5 Increasing the accountability and the accuracy of the results produced by computational methods

Undeniably, dynamic PPI networks describe cellular interactions in a more realistic manner. Therefore, the computational methods, customized to suit such networks, would certainly produce analytical results with higher accuracy and accountability. Here, we namely consider the algorithms designed to detect protein complexes in protein interaction networks. The integration of temporal, spatial or contextual biological information with PPI data as a means to show the PPI dynamics, can be viewed as a kind of clustering based on temporal, spatial and/or contextual attributes. Hence, the proteins and their interconnections can be grouped based on the integrated conditions and a protein complex-

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

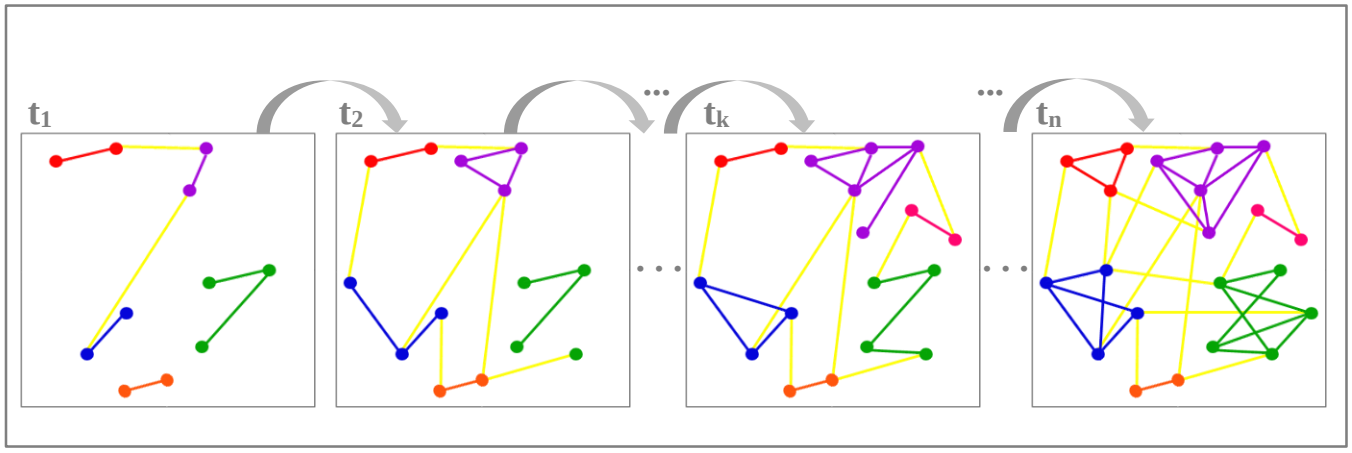


Fig. 1 Snapshots of a hypothetical PPI network, capturing its dynamics at different time points/stages. Each schema includes the available proteins at a certain stage, along with their interconnections. Nodes and edges of similar colors correspond to the same protein complexes, whereas the rest of the edges are represented in yellow.

detection method shall be applied accordingly and with a generalization capability indeed. Once this is achieved, the rates of false positives and false negatives will certainly decrease at the level of the detected complexes and at the level of their protein members as well. Consequently, the overall accuracy of the results will be higher than those scored by methods applied on static networks. The former potentially applies to other exploratory approaches of PPI networks.

3 Modeling Dynamic PPI Networks

A single scheme is usually used to represent a static PPI network with all its components. In contrast, a dynamic PPI network can be visualized by a series of schemes representing snapshots of the network state corresponding to different stages and/or locations of molecular activities, as shown in Fig. 1. The interpretation of a dynamic interaction network and its state transitions depends on the types of data which are used to biologically-condition PPI events. We will hereafter highlight some of the concepts and the approaches to model the dynamicity of protein interaction networks and we will particularly relate them to the problem of detecting protein complexes in PPI networks.

The advancements in experimental techniques are gradually allowing in-depth explorations of biological systems. The resultant progresses can broaden our understanding of biology through the integration of various types of generated information and by consistently developing computational tools to expand our knowledge.

Gene expression datasets are subsequent products which consist of quantitative measurements of RNA species in cellular compartments across different conditions [33]. Genome-wide expression levels can now be studied [34]. Time-series gene expression data report quantities of RNA across different time points in cellular processes. It is believed that genes with correlated expressions across subsets of conditions most likely interact. When combined with PPI data to model the interaction dynamics, it can potentially reveal the processes which underline the formation of protein complexes. For instance, that was done in [35] where it was shown that a just-in-time mechanism elapsing through continuous time points delineates the formation of most complexes. The statistical 3-sigma principle was then used by the works presented in [35] and [36] to define the active time points of proteins based on their gene expression levels and consequently, introduce approaches to detect and refine protein complexes. The core-attachment view of complexes was recently considered in [37]; based on gene expression data, the identification of a protein complex was split into two main parts: a static core consisting of proteins expressed throughout the whole cell cycle and a short-lived dynamic attachment. The results of these approaches were better than the ones tested on static networks. Kim et al. [38] highlighted some of the computational methods used to infer dynamic networks from expression data based on statistical dependence to classify nodes and/or edges as active or inactive. These methods include: Bayesian networks [39], relevance networks [40], Markov Random Fields [41], ordinary differential equations [42] and logic-based models [43].

As they are conditioned by time, PPIs are space-dependent as well. In other words, the occurrence of a protein interaction is also subject to the co-localization of its interacting partners in cellular components [44]. Actually, a failed interaction caused by inappropriate protein localizations could be pathological. Consequently, subcellular localization annotations [45] can be used to model dynamic PPI networks based on spatial constraints. Indeed, the formation of protein complexes is also influenced by the localization settings of proteins. According to that, it is certainly beneficial to incorporate the spatial dynamics towards improving complex-detection approaches. Various methods aim at studying and collecting spatial movements about proteins [46]. However, in addition to mathematical modeling techniques, further

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

approaches to appropriately integrate spatial protein dynamics in PPI networks are still required.

Gene ontology annotations [47], which provide information about genes that are shared across species, can also infer the dynamics of PPI networks [48]. As an indicator of interaction probability, various weighting schemes were introduced to assign PPI weights based on the similarity degrees of gene ontology terms between interacting partners. Among these approaches are SWEMODE [49], which detects communities within PPI networks based on weighted clustering coefficient and weighted average nearest-neighbors degree measures, and OIIP [48], which is a method to detect protein complexes in PPI networks by assigning node and edge weights based on the size of gene annotations.

Gene expression, spatial annotation and gene ontology annotation data could credibly contribute to the incremental attempts to model dynamic PPI networks.

Forthcoming approaches are expected to profit from these data among other types of biological information. Specifically, the integration of biological attributes enhances the computational methods designed to detect protein complexes in protein interaction networks. It not only participates in uncovering the mechanisms of protein-complex formation but also points out useful details for the design of such methods. In addition, the former may help categorize protein complexes and could be informative regarding their building blocks as well.

4 Datasets and Evaluation Measures

The datasets which could be used to enrich PPI networks in order to model their dynamic aspects, such as gene expression and gene ontology data, typically describe the variations of protein activities and/or quantities across sets of conditions. The resulting network analysis ought to consider these conditions. For example, the detected protein complexes in a PPI network enriched by time-series gene expression data would most likely be adherent to the conditions across which the gene expression data were generated. Therefore, reference protein-complex sets which were used to evaluate previous approaches that work on static networks, such as MIPS [50] and CYC2008 [51], may not be the best choice for dynamic networks. Accordingly, reference sets tailored to match input datasets and their conditionality could be more convenient in such cases. Similarly, issues regarding the choice of evaluation measures arise when shifting to dynamic PPI networks. The formulae used to evaluate the accuracy, sensitivity and specificity in addition to other qualities of previous approaches are not the same [8, 52]. Strong evaluation scores include the number of complexes in the reference catalog that are matched with at least one of the predicted complexes with an overlap score, *w*, greater than a certain threshold; the clustering-wise sensitivity (S_n);the clustering-wise positive predictive value (*PPV*); the geometric accuracy (*Acc*); and the maximum matching ratio (*MMR*) which shows how accurately the predicted complexes represent the reference complexes by dividing the total weight of the maximum matching by the number of reference complexes. Given *m* predicted complexes and *n* reference complexes, the corresponding formulae are given by the following equations, where t_{ij} represents the number of proteins that are found in both predicted complex *m* and reference complex *n*.

$$w(A,B) = \frac{|A \cap B|^2}{|A||B|} \quad (1)$$

$$S_n = \frac{\sum_{i=1}^{n} max_{j=1}^{m} t_{ij}}{\sum_{i=1}^{n} n_i} \quad (2)$$

$$PPV = \frac{\sum_{j=1}^{m} max_{i=1}^{n} t_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{n} t_{ij}} \quad (3)$$

$$Acc = \sqrt{S_n \times PPV} \quad (4)$$

5 Conclusion

The realization of dynamic protein interaction networks is a natural evolution which leverages computational methods for biology. It could typically be acquired by investing in recent biological data generated by advanced experimental techniques. These data include, but are not limited to, gene expression, subcellular localization annotation and gene ontology terms annotation datasets which provide temporal, spatial and contextual information about protein interactions throughout cellular processes. With emphasis on the algorithms for the detection of protein complexes, by modeling the dynamics of PPI networks, we could: reproduce the mechanisms of protein-complex formation more realistically; potentially uncover new biological facts about complexes; overcome data limitations existing in most experimental datasets; categorize modules deduced from PPI networks; and finally, increase the accuracy and value of the detected results. Accordingly, novel algorithms for the detection of protein complexes in dynamic protein interaction networks are expected to appear.

6 References

[1] Durbin, R.M., Abecasis, G.R., Altshuler, D.L., et al. "A map of human genome variation from population-scale sequencing". Nature, Vol. 467, 1061–1073, Oct. 2010.

[2] Gavin, A.C., Aloy, P., Grandi, P., et al. "Proteome survey reveals modularity of the yeast cell machinery". Nature, 440, 631–636, Mar. 2006.

[3] Adelmant, G., and Marto, J.A. "Protein complexes: the forest and the trees". Expert Rev. Proteomics, 6(1), 5–10, Feb. 2009.

[4] Dongen, S. "Graph clustering by flow simulation". PhD Thesis. University of Utrecht, Amsterdam, 2000.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[5] Bader, G.D., and Hogue, C.W.V. "An automated method for finding molecular complexes in large protein interaction networks". BMC Bioinformatics, 4:2, Jan. 2003.

[6] Guimei, L., Wong, L., and Chua, H.N. "Complex discovery from weighted PPI networks". Bioinformatics, 25(15), 1891 – 1897, May 2009.

[7] Frey, B.J., and Dueck, D. "Clustering by passing messages between data points". Science, 315(5814):972 – 976, Feb. 2007.

[8] Nepusz, T., Yu, H., and Paccanaro, A. "Detecting overlapping protein complexes in protein-protein interaction networks". Nature Methods, 9, 471 – 472, Mar. 2012.

[9] King, A.D., Przulj, N., and Jurisica, I. "Protein complex prediction via cost-based clustering". Bioinformatics, 20(17), 3013 – 3020, June 2004.

[10] Przulj, N., Wigle, D.A., Jurisica, I. "Functional topology in a network of protein interactions". Bioinformatics, 20(3), 340 – 348, Feb. 2004.

[11] Macropol, K., Can, T., and Singh, A.K. "RRW: repeated random walks on genome-scale protein networks for local cluster discovery". BMC Bioinformatics, 10:283, Sep. 2009.

[12] Adamcsek, B., Palla, G., Farkas, I.J., et al. "CFinder: locating cliques and overlapping modules in biological networks". Bioinformatics, 22(8), 1021 – 1023, Apr. 2006.

[13] Zaki, N.M., Berengueres, J., and Efimov, D. "Detection of protein complexes using a protein ranking algorithm". Proteins: Structure, Function, and Bioinformatics, 80(10), 2459 – 2468, Oct. 2012.

[14] Zaki, N.M., Berengueres, J., and Efimov, D. "Prorank: A method for detecting protein complexes". In Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Conference (GECCO '12), Philadelphia. Edited by Terence Soule: ACM, New York, 209 – 216, July 2012.

[15] Hanna, E.M., and Zaki, N.M. "ProRank+: A Method for Detecting Protein Complexes in Protein Interaction Networks". In Proceedings of the 5th International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS'14), Angers, Loire Valley, France, Mar. 2014.

[16] Efimov, D., Zaki, N.M., and Berengueres, J. "Detecting protein complexes from noisy protein interaction data". In Proceedings of the 11th International Workshop on Data Mining in Bioinformatics (BIOKDD '12), 1-7, Aug. 2012.

[17] Zaki, N.M., Dmitry, D., and Berengueres, J. "Protein Complex Detection using Interaction Reliability Assessment and Weighted Clustering Coefficient". BMC Bioinformatics, 14:163, May 2013.

[18] Fields, S., and Song, O. "A novel genetic system to detect protein–protein interactions". Nature, 340, 245 – 246, July 1989.

[19] Collins, M.O., and Choudhary, J.S. "Mapping multiprotein complexes by affinity purification and mass spectrometry". Curr. Opin. Biotechnol., 19, 324 – 330, Aug. 2008.

[20] Kim, T.H., and Ren, B. "Genome-wide analysis of protein–DNA interactions". Annu. Rev. Genomics Hum. Genet., 7, 81 – 102, Sep. 2006.

[21] Johnson, D.S., Mortazavi, A., Myers, R.M., et al. "Genome-wide mapping of in vivo protein–DNA interactions". Science, 316, 1497 – 1502, June 2007.

[22] Levy, E.D., and Pereira-Leal, J.B. "Evolution and dynamics of protein interactions and networks". Curr. Opin. Struct. Biol., 18, 349 – 357, June 2008.

[23] Przytycka, T.M., Singh, M., Slonim, D.K. "Toward the dynamic interactome: it's about time". Briefings in Bioinformatics, 11, 15 – 29, Jan. 2010.

[24] Han, J.D., Bertin, N., Hao, T., et al. "Evidence for dynamically organized modularity in the yeast proteinprotein interactionnetwork". Nature, 430, 88 – 93, July 2004.

[25] Brun, C., Chevenet, F., Martin, D., et al. "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network". Genome Biol., 5(1):R6, Dec. 2003.

[26] Chua, H., Ning, K., Sung, H.W., et al. "Using indirect protein-protein interactions for protein complex prediction". J. Bioinform. Comput. Biol., 6, 435 – 466, Jan. 2008.

[27] Hon, N.C., Sung, W.K., and Wong, L. "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions". Bioinformatics, 22, 1623 – 1630, July 2006.

[28] Chen, J., and Yuan, B. "Detecting functional modules in the yeast protein-protein interaction network". Bioinformatics, 22, 2283 – 2290, July 2006.

[29] Xu, B., Lin, H., and Yang, Z. "Ontology integration to identify protein complex in protein interaction networks". Proteome Sci., 9:S7, Oct. 2011.

[30] Von Mering, C., Krause, R., Snel, B., et al. "Comparative assessment of large-scale data sets of protein-protein interactions". Nature, 417, 399 – 403, May 2002.

[31] Spirin, V., and Mirny LA. "Protein complexes and functional modules in molecular networks". PNAS, 100, 12123 – 12128, Oct. 2003.

[32] Zaki, N.M., and Mora, A. "A comparative analysis of computational approaches and algorithms for protein subcomplex identification". Scientific Reports, 4: 4262, nature group, Mar. 2014.

[33] Lovén, J., Orlando, D.A., Sigova, A.A., et al. "Revisiting global gene expression analysis". Cell, 151(3), 476 – 482, Oct. 2012.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[34] Secrier, M., and Schneider, R. "Visualizing time-related data in biology, a review". Briefings in Bioinformatics, Software Review, Apr. 2013.

[35] Wang, J., Peng, X., Xiao, Q., et al. "An effective method for refining predicted protein complexes based on protein activity and the mechanism of protein complex formation". BMC Systems Biology, 7:28, Mar. 2013.

[36] Wang, J., Peng, X., Li, M. "Active Protein Interaction Network and Its Application on Protein Complex Detection". In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 37 – 42, Nov. 2011.

[37] Li, M., Chen, W., Wang, J., et al. "Identifying dynamic protein complexes based on gene expression profiles and PPI Networks". Biomed Research International, Mar. 2014.

[38] Kim, Y., Han, S., Choi, S., et al. "Inference of dynamic networks using time-course data". Briefings in Bioinformatics, 15(2), 212 – 228, May 2013.

[39] Friedman, N., Linial, M., Nachman, I., et al. "Using Bayesian networks to analyze expression data". J. Comp. Biol., 7, 601 – 620, July 2004.

[40] Remondini, D., O'Connell, B., Intrator, N., et al. "Targeting c-Myc-activated genes with a correlation method: detection of global changes in large gene expression network dynamics". PNAS, 102, 6902 – 6906, May 2005.

[41] Song, L., Kolar, M., and Xing, E.P. "KELLER: estimating time-varying interactions between genes". Bioinformatics, 25, i128 – i136, June 2009.

[42] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., et al. "How to infer gene networks from expression profiles". Mol. Syst. Biol, 3:122, Feb. 2007.

[43] Morris, M.K., Saez-Rodriguez, J., Sorger, P.K., et al. "Logic-based models for the analysis of cell signaling networks". Biochemistry, 49, 3216 – 3224, Mar. 2010.

[44] Park, S., Yang, J.S., Shin, Y.E., et al. "Protein localization as a principal feature of the etiology and comorbidity of genetic diseases". Mol. Syst. Biol., 7:494, May 2011.

[45] De Lichtenberg, U., Jensen, L.J., Brunak, S., et al. "Dynamic complex formation during the yeast cell cycle". Science, 307, 724 – 727, Feb. 2005.

[46] Lee, Y.H., Tan, H.T., and Chung, M.C. "Subcellular fractionation methods and strategies for proteomics". Proteomics, 10, 3935 – 3956, Nov. 2010.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Erythritol promotes interfacial stability of GAPDH

Norbert W. Seidler[1] **and Jane M. Jones** [1]
[1] Department of Biochemistry, Kansas City University of Medicine and Biosciences, 1750 Independence Avenue, Kansas City, Missouri, USA

Abstract *- GAPDH is involved in glycolysis, but is also know to translocate to the nucleus and participate in reactions leading to the cell's demise. One of the key feature's to understanding the multifunctionality of GAPDH is to increase our understanding of the oligomeric dynamics of this protein. Though it acts as a tetramer in its role as an oxidoreductase in glycolysis occurring in the cytosol, it becomes translocated to the nucleus as a dimeric species under various stress inducers. Cellular metabolites or xenobiotics, acting as kosmotropes and chaotropes, would likely impact the conformational trajectories that ultimately decide the net subpopulations of stable and metastable oligomeric structures of GAPDH. We were interested in examining the effects of erythritol, a polyol with kosmotropic properties, and isoflurane, a volatile anesthetic agent with chaotropic properties. In addition from looking at activity and thermal stability, we assessed the effects of erythritol on the interfacial dynamics of GAPDH subunit-subunit interactions using molecular dynamics. We observed that erythritol stabilized the protein and that the computational analysis confirmed the experimental data.*

1 Introduction

The ability of polyols (i.e. compounds with multiple hydroxymethyl groups) to stabilize proteins and the mechanism by which this is accomplished have been known for some time [4]. The water-ordering ability of polyols is the driving force of protein stabilization, and the effects of polyols on protein-water interactions are far more influential than effects on exposed hydrophobic groups of unfolded protein [5]. The hydrophobic interaction within the protein appears to be strengthened by the polyol's preferential interaction with solvent (i.e. water molecules) [6]. The hydration shell that exists at the protein's surface topology is of particular importance in describing the manner by which polyols exert their effects [2,3]. The unfavorable interaction of polyols with the exposed nonpolar groups promotes a protein hydration that enhances stability as shown by the competing effects of the chaotropic agent guanidine hydrochloride [7]. We previously proposed that the preconditioning effects of volatile anesthetic agents may be due to their chaotropic (i.e. water-disordering) effects on proteins [8, 9] with the highly abundant multi-functional protein glyceraldehyde 3-phosphate dehydrogenase (GAPDH) acting as a pivotal signaling hub for comprehensive cellular communication [10]. In the present study, we were interested in the effects of erythritol, a food additive and four-carbon polyol, on GAPDH. We experimentally examined the competing effects of erythritol and the anesthetic agent isoflurane on the activity and thermal stability of GAPDH. We previously observed that isoflurane alters the interfacial dynamics of GAPDH, presumably by promoting the formation of a dimeric species that enters into a new configuration (i.e. decameric structure) [11]. Additionally, in the current study, we used computational approaches to examine the way erythritol interacts with GAPDH. Native GAPDH is an asymmetric homotetramer, consisting of identically sequenced polypeptide chains. Curiously, each chain is conformationally unique and is designated as either subunit -O, -P, -Q and -R, depending on its position in the tetramer. The subunits interact across three axes: *P*-, *R*- and *Q*-axis, conventionally italicized to distinguish them from the subunits (Figure 1).

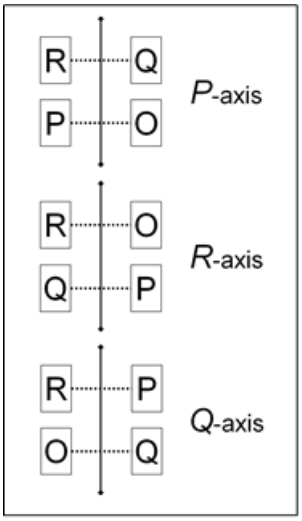


Figure 1: Schematic illustration of the subunit interfaces in GAPDH. The four types of subunits, each of which exhibits a unique conformation, are given as rectangles identified by the letters O, P Q and R. Each solid vertical line represents an interface with a pseudo-two-fold axis of symmetry. The horizontal dotted lines show the interactions of the respective subunits across the axes. The *R*-axis, for example, defines the interactions between subunits O and R as well as those between P and Q.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

GAPDH's multifunctionality may be in part due to the dynamic oligomeric properties of this molecule. Particularly of interest is the property of initiating the cell death cascade upon nuclear translocation, which is associated with stable dimer formation, presumably O-P or Q-R dimers since the *P*-axis is the most conserved and consists by far of the most interfacial contacts. Relevant to this point is the degree of stability at the *R*-axis, which also exhibits a fair number of interfacial contacts. one would consider the *Q*-axis as the most labile, and therefore we focused our interest in the dynamics at the *R*-axis.

We carefully examined a crystal structure (i.e. yeast GAPDH) in the public database that was co-crystalized with erythritol, in order to define the polyol binding site. We also performed molecular dynamics simulations using P-Q dimers of human GAPDH in the presence and absence of erythritol, in order to assess the effects of this polyol on protein stability.

2 Methods

2.1 Experimental Approach

Preparation of GAPDH - Rabbit GAPDH (Sigma Aldrich; cat. no. G2267) was prepared in a buffer containing 50mM sodium phosphate (pH 7.4) and 5mM EDTA and was filtered (Millipore Millix-HV, 0.45µm PVDF, pre-wetted with water and then buffer) to ensure the presence of 100% native (i.e. non-denatured) protein. Aged GAPDH was prepared by storage at 5°C for five days, which corresponds to the cellular half-life of this protein [12]. The protein sequence of rabbit GAPDH is 95% identical to that of human GAPDH.

Treatment of GAPDH - Samples of GAPDH (5µM), which were either freshly-prepared or aged, were incubated with and without erythritol (100mM) for 5min at room temperature prior to exposure to isoflurane (0.25mM for 10min at room temperature under mild agitation and shielded from light). Samples were then degassed prior to determination of enzyme activity or thermal instability. Data was analyzed using t-tests with 90% confidence limits.

Oxidoreductase activity of GAPDH - Enzyme activity was measured spectrophotometrically. Measurements were made after exposure of GAPDH to isoflurane (0.2mM for 10min at room temperature under slight agitation) with and without erythritol (100mM). Final assay cuvettes contained equivalent micromolar concentrations of erythritol.

Thermal instability of GAPDH - We previously showed that freshly-prepared GAPDH exhibits a Tm (i.e. temperature at which half of the protein is denatured) of 54.7°C, using a heating rate of 1°C per minute [13]. In the current study, the extent of heat denaturation of GAPDH samples was assessed following incubation at 46°C for 5min. Denaturation was measured by optical density at 450nm using a spectrophotometer compared to controls: 0% denatured (i.e. no heat exposure) and 100% denatured (i.e. exposed to 95°C).

2.2 Computational Approach

Erythritol binding site in GAPDH - We examined the structure database for GAPDH molecules co-crystalized with erythritol. Of the approximately 150 crystal structures of GAPDH in the Protein Data Bank (i.e. accessible through http://www.ncbi.nlm.nih.gov/Structure) only one structure was found that contained the polyol erythritol. Several structures contain smaller hydroxyl-containing molecules (i.e. glycerol and 1,2-ethanediol). The structure that contained erythritol is the yeast-derived GAPDH (isoform 3 from *Saccharomyces cerevisiae*). The ID numbers for this structure are 96311 (i.e. MMDB ID) and 3PYM (i.e. PDB ID). The 3PYM crystal structure contains tetrameric GAPDH, four molecules of NAD^+, four erythritols and two Na^+ atoms. We examined the erythritol binding site, identifying amino acid residues in close proximity to this region. Yeast GAPDH (i.e., 3PYM.pdb, by sequence, is 65% identical and 79% homologous to human GAPDH (i.e., 1UBF.pdb).

Molecular dynamics - We used ChemBioDraw Ultra 12.0 (www.cambridgesoft.com) to generate erythritol structures and ChemBio3D Ultra 12.0 to perform molecular dynamics simulations with the human GAPDH structures (i.e. 1U8F.pdb) obtained from NCBI. The tetrameric GAPDH was copied to the window and the subunits O and R (with its corresponding ligand NAD^+) were deleted, leaving subunits P and Q, each with its bound NAD^+). All water molecules were kept (i.e. 911 solvent molecules). The file was duplicated and erythritol was inserted in one of the files. We used the MM2 force field method. Energy minimization was performed prior to each molecular dynamics simulation. The step and frame intervals were 1 and 10fs, respectively. The heating rate was 1.0kcal/atom/ps and the target temperature was set at 300°K. We chose to investigate the distances across the *R*-axis, since residues in this area were identified as those found in the erythritol binding site in yeast GAPDH. The interacting partners (i.e. amino acid residues from the P-subunit that interact with residues from the Q-subunit), which we used in our computations, were taken from [14]. We obtained inter-residue distances of the original P-Q dimer crystal structure, as well as at the start of the simulations, after energy minimization. We compared these distances with those taken after 100 iterations and after 1000 iterations.

3 Results

3.1 Experimental Effects of Erythritol on GAPDH

In experimentally measuring the effects of erythritol on oxidoreductase activity of rabbit GAPDH, we observed that pre-incubation of GAPDH with erythritol had no effect on the activity of either freshly-prepared (Figure 2A) or aged

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

control samples. Isoflurane also had no effect on the activity of freshly-prepared GAPDH (Figure 2A). However, when aged GAPDH was exposed to isoflurane under the same conditions, we observed a significant reduction in enzymatic activity (Figure 2B). Erythritol protected aged GAPDH from the deleterious effects of isoflurane (Figure 2B). Interestingly, the control activity of freshly prepared GAPDH exceeded aged GAPDH (24nmol/min per mg prot ±1.2 vs 22 ±1.4 for freshly-prepared and aged samples, respectively; df = 4; p = 0.097) when using a 90% confidence limit.

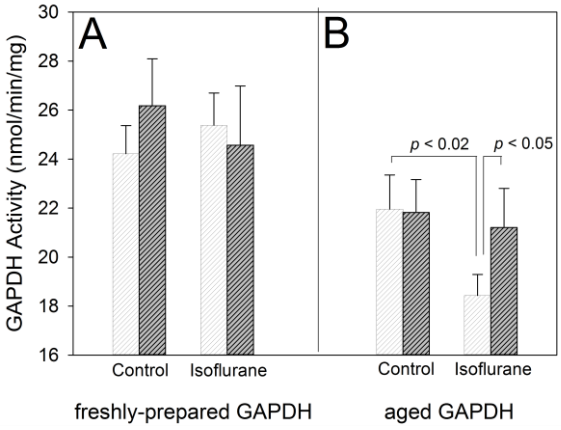


Figure 2. Effects of erythritol and isoflurane on the enzymatic activity of freshly-prepared and aged GAPDH. Samples of GAPDH, which were either freshly-prepared (A) or aged (B), were incubated with (gray-hashed bars) and without (white-hashed bars) erythritol prior to exposure to isoflurane. Data represent mean ± SD of three measurements.

Erythritol prevented heat denaturation of GAPDH under all conditions that were tested (Figure 3). In contrast, isoflurane increased the susceptibility to heat denaturation of both freshly-prepared and aged GAPDH. Upon incubation with erythritol, GAPDH was protected from the deleterious effects of isoflurane in both freshly-prepared and aged samples. Aged GAPDH was more susceptible to heat denaturation than the GAPDH of freshly-prepared samples (17.0% denaturation ±0.00 vs 22.4 ±0.21, for freshly-prepared and aged samples, respectively; df = 2; $p < 0.0005$), consistent with the effects of *in vitro* aging on enzyme activity (Figure 2).

Curiously, the effects of isoflurane and age (i.e. storage of protein) were additive, suggest that they both negatively affect the conformational integrity of GAPDH. In assessing both activity and susceptibility to heat denaturation in the presence of the anesthetic agent, isoflurane, we saw that the aged protein is particularly labile to the chaotropic properties of this agent. The implications of these observations would suggest that proteins in the elderly (given, of course, that protein turnover and renewal is compromised - a common feature of organismal aging) may be at greater risk of structural and functional alteration. We intend to pursue this concept further.

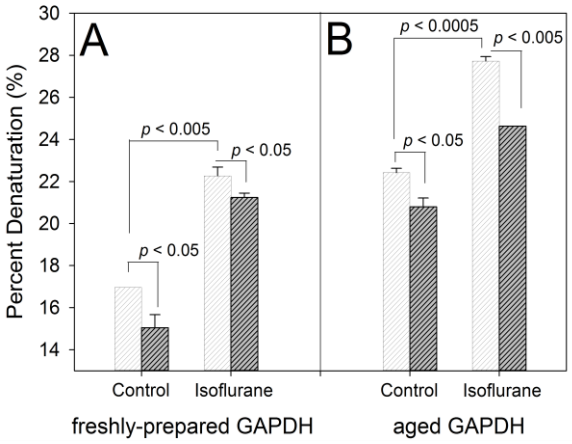


Figure 3. Effects of isoflurane and erythritol on thermal instability of freshly-prepared and aged GAPDH. Freshly-prepared (A) or aged (B) GAPDH was incubated with (gray hashed bars) and without (white hashed bars) erythritol prior to exposure to isoflurane). Data represent mean ± SD of two measurements.

3.2 Erythritol Binding Site

The only erythritol-containing GAPDH crystal structure in the NCBI/Structure database (i.e. 3PYM.pdb) contains tetrameric GAPDH, four NAD^+, four erythritols and two Na^+ atoms. The structure was schematically re-drawn (Figure 4) to show the four subunits and the binding relationship of the four erythritols to these subunits.

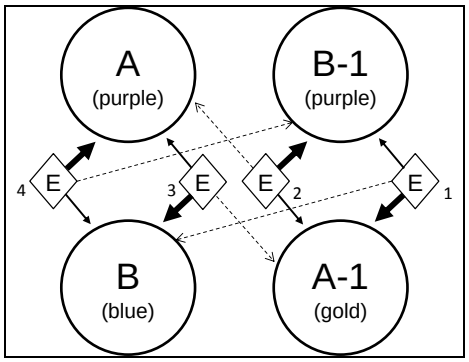


Figure 4. Schematic illustration of yeast GAPDH with bound erythritol. The four subunits (given as circles) are identified as indicated in the database (3PYM.pdb). The four erythritols (given as diamonds) are numbered. Each erythritol molecule forms interactions with three different subunits as indicated by different-sized arrows to reflect the degree of contact.

The four subunits of yeast GAPDH (i.e. 3PYM.pdb) can be compared to the rabbit or human GAPDH. The A, A-1, B and

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

B-1 subunits (Figure 4) in yeast GAPDH match to the conventional O, P, Q and R subunits in human GAPDH. Hence, erythritol molecules #2 and #4 would be expected to stabilize the interactions between subunits A and B-1 (or, by analogy, O and R), that is to say, across the *R*-axis. And, erythritols #1 and #3 would stabilize the interactions between subunits B and A-1 (or, by analogy, Q and P), likewise, across the *R*-axis. We were interested in identifying the amino residues that constitute the erythritol binding site. We looked at each of these sites, using Cn3D 4.3.1 structure viewer. After highlighting one of the erythritols, we selected to identify all residues within 7Å, which approximates the H-bonding distance between atoms (i.e. those of erythritol and the protein) with only one interconnecting water molecule. In this manner, we were able to determine the amino acid residues and their subunits that represent the cavity in which erythritol sits. We followed this procedure for each of the four erythritol molecules and tabulated the results in Table 1.

Table 1. Identity of amino acid residues at the erythritol binding site. The residues at each of the four erythritol binding sites are given.

Erythritol Molecule	Subunit A-1 (gold)	Subunit B-1 (purple)	Subunit A (purple)	Subunit B (blue)
1	Asn38-Met44 Lys46-Tyr47 Glu57-His60	Thr275-Ala278	-	Trp194
2	Thr275 Asp277-Ala278	Asp39Tyr47 Arg53 Gly56-His60	Trp194	-
3	Trp194	-	Thr275 Asp277-Ala278	Asp39-Tyr47 Arg53 Gly56-His60
4	-	Trp194	Asn38-Met44 Lys46-Tyr47 Glu57-His60	Thr275-Ala278

There are a total of 21 different residues at the erythritol binding site with some slight differences among the sites. These residues represent three distinct regions of the protein (i.e. aa38-60; aa194; and aa275-278), which is the yeast ortholog. We compared the erythritol binding site in yeast GAPDH with the corresponding sequence in human GAPDH (Figure 5). Curiously, the overall homology between the full sequences is 79%, while the homology of the 21 residues that make up this site, drops to 62%, suggesting evolutionary pressures may have contributed to a significant change in ability to bind erythritol. Given this observation, we would expect that erythritol binding to human GAPDH is decreased compared to its affinity to yeast GAPDH. Conversely, human GAPDH may bind erythritol with greater affinity. In the next section, we describe the role of amino acid residues Gly193, Asp39 and Asn41 of human GAPDH in binding erythritol as evidenced by molecular dynamics simulations.

```
yeast  36  ~ITNDYAAYMFKYDSTHGRYAGEVSHDD~ 62
           I  +Y  YMF+YDSTHG++ G V  ++
human  38  ~IDLNYMVYMFQYDSTHGKFHGTVKAEN~ 64

yeast 192  ~KDWRG~ 196
           K WR
human 194  ~KLWRD~ 198

yeast 275  ~YTEDAVV~ 278
           YTE  VV
human 277  ~YTEHQVV~ 278
```

Figure 5: Comparison of the erythritol binding site residues. The sequences of yeast GAPDH (i.e., 3PYM.pdb) are compared with those of human GAPDH (i.e. 1U8F.pdb). The 21 residues are shaded and bolded. Indentities are indicated in the intervening line. Similar amino acids are given a plus symbol (+) to show homology.

3.3 Molecular Dynamics Simulations

We used molecular dynamics simulations to examine whether or not erythritol would interact with human GAPDH and whether it would have a stabilizing effect. Since we were particularly interested in the R-axis, we started with a P-Q dimer of human GAPDH (i.e. 1U8F.pdb). There are 31 P-subunit residues (Figure 6) that make contact (i.e. largely reciprocal interactions) with 32 residues on the Q-subunit. We measured 75 inter-subunit distances: (a) in the original crystal state, (b) after energy minimization, and (c) after 1000 iterations.

```
MGKVKVGVNGFGRIGRLVTRAAFNSGKVDI 30
VAINDPFIDLNYMVYMFQYDSTHGKFHGTV 60
KAENGKLVINGNPITIFQERDPSKIKWGDA 90
GAEYVVESTGVFTTMEKAGAHLQGGAKRVI 120
ISAPSADAPMFVMGVNHEKYDNSLKIISNA 150
SCTTNCLAPLAKVIHDNFGIVEGLMTTVHA 180
ITATQKTVDGPSGKLWRDGRGALQNIIPAS 210
TGAAKAVGKVIPELNGKLTGMAFRVPTANV 240
SVVDLTCRLEKPAKYDDIKKVVKQASEGPL 270
KGILGYTEHQVVSSDFNSDTHSSTFDAGAG 300
IALNDHFVKLISWYDNEFGYSNRVVDLMAH 330
MASKE 335
```

Figure 6: Human GAPDH sequence (Uniprot P04406) highlighting the residues at the R-axis interface. The amino acid residues (shown black and gray shaded) were taken from [14], and are defined as those P-subunit residues that are within 5Å of the neighboring Q-subunit, as determined by structural analysis of lobster GAPDH (see Table 7.2 in [15]). Three of these human GAPDH residues (shown gray-shaded) are homologous to the lobster ortholog. Interestingly, the amino acid residues that are defined as the erythritol binding site (See Figure 5), overlap these regions.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Notice that there are four regions where the protein folds in the direction of, and becomes associated with, the P/Q interface: region 1 spans from residue 12 to 16; region 2, from 38 to 52; region 3, from 181 to 205; and region 4, from 236 to 238. Also residue 317 on the Q-subunit is involved in the interactions. Inter-subunit distances were measured before and after molecular dynamics simulations (Table 2). The data was grouped according to the regions described in the text. In determining the averages, the distances from regional interface 3/1, 3/2, 3/3 and 3/4 involved 12, 39, 17 and 7 inter-subunit distances, respectively. The means and SEM of these respective data points are presented in angstrom (Å). Two-tailed paired t-tests were performed with consideration of an α value of 0.10, comparing the start distances with those after 1000 iterations. Using an α value of 0.10 for statistical significance increases the probability of a Type I error (i.e. that the claim of a change in inter-facial distance is false). But, this gave us a better likelihood of not missing a difference. We think that this design is appropriate particularly given that only main chain α-carbons were used as measurement posts rather than all atoms of the side chains of interfacial residues.

Table 2: Erythritol prevents the separation of the subunits at the regional interfaces across the *R*-axis.

Regional Interface	Distance at the Start	Distance after 1000 Iterations	
		Control	Erythritol
3/1	8.8 ±0.46	8.2 ±0.54 $p < 0.01$	8.4 ±0.60 Not significant
3/2	8.6 ±0.21	9.1 ±0.31 $p = 0.085$	8.5 ±0.24 Not significant
3/3	7.1 ±0.29	8.3 ±0.38 $p < 0.01$	7.5 ±0.30 Not significant
3/4	8.9 ±0.63	9.5 ±0.71 $p = 0.051$	10.3 ±0.51 $p < 0.05$

Interestingly, when we carefully looked at the location of erythritol binding after 1000 interactions, we found that the erythritol molecule was bound at the regional interface that included Q-subunit (i.e. region 3) and P-subunit (i.e. region 2). Therefore, we measured the distances across this regional interface as well as the reciprocal set of interactions (i.e. P-subunit region 3 to Q-subunit region 2) to assess erythritol's effect (Table 3). Since the interfacial interactions exhibit a reciprocal property (i.e. P region 3 to Q region 2 on one side and Q region 3 to P region 2 on the flip side), and since the single erythritol molecule bound to one of these locations, the unoccupied site represents the control, to which the site that binds erythritol can be compared.

Table 3: Bilateral effects of a single molecule of erythritol.

Reciprocal Interactions (region 3 to region 2)	Distance after 1000 Iterations		Statistical Comparison
	Control	Erythritol	
Q-subunit (region 3) to P-subunit (region 2)	9.5 ± 0.39	7.8 ± 0.31	$p < 0.001$
P-subunit (region 3) to Q-subunit (region 2)	8.7 ± 0.49	9.2 ± 0.29	$p = 0.209$ Not Significant

The data are presented in angstrom (Å) as means ±SEM of 19 distances (i.e. across Q region 3 to P region 2) and 18 distances (i.e. across P region 3 to Q region 2). Note that the side containing bound erythritol showed significance difference, suggesting that erythritol not only prevented separation of the subunits, but enhanced its stability. This bilateral activity is indicative of erythritol's kosmotropic properties.

We also determined which specific atoms may be involved in binding erythritol to the Q region 3/P region 2 bilateral side of the *R*-axis (Figure 7).

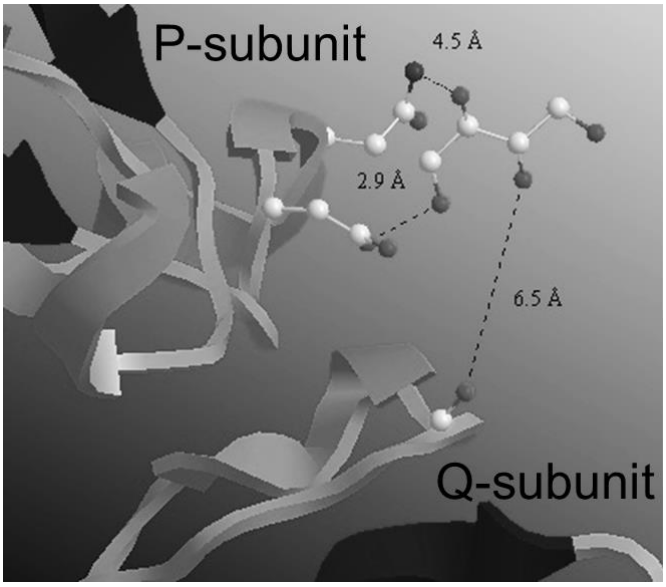


Figure 7: Interactions of erythritol and the P-Q dimer of GAPDH. The image illustrates the subunit interactions with erythritol following molecular dynamics simulation (i.e. 1000 interations). The solvent water molecules were left out for clarity. The .jpg color image was converted to grayscale and brightness and contrast was adjusted to improve visibility.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

The amino acid residue Gly193 (specifically, the main chain carbonyl oxygen atom) on the Q-subunit is 6.5 Å from the C-2 (i.e. second carbon) hydroxyl oxygen atom of erythritol. The amino acid residue Asp39 (specifically, the side chain oxygen atom) on the P-subunit is 2.9 Å from the C-4 (i.e. fourth carbon) hydroxyl oxygen atom of erythritol.

The amino acid residue Asn41 (specifically, the side chain amide nitrogen atom) on the P-subunit is 4.5 Å from the C-3 (i.e. third carbon) hydroxyl oxygen atom.
Interestingly, a water molecule (#279) was found positioned between Gly193 (i.e. main chain oxygen) and erythritol (i.e. C-2 hydroxyl oxygen), specifically 2.1 and 3.8 Å.

Additionally, two other water molecules (#384 and #304) were positioned 4.7 and 3.5 Å from the C-3 hydroxyl oxygen of erythritol.

4 Discussion

Molecular dynamics simulations with chymotrypsin inhibitor protein and polyols, such as xylitol, offered a molecular basis of protein stabilization [1]. The authors state that there is a preferential hydration at the surface of the protein and that the polyols cluster at a distance beyond this initial water shell, concluding that the water structure becomes more ordered (i.e. polyols decrease the entropy of water in the first shell of hydration). Their studies involved simulation at 363°K, where they could observe thermal unfolding, which was prevented by the polyols. They did not examine the effects of erythritol. In the present study, we also confirmed experimentally that erythritol prevented thermal instability of GAPDH (Figure 3). The stabilization of GAPDH also may be due in part to the prevention of the subunits to disengage, as evidenced by the molecular dynamics simulations using P-Q dimer of GAPDH and a single erythritol molecule (Table 2). In the control simulations, the regional interface of residues 12 to 16 (i.e. region 1) with residues 181 to 205 (i.e. region 3) exhibited on average a decreased distance, while all other regional interfaces exhibited an increased average distance after 1000 iterations, reaching a temperature below 300°K. In the erythritol simulations, the average distances across these regional interfaces did not change with the exception of the reciprocal distances associated with regions 3 and 4, which increased. Of the total 63 amino acid residues that are involved in inter-subunit contact at the *R*-axis, there are only 10 residues at the 3/4 regional interface. In effect, 53 residues have remained unchanged relative to one another in the P-Q dimer with a single molecule of erythritol.

5 Conclusion

Our computational and experimental approaches to investigating the protein stabilizing effects of the polyol, erythritol, have corroborated one another. Erythritol prevented heat denaturation of GAPDH using an experimental approach at the bench, and erythritol as a single molecule prevented the separation of a single P-Q dimer of GAPDH in molecular dynamics simulations. our findings support the concept that erythritol behaves as a kosmotrope and that isoflurane exhibits chaotropic properties.

The authors gratefully acknowledge the efforts of Sharon White in her clerical role in preparing this manuscript.

6 References

[1] Liu FF, Ji L, Zhang L, Dong XY, Sun Y. Molecular basis for polyol-induced protein stability revealed by molecular dynamics simulations. J Chem Phys 2010;132(22):225103.

[2] Ortbauer M, Popp M. Functional role of polyhydroxy compounds on protein structure and thermal stability studied by circular dichroism spectroscopy. Plant Physiol Biochem 2008;46(4):428-34.

[3] Shimizu S, Smith DJ. Preferential hydration and the exclusion of cosolvents from protein surfaces. J Chem Phys 2004;121(2):1148-54.

[4] Back JF, Oakenfull D, Smith MB. Increased thermal stability of proteins in the presence of sugars and polyols. Biochemistry 1979;18(23):5191-6.

[5] Gekko K, Morikawa T. Thermodynamics of polyol-induced thermal stabilization of chymotrypsinogen. J Biochem 1981;90(1):51-60.

[6] Gekko K. Calorimetric study on thermal denaturation of lysozyme in polyol-water mixtures. J Biochem 1982;91(4):1197-204.

[7] Gekko K, Ito H.Competing solvent effects of polyols and guanidine hydrochloride on protein stability. J Biochem 1990;107(4):572-7.

[8] Baker MR, Benton SK, Theisen CS, McClintick CA, Fibuch EE, Seidler NW Isoflurane's effect on protein conformation as a proposed mechanism for preconditioning. Biochemistry Res Int 2011;2011:739712.

[9] Ferns JE, Theisen CS Fibuch EE, Seidler NW Protection against protein aggregation by alpha-crystallin as a mechanism of preconditioning. Neurochem Res 2012;37(2):244-252.

[10] Seidler NW. GAPDH in Anesthesia. Adv Exp Med Biol 2013;985:269-291

[11] Pattin AE1, Ochs S, Theisen CS, Fibuch EE, Seidler NW. Isoflurane's effect on interfacial dynamics in GAPDH

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

influences methylglyoxal reactivity. Arch Biochem Biophys 2010;498(1):7-12.

[12] Seidler NW. Basic biology of GAPDH. Adv Exp Med Biol 2013;985:1-36.

[13] Seidler NW, Yeargans GS. Effects of thermal denaturation on protein glycation. Life Sci 2002;70(15):1789-99.

[14] Moras D, Olsen KW, Sabesan MN, Buehner M, Ford GC, Rossmann MG. Studies of asymmetry in the three-dimensional structure of lobster D-glyceraldehyde-3-phosphate dehydrogenase. J Biol Chem 1975;250(23):9137-62.

[15] Seidler NW. Dynamic oligomeric properties. Adv Exp Med Biol 2013;985:207-247

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Computational Methods for Protein-Protein Interaction Prediction

Maad Shatnawi
College of Information Technology, United Arab Emirates University, Al-Ain, Abu Dhabi, UAE
E-mail: shatnawi@uaeu.ac.ae

Abstract— *Protein-protein interactions (PPI) occur at almost every level of cell functions. The identification of interactions among proteins provides a global picture of cellular functions and biological processes. It is also an essential step in the construction of PPI networks for human and other organisms. PPI prediction has been considered to be a promising alternative to the traditional drug design techniques. The identification of possible viral-host protein interactions can lead to a better understanding of infection mechanisms and, in turn, to the development of several medication drugs and treatment optimization. This paper investigates most of the relevant existing computational approaches carried out towards this perspective and provides a comparison of these approaches. Technical challenges, unresolved issues, and future research directions in this area are also discussed.*

Keywords: Protein-protein interaction prediction, PPI, protein sequences, computational techniques

1. Introduction

Proteins are the building blocks of all living organisms. A protein is a polypeptide which is a linear polymer of several amino acids (AAs) connected by peptide bonds. The primary structure of a protein is the linear sequence of its AAs. The secondary structure of a protein is the general three-dimensional form of its local parts without displaying specific atomic positions, which are considered to be a tertiary structure. Domains are the basic functional units of protein tertiary structures. A protein interacts with other proteins in order to perform certain functions and tasks. Protein-protein interaction (PPI) occurs at almost every level of cell functions. The identification of interactions among proteins provides a global picture of cellular functions and biological processes. Since most biological processes involve one or more protein-protein interactions (PPIs), precisely identifying the set of interacting proteins in an organism is very useful for deciphering the molecular mechanisms underlying given biological functions and for assigning functions to unknown proteins based on their interacting partners [1]–[3]. Protein interaction prediction is also a fundamental step in the construction of PPI networks for human and other organisms. The identification of specific disease-related protein interactions can lead to the development of a number of medication drugs. Abnormal PPIs have implications in several neurological disorders such as Creutzfeld-Jacob and Alzheimer [4]–[6]. Therefore, the development of accurate and reliable methods for identifying PPIs has very important impacts in several protein research areas.

In this paper we explore the two main categories of computational PPI prediction methods; sequence-based and structure-based methods. We provide a comprehensive comparative study of the existing approaches in PPI prediction and discuss the technical challenges and unresolved issues in this field. The rest of this paper is organized as follows. Next section addresses the key technical challenges that face PPI prediction and the open issues in this field. Section 3 discusses the evaluation measures that are typically used in PPI prediction. In Section 4, which is the focus of our paper, comprehensive description and comparison are presented for the most current computational PPI prediction methods. Concluding remarks are presented in Section 5.

2. Technical Challenges and Open Issues

There are several technical challenges that face computational analysis of protein sequences in general and PPI prediction in particular. First, there have been a huge amount of newly discovered protein sequences in the past genomic era. Second, Protein chains are typically long which are difficult, time-consuming, and expensive to characterize by experimental methods. Third, the availability of large, comprehensive, and accurate benchmark datasets is required for the training and evaluation of prediction methods.

One of the challenges of protein prediction methods is protein representation. Protein prediction methods vary in protein representation and feature extraction in order to build their classification models. There are two kinds of models that were generally used to represent protein samples; the sequential model and the discrete model. The most and simplest sequential model for a protein is its entire AA sequence. However, this representation does not work well when the query protein does not have high sequence similarity to any attribute-known proteins. Several non-sequential models, or discrete models, have been proposed. The simplest discrete model is AA composition which is the normalized occurrence frequencies of the twenty native amino acids in a protein. However, all the sequence-order knowledge will be lost using this representation which, in turn, will negatively affect the prediction quality [7]. Some approaches use AA physiochemical properties. Other approaches use pairwise similarity. Some approaches are template-based while others are statistical-based.

There are various challenges that face machine-learning (ML) protein prediction methods. Selecting the best ML approach is a great challenge. There is a variety of techniques that diverse in accuracy, robustness, complexity, computational cost, data diversity, over-fitting, and ability to deal with missing attributes and different features. Most ML approaches of protein sequences are computationally expensive and often suffer from low prediction accuracy. They are further susceptible to overfitting [8].

Furthermore, most PPI prediction approaches have achieved reasonable performance on balanced datasets con-

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

taining equal number of interacting and non-interacting protein pairs. However, this ratio is highly unbalanced in nature and these approaches have not been comprehensively assessed with respect to the effect of the large number of non-interacting pairs in realistic datasets. In addition, since highly unbalanced distributions usually lead to large datasets, more efficient prediction methods are required to handle such challenging tasks.

3. Evaluation Metrics

There are several assessment measures that are used to evaluate a PPI predictor and compare it with other approaches. The most frequently used evaluation metrics in this field are accuracy, sensitivity, specificity, precision, F1, and MCC. Accuracy ($Ac = \frac{TP+TN}{TP+TN+FN+FP}$, where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively) is defined as the proportion of correctly predicted interacting and non-interacting protein pairs to all of the protein pairs listed in the dataset. Sensitivity, or recall ($R = \frac{TP}{TP+FN}$), is defined as the proportion of correctly predicted interacting protein pairs to all of the interacting protein pairs listed in the dataset. Precision ($P = \frac{TP}{TP+FP}$) is defined as the proportion of correctly predicted interacting protein pairs to all of the predicted interacting protein pairs. Specificity ($Sp = \frac{TN}{TN+FP}$) is defined as the proportion of correctly predicted non-interacting protein pairs to all the non-interacting protein pairs listed in the dataset. The F-measure ($F1 = \frac{2*P*R}{P+R}$) is an evaluation metric that combines precision and recall into a single value and is defined as the harmonic mean of precision and recall [9], [10]. Mathew correlation coefficient ($MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$) is a measure that balances prediction sensitivity and specificity. MCC ranges from -1, indicating an inverse prediction, through 0, which corresponds to a random classifier, to +1 for perfect prediction.

4. Computational Approaches

PPI prediction has been studied extensively by several researchers and a large number of approaches have been proposed. These approaches can be classified into physiochemical experimental and computational approaches. Physiochemical experimental techniques identify the physiochemical interactions between proteins which, in turn, are used to predict the functional relationships between them. These techniques include; yeast two-hybrid based methods [11], mass spectrometry [12], Tandem Affinity Purification [13], protein chips [14], and hybrid approaches [15]. Although these techniques have succeeded in identifying several important interacting proteins in several species such as Yeast, Drosophila, and Helicobacter-pylori [16], they are computationally expensive and significantly time consuming, and so far the predicted PPIs have covered only a small portion of the complete PPI network. As a result, the need of computational tools has been increased in order to validate physiochemical experimental results and to predict non-discovered PPIs [1], [17].

Several computational methods have been proposed for PPI prediction and can be categorized according to the used protein features into sequence-based and structure-based methods. Sequence-based methods utilize AA features while structure-based methods use three-dimensional structural features [18]. This section provides an overview and discussion of some of the current computational sequence-based and structure-based PPI prediction approaches.

4.1 Sequence-Based Approaches

Sequence-based PPI prediction methods utilize AA features such as hydrophobicity, physiochemical properties, evolutionary profiles, AA composition, AA mean, or weighted average over a sliding window [18]. This section presents and evaluates some of the existing sequence-based approaches.

4.1.1 PIPE

Protein-protein Interaction Prediction Engine (PIPE) was developed by Pitre *et al.* [19] based on the assumption that interactions between proteins occur by a finite number of short polypeptide sequences observed in a database of known interacting protein pairs. These sequences are typically shorter than the classical domains and reoccur in different proteins within the cell. PIPE uses protein primary structure to estimate the likelihood of an interaction between a pair of the yeast Saccharomyces cerevisiae proteins by measuring the reoccurrence of these short polypeptides within known interacting proteins pairs. To determine whether two proteins A and B interact, the two query proteins are scanned for similarity to a database of known interacting proteins pairs. For each known interacting pair (X, Y), PIPE uses sliding windows to compares the AA residues in protein A against that in X and protein B against Y, and then measures how many times a window of protein A finds a match in X and at the same time a window in protein B matches a window in Y. These matches are counted and added up in a 2D matrix. A positive protein interaction is predicted when the reoccurrence count in certain cells of the matrix exceed a predefined threshold value. PIPE was evaluated on a randomly selected set of 100 interacting yeast protein pairs and 100 non-interacting proteins from the database of interacting proteins (DIP) [20] and MIPS [21] databases. PIPE showed a prediction sensitivity of 0.61 and specificity of 0.89.

Since PIPE is based on protein primary structure information without any previous knowledge about the higher structure, domain composition, evolutionary conservation or the function of the target proteins. It can identify interactions of protein pairs for which limited structural information is available. The limitations of PIPE are as follows. PIPE is computationally intensive and requires hours of computation per protein pair as it scans the interaction library repeatedly every time. Second, PIPE shows weakness in detecting novel interactions among genome wide large-scale datasets as it reported a large number of false positives. Third, PIPE was evaluated on uncertain data of interactions that were determined using several methods, each having a limited accuracy.

Pitre *et al.* [22] then developed PIPE2 as an improved and more efficient version of PIPE which showed a specificity of 0.999. PIPE2 represents AA sequences in a binary code

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

which speeds up searching the similarity matrix. Unlike the original PIPE that scans the interaction database repeatedly every time, PIPE2 pre-computes all window comparisons in advance and stores them on a local disk.

Although PIPE2 achieves a high specificity, it has a large number of false positives with a sensitivity of 0.146 only. False positives rate can be reduced by incorporating other information about the target protein pairs including sub-cellular localization or functional annotation. A major limitation of PIPE2 is that it relies exclusively on a database of pre-existing interaction pairs for the identification of re-occurring short polypeptide sequences and in the absence of sufficient data, PIPE2 will be ineffective. PIPE2 is also less effective for motifs that span discontinuous primary sequence as it does not account for gaps within the short polypeptide sequences.

4.1.2 Auto Covariance

Guo *et al.* [23] proposed a sequence-based method using Auto Covariance (AC) and Support Vector Machines (SVM). AA residues were represented by seven physicochemical properties. These properties are hydrophobicity, hydrophilicity, volumes of side chains, polarity, polarizability, solvent-accessible surface area, and net charge index of AA side chains. AC counts for the interactions between residues a certain distance apart in the sequence. AA physicochemical properties were analyzed by AC based on the calculation of covariance. A protein sequence was characterized by a series of ACs that covered the information of interactions between each AA and its 30 vicinal AAs in the sequence. Finally, a SVM model with a Radial Basis Function (RBF) kernel was constructed using the vectors of AC variables as input. The optimization experiment demonstrated that the interactions of one AA and its 30 vicinal AAs would contribute to characterizing the PPI information. The software and datasets are available at http://www.scucic.cn/Predict_PPI/index.htm. A dataset of 11,474 yeast PPIs extracted from DIP [24] was used to evaluate the model and the average prediction accuracy, sensitivity, and precision achieved are respectively 0.86, 0.85, and 0.87.

AC includes the information of interactions between AAs a longer distance apart in the sequence taking the neighboring effect into account. The long-range interactions are important for representing the PPI information. The use of SVM as a predictor is another advantage. SVM is the state of the art ML technique and has many benefits and overcomes many limitations of other techniques. SVM has strong foundations in statistical learning theory [25] and has been successfully applied in various classification problems [26]. SVM offers several related computational advantages such as the lack of local minima in the optimization [27].

4.1.3 Pairwise Similarity

Zaki *et al.* [1] proposed a PPI predictor based on pairwise similarity and protein primary structure. Each protein sequence is represented by a vector of pairwise similarities against large AA subsequences created by a sliding window which passes over concatenated protein training sequences. Each coordinate of this vector is the E-value of the Smith-Waterman (SW) score [28]. These vectors are then used to compute the kernel matrix which is exploited in conjunction with a RBF-kernel SVM. Two proteins may interact by the means of the scores similarities they produce [29], [30]. Each sequence in the testing set is aligned against each sequence in the training set, count the number of positions that have identical AAs, and then divide by the total length of the alignment.

The method was evaluated on 100 interacting yeast *Saccharomyces cerevisiae* protein pairs within 157 proteins and 100 non-interacting protein pairs within 77 proteins from the DIP [20] and MIPS [21] databases. The method achieved a sensitivity of 0.81 and a specificity of 0.744.

SW alignment score provides a relevant measure of similarity between proteins. Therefore protein sequence similarity typically implies homology, which in turn may imply structural and functional similarity [31]. SW scores parameters have been optimized over the past two decades to provide relevant measures of similarity between sequences and they now represent core tools in computational biology [32]. The use of SVM as a predictor is another advantage. This work can be improved by combining knowledge about gene ontology, inter-domain linker regions, and interacting sites to achieve more accurate prediction.

4.1.4 AA Composition

Roy *et al.* [33] examine the role of amino acid composition (AAC) in PPI prediction and it performance against well-known features such as domains, tuple feature, and signature product feature. Every protein pair is represented by AAC and domain features. AAC is represented by monomer and dimer features. Monomer features capture composition of individual amino acids, whereas dimer features capture composition of pairs of consecutive AAs. To generate the monomer features, a 20-dimensional vector representing the normalized proportion of the 20 AAs in a protein is created. The real-valued composition is then discretized into 25 bits producing a set of 500 binary features. To generate the dimer features, a 400-dimensional vector of all possible AA pairs were extracted from the protein sequence and discretized into 10 bits producing a set of 4000 binary features. The domains are represented as binary features with each feature identified by a domain name. To compare AAC against other non-domain sequence-based features, tuple features [34] and signature products [35] were obtained. The tuple features were created by grouping AAs into six categories based on their biochemical properties, and then creating all possible strings of length 4 using these categories. The signature products were obtained by first extracting signatures of length 3 from the individual protein sequences. Each signature consists of a middle letter and two flanking AAs represented in alphabetical order. Thus two 3-tuples with the first and third amino acid letter permuted have the same signature. The signatures were used to construct a signature kernel specifying the inner product between two proteins.

The proposed approach was examined using three ML classifiers (logistic regression, SVM, and the Naive Bayes) on PPI datasets from yeast, worm and fly. Three datasets for yeast *S. cerevisiae* were extracted from the General Repository for Interaction Datasets (GRID) database [36], TWOHYB (Yeast Two-hybrid), AFFMS (Affinity pull down

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

with mass spectrometry), and PCA (protein complementation assay). In addition to that, a dataset each for worm, *C. elegans* (Biogrid dataset) [37] and fly, *D. melanogaster* [36] were used. The authors reported that AAC features performed almost equivalent contribution as domain knowledge across different datasets and classifiers which indicated that AAC captures significant information for identifying PPIs.

Most computational PPI prediction methods use domain knowledge as they capture conserved information of interaction surfaces. However, approaches relying on domains as features cannot be applied to proteins without any domain information. On contrast, AAC is simple feature, computationally cheep, applicable to any protein sequence, and can be used when there is lack of domain information. AAC can be used alone or combined with other features to predict new and validate existing interactions.

4.1.5 AA Triad

Yu *et al.* [38] proposed a probability-based approach of estimating triad significance to alleviate the effect of AA distribution in nature. The relaxed variable kernel density estimator (RVKDE) [39] is employed to predict PPIs based on AA triad information. The method is summarized as follows. Each protein sequence is represented as AA triads by considering every three continuous residues in the protein sequence as a unit. To reduce feature dimensionality vector, the 20 AA types were categorized into seven groups based on their dipole strength and side chain volumes [16]. The method then scans triads one by one along the sequence, and each scanned triad is counted in an occurrence vector, O. Subsequently, a significance vector, S, is proposed to represent a protein sequence by estimating the probability of observing less occurrences of each triad than the one that is actually observed in O. Each PPI pair is then encoded as a feature vector by concatenating the two significance vectors of the two individual proteins. Finally, the feature vector is used to train a RVKDE PPI predictor. The method was evaluated on 37,044 interacting pairs within 9,441 proteins from the Human Protein Reference Database (HPRD) [40], [41]. Datasets with different positive-to-negative ratios (from 1:1 to 1:15) were generated with the same positive instances and distinct negative sets, which are obtained by randomly sampling from the negative instances. The authors concluded that the degree of dataset imbalance is important to PPI predictor behavior. With 1:1 positive-to-negative ratio, the proposed method achieves 0.81 sensitivity, 0.79 specificity, 0.79 precision, and 0.8 F-measure. These evaluation measures drop as the data gets more imbalanced to reach 0.39 sensitivity, 0.97 specificity, 0.495 precision, and 0.44 F-measure with 1:15 positive-to-negative ratio.

RVKDE is a ML algorithm that constructs a RBF neural network to approximate the probability density function of each class of objects in the training dataset. One main distinct feature of RVKDE is that it takes an average time complexity of $O(nlogn)$ for the model training process, where n is the number of instances in the training set. In order to improve the prediction efficiency, RVKDE considers only a limited number of nearest instances within the training dataset to compute the kernel density estimator of each class. One important advantage of RVKDE, in comparison with SVM, is that the learning algorithm generally takes far less training time with an optimized parameter setting. In addition to that, the number of training samples remaining after a data reduction mechanism is applied is quite close to the number of support vectors of SVM algorithm. Unlike SVM, RVKDE is capable of classifying data with more than two classes in one single run [39].

Table 1 summarizes these sequence-based approaches including the features that are used, the technique and/or the tools applied, and the validation datasets used.

4.2 Structure-Based Approaches

Structure-based PPI prediction methods use three-dimensional structural features such as domain information, solvent accessibility, secondary structure states, and hydrophobic and polar surface locations [18]. This section presents and evaluates some of the state-of-the-art structure-based approaches.

4.2.1 Struct2Net

Singh *et al.* [42] introduced Struct2Net as a structure-based PPI predictor. The method predicts interactions by threading each pair of protein sequences into potential structures in the Protein Data Bank (PDB) [43]. Given two protein sequences (or one sequence against all sequences of a species), Struct2Net threads the sequence to all the protein complexes in the PDB and then chooses the best potential match. Based on this match, it uses logistic regression technique to predict whether the two proteins interact.

Later on, Singh *et al.* [44] introduced Struct2Net as a web server with multiple querying options which is available at http://struct2net.csail.mit.edu. Users can retrieve Yeast, fly, and human PPI predictions by gene name or identifier while they can query for proteins of other organisms by AA sequence in FASTA format. Struct2Net returns a list of interacting proteins if one protein sequence is provided and an interaction prediction if two sequences are provided. When evaluated on yeast and fly protein pairs, Struct2Net achieves a recall of 0.80 with a precision of 0.30.

4.2.2 PRISM

Tuncbag *et al.* [45] developed PRISM as a template-based PPI prediction method based on information regarding the interaction surface of crystalline complex structures. The two sides of a template interface are compared with the surfaces of two target monomers by structural alignment. If regions of the target surfaces are similar to the complementary sides of the template interface, then these two targets are predicted to interact with each other through the template interface architecture. The method can be summarized as follows. First, interacting surface residues of target chains are extracted using Naccess [46]. Second, complementary chains of template interfaces are separated and structurally compared with each of the target surfaces by using MultiProt [47]. Third, the structural alignment results are filtered according to threshold values, and the resulting set of target surfaces is transformed into the corresponding template interfaces to form a complex. Finally, the Fiber-Dock [48] algorithm is used to refine the interactions to introduce

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 1: Sequence-based PPI prediction approaches.

Approach	Extracted Features	Technique/Tool	Datasets
PIPE (Pitre *et al.* 2006), PIPE2 (Pitre *et al.* 2008)	Short AA polypeptides	Similarity measure	Yeast protein (DIP and MIPS)
Auto Covariance (Guo *et al.* 2008)	AA physicochemical properties	Auto covariance, SVM	Yeast protein (DIP and MIPS)
Pairwise Similarity (Zaki et al 2009)	Pairwise similarity	SVM	Yeast protein
AA Composition (Roy *et al.* 2009)	AAC	Logistic regression, SVM, Naive Bayes	Yeast protein (GRID), worm protein (Li *et al.* 2004), fly protein (Biogrid)
AA Triad (Yu *et al.* 2010)	AA triad information	RVKDE	Human protein (HPRD)

flexibility, compute the global energy of the complex, and rank the solutions according to their energies. When the computed energy of a protein pair is less than a threshold of -10 kcal/mol, the pair is determined to interact.

PRISM has been applied for predicting PPIs in a human apoptosis pathway [49] and a p53- protein-related pathway [50], and has contributed to the understanding of the structural mechanisms underlying some types of signal transduction. PRISM obtained a precision of 0.231 when applied to a human apoptosis pathway that consisted of 57 proteins.

Because PRISM is a template-based method, its prediction accuracy depends on the template dataset prepared. Only PPIs whose interaction surface structures are conserved are expected to be predicted.

4.2.3 MEGADOCK

Ohue *et al.* [51] developed MEGADOCK as a protein-protein docking software package using the real Pairwise Shape Complementarity (rPSC) score. First, they conducted rigid-body docking calculations based on a simplified energy function considering shape complementarities, electrostatics, and hydrophobic interactions for all possible binary combinations of proteins in the target set. Using this process, a group of high-scoring docking complexes for each pair of proteins were obtained. Then, ZRANK [52] was applied for more advanced binding energy calculation and re-ranked the docking results based on ZRANK energy scores. The deviation of the selected docking scores from the score distribution of high-ranked complexes was determined as a standardized score (Z-score) and was used to assess possible interactions. Potential complexes that had no other high-scoring interactions nearby were rejected using structural differences. Thus binding pairs that had at least one populated area of high-scoring structures were considered. MEGADOCK has been applied for PPI prediction for 13 proteins of a bacterial chemotaxis pathway [53], [54] and obtained a precision of 0.4. MEGADOCK is available at http://www.bi.cs.titech.ac.jp/megadock.

One of the limitations of this approach is the demerit of generating false-positives for the cases in which no similar structures are seen in known complex structure databases.

4.2.4 Meta Approach

Ohue *et al.* [55] proposed a PPI prediction approach based on combining template-based and docking methods. The approach applies PRISM [45] as a template-matching method and MEGADOCK [51] as a docking method. A protein pair is considered to be interacting if both PRISM and MEGADOCK predict that this protein pair interacts. When applied to the human apoptosis signaling pathway, the method obtained a precision of 0.333, which is higher than that achieved using individual methods (0.231 for PRISM and 0.145 for MEGADOCK), while maintaining an F1 of 0.285 comparable to that obtained using individual methods (0.296 for PRISM, and 0.220 for MEGADOCK).

Meta approaches have already been used in the field of protein tertiary structure prediction [56], and critical experiments have demonstrated improved performance of Meta predictors when compared with individual methods. The Meta approach has also provided favorable results in protein domain prediction [57] and the prediction of disordered regions in proteins [58]. Although some true positives may be dropped by this method, the remaining predicted pairs are expected to have higher reliability because of the consensus between two prediction methods that have different characteristics.

4.2.5 PrePPI

Zhang *et al.* [59] proposed PrePPI (Predicting Protein-Protein Interactions) as a structural alignment PPI predictor based on geometric relationships between secondary structure information. Given a pair of query proteins A and B, representative structures for the individual subunits (M_A, M_B) are taken from the PDB (Protein Data Bank) [43] or from the ModBase [60] and SkyBase [61] homology model databases. Close and remote structural neighbors are found for each subunit. A template for the interaction exists if a PDB or PQS [62] structure contains interacting pairs that are structural neighbors of M_A and M_B. A model is constructed by superposing the individual subunits, M_A and M_B, on their corresponding structural neighbors. The likelihood for each model to represent a true interaction is then calculated using a Bayesian Network trained on 11,851 yeast interactions and 7,409 human interactions datasets. Finally the structure-derived score is combined with non-

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

structural information, including co-expression and functional similarity, into a naive Bayes classifier.

A common limitation of all structure-based PPI prediction approaches is the low coverage as the number of known protein structures is much smaller than the number of known protein sequences, and therefore, such approaches fail when there is no structural template available for the queried protein pair. Table 2 summarizes these structure-based approaches including the features that are used, the technique and/or the tools applied, and the validation datasets used.

5. Conclusion

This survey is focused on protein-protein interaction prediction. The main challenges that face PPI prediction were presented. We investigated several relevant existing approaches and provided a comparison of them. It is clearly noticed that PPI prediction still needs much research effort in order to achieve reasonable prediction accuracy.

One of the issues in the PPI prediction methods is that they do not use a uniform dataset and evaluation measure. We recommend creating a standard benchmark dataset taking into consideration the biological properties of proteins and examining the performance of all these methods on this benchmark dataset using a well-defined evaluation measures. This will allow us to compare the performance of these prediction methods in a fair and uniform fashion. This work can also be extended by investigating more recently published PPI prediction techniques, analyze them in depth, and compare their performance on a uniform dataset according to a uniform evaluation metrics. More focus should be given to the techniques which incorporate biological knowledge such as structural and functional information into the prediction process.

References

[1] N. Zaki, S. Lazarova-Molnar, W. El-Hajj, and P. Campbell, "Protein-protein interaction based on pairwise similarity," *BMC bioinformatics*, vol. 10, no. 1, p. 150, 2009.

[2] I. Xenarios and D. Eisenberg, "Protein interaction databases," *Current Opinion in Biotechnology*, vol. 12, no. 4, pp. 334–339, 2001.

[3] W. K. Kim, J. Park, J. K. Suh, *et al.*, "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair," *Genome Informatics Series*, pp. 42–50, 2002.

[4] G. D. Bader and C. W. Hogue, "Analyzing yeast protein–protein interaction data obtained from different sources," *Nature biotechnology*, vol. 20, no. 10, pp. 991–997, 2002.

[5] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[6] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 63, no. 3, pp. 490–500, 2006.

[7] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011.

[8] J. C. Melo, G. Cavalcanti, and K. Guimaraes, "Pca feature extraction for protein structure prediction," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 4. IEEE, 2003, pp. 2952–2957.

[9] Y. Sasaki, "The truth of the f-measure," *Teach Tutor mater*, pp. 1–5, 2007.

[10] D. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[11] P. L. Bartel and S. Fields, *The yeast two-hybrid system*. Oxford University Press, 1997.

[12] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, *et al.*, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[13] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin, "A generic protein purification method for protein complex characterization and proteome exploration," *Nature biotechnology*, vol. 17, no. 10, pp. 1030–1032, 1999.

[14] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, *et al.*, "Global analysis of protein activities using proteome chips," *science*, vol. 293, no. 5537, pp. 2101–2105, 2001.

[15] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, *et al.*, "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules," *Science*, vol. 295, no. 5553, pp. 321–324, 2002.

[16] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein–protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[17] A. Szilágyi, V. Grimm, A. K. Arakaki, and J. Skolnick, "Prediction of physical protein–protein interactions," *Physical biology*, vol. 2, no. 2, p. S1, 2005.

[18] A. Porollo and J. Meller, "Computational methods for prediction of protein-protein interaction sites," *Protein-Protein Interactions-Computational and Experimental Tools; W. Cai and H. Hong, Eds. InTech*, vol. 472, pp. 3–26, 2012.

[19] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, *et al.*, "Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC bioinformatics*, vol. 7, no. 1, p. 365, 2006.

[20] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.

[21] H.-W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil, "Mips: a database for genomes and protein sequences," *Nucleic acids research*, vol. 30, no. 1, pp. 31–34, 2002.

[22] S. Pitre, C. North, M. Alamgir, M. Jessulat, A. Chan, X. Luo, J. Green, M. Dumontier, F. Dehne, and A. Golshani, "Global investigation of protein–protein interactions in yeast saccharomyces cerevisiae using re-occurring short polypeptide sequences," *Nucleic acids research*, vol. 36, no. 13, pp. 4286–4294, 2008.

[23] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[24] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.

[25] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[26] N. Zaki, S. Wolfsheimer, G. Nuel, and S. Khuri, "Conotoxin protein classification using free scores of words and support vector machines," *BMC bioinformatics*, vol. 12, no. 1, p. 217, 2011.

[27] V. N. Vapnik, "Statistical learning theory (adaptive and learning systems for signal processing, communications and control series)," 1998.

[28] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[29] N. Zaki, S. Deris, and H. Alashwal, "Protein-protein interaction detection based on substring sensitivity measure," *International journal of biomedical sciences*, vol. 2, no. 1, pp. 148–154, 2006.

[30] N. Zaki, "Protein-protein interaction prediction using homology and inter-domain linker region information," *Advances in Electrical Engineering and Computational Science*, vol. 67, no. 4, pp. 635–645, 2007.

[31] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *Journal of computational biology*, vol. 10, no. 6, pp. 857–868, 2003.

[32] H. Saigo, J.-P. Vert, N. Ueda, and T. Akutsu, "Protein homology detection using string alignment kernels," *Bioinformatics*, vol. 20, no. 11, pp. 1682–1689, 2004.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 2: Structure-based PPI prediction approaches.

Approach	Extracted Features	Technique/Tool	Datasets
Struct2Net (Singh *et al.* 2006, 2010)	Homology with known protein complexes in PDB	Logistic regression	Yeast, Fly ,and Human protein
PRISM (Tuncbag *et al.* 2011)	Interaction surface of crystalline complex structures	Naccess, MultiProt, Fiber-Dock	Human Protein (Ozbabacan et al. 2012, Tuncbag *et al.* 2009)
MEGADOCK (Ohue *et al.* 2013a)	Shape complementarities, electrostatics, and hydrophobic interactions	rPSC, ZRANK	Bacterial protein (Ohue *et al.* 2012, Matsuzaki *et al.*. 2013)
Meta Approach (Ohue *et al.* 2013b)	Interaction surface of crystalline complex structures, shape complementarities, electrostatics, and hydrophobic interactions	PRISM, MEGADOCK	Human protein
PrePPI (Zhang *et al.* 2012)	Secondary structure	Bayesian networks, Naive Bayes	Yeast protein, Human protein

[33] S. Roy, D. Martinez, H. Platero, T. Lane, and M. Werner-Washburne, "Exploiting amino acid composition for predicting protein-protein interactions," *PloS one*, vol. 4, no. 11, p. e7813, 2009.

[34] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein–protein interactions from protein sequences," *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[35] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein–protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[36] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[37] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, *et al.*, "A map of the interactome network of the metazoan c. elegans," *Science*, vol. 303, no. 5657, pp. 540–543, 2004.

[38] C.-Y. Yu, L.-C. Chou, and D. T. Chang, "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins," *BMC bioinformatics*, vol. 11, no. 1, p. 167, 2010.

[39] Y.-J. Oyang, S.-C. Hwang, Y.-Y. Ou, C.-Y. Chen, and Z.-W. Chen, "Data classification with radial basis function networks based on a novel kernel density estimation algorithm," *Neural Networks, IEEE Transactions on*, vol. 16, no. 1, pp. 225–236, 2005.

[40] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, *et al.*, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome research*, vol. 13, no. 10, pp. 2363–2371, 2003.

[41] G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, *et al.*, "Human protein reference databaseâĂŤ2006 update," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D411–D414, 2006.

[42] R. Singh, J. Xu, and B. Berger, "Struct2net: Integrating structure into protein-protein interaction prediction." in *Pacific Symposium on Biocomputing*, vol. 11. Citeseer, 2006, pp. 403–414.

[43] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[44] R. Singh, D. Park, J. Xu, R. Hosur, and B. Berger, "Struct2net: a web service to predict protein–protein interactions using a structure-based approach," *Nucleic acids research*, vol. 38, no. suppl 2, pp. W508–W515, 2010.

[45] N. Tuncbag, A. Gursoy, R. Nussinov, and O. Keskin, "Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism," *Nature protocols*, vol. 6, no. 9, pp. 1341–1354, 2011.

[46] S. J. Hubbard and J. M. Thornton, "Naccess," *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, vol. 2, no. 1, 1993.

[47] M. Shatsky, R. Nussinov, and H. J. Wolfson, "A method for simultaneous alignment of multiple protein structures," *Proteins: Structure, Function, and Bioinformatics*, vol. 56, no. 1, pp. 143–156, 2004.

[48] E. Mashiach, R. Nussinov, and H. J. Wolfson, "Fiberdock: flexible induced-fit backbone refinement in molecular docking," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 6, pp. 1503–1519, 2010.

[49] S. E. Acuner Ozbabacan, O. Keskin, R. Nussinov, and A. Gursoy, "Enriching the human apoptosis pathway by predicting the structures of protein–protein complexes," *Journal of structural biology*, vol. 179, no. 3, pp. 338–346, 2012.

[50] N. Tuncbag, G. Kar, A. Gursoy, O. Keskin, and R. Nussinov, "Towards inferring time dimensionality in protein–protein interaction networks by integrating structures: the p53 example," *Molecular BioSystems*, vol. 5, no. 12, pp. 1770–1778, 2009.

[51] M. Ohue, Y. Matsuzaki, N. Uchikoga, T. Ishida, and Y. Akiyama, "Megadock: An all-to-all protein-protein interaction prediction system using tertiary structure data." *Protein and peptide letters*, 2013.

[52] B. Pierce and Z. Weng, "Zrank: reranking protein docking predictions with an optimized energy function," *Proteins: Structure, Function, and Bioinformatics*, vol. 67, no. 4, pp. 1078–1086, 2007.

[53] M. Ohue, Y. Matsuzaki, T. Ishida, and Y. Akiyama, "Improvement of the protein–protein docking prediction by introducing a simple hydrophobic interaction model: An application to interaction pathway analysis," in *Pattern Recognition in Bioinformatics*. Springer, 2012, pp. 178–187.

[54] Y. Matsuzaki, M. Ohue, N. Uchikoga, and Y. Akiyama, "Protein-protein interaction network prediction by using rigid-body docking tools: Application to bacterial chemotaxis." *Protein and peptide letters*, 2013.

[55] M. Ohue, Y. Matsuzaki, T. Shimoda, T. Ishida, and Y. Akiyama, "Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods," in *BMC Proceedings*, vol. 7, no. Suppl 7. BioMed Central Ltd, 2013, p. S6.

[56] H. Zhou, S. B. Pandit, and J. Skolnick, "Performance of the pro-sp3-tasser server in casp8," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. S9, pp. 123–127, 2009.

[57] H. K. Saini and D. Fischer, "Meta-dp: domain prediction meta-server," *Bioinformatics*, vol. 21, no. 12, pp. 2917–2920, 2005.

[58] T. Ishida and K. Kinoshita, "Prediction of disordered regions in proteins based on the meta approach," *Bioinformatics*, vol. 24, no. 11, pp. 1344–1348, 2008.

[59] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, *et al.*, "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[60] U. Pieper, N. Eswar, F. P. Davis, H. Braberg, M. S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B. M. Webb, D. Eramian, *et al.*, "Modbase: a database of annotated comparative protein structure models and associated resources," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D291–D295, 2006.

[61] N. Mirkovic, Z. Li, A. Parnassa, and D. Murray, "Strategies for high-throughput comparative modeling: Applications to leverage analysis in structural genomics and protein family organization," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 4, pp. 766–777, 2007.

[62] K. Henrick and J. M. Thornton, "Pqs: a protein quaternary structure file server," *Trends in biochemical sciences*, vol. 23, no. 9, pp. 358–361, 1998.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

SESSION

MACHINE LEARNING, DATA AND INFORMATION MINING + SIGNAL PROCESSING AND DATA QUALITY ENHANCEMENT + DATABASES

Chair(s)

TBA

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

BIOBase: An Adaptable Biometric Tool & Database

Christine Brown, Karl Ricanek and Devon M. Simmonds,
University of North Carolina Wilmington, 601 S. College Rd.
601 South College Road, Wilmington, North Carolina, 28403
{ cp0304, ricanekk, simmondsd}@uncw.edu

Abstract

BIOBase is an adaptable biometric database coupled with a database frontend that allows access to and view of the data in the database. The database is designed to enable access to multiple biometric modalities in a single database and the addition of different biometric modalities to the database without having to change any of the database tables. BIOBase reduces scripting and enhances the ease with which users can access the data. In addition, users can select and copy files to be used in an experiment to a location of their choice as well as reload and modify data from an old experiment. This paper outlines the design of the database using a model-driven approach. Results and lessons learned are presented.

I. INTRODUCTION

Biometrics is the study of how persons may be identified using physiological or behavioral characteristics [8]. To help researchers accomplish their tasks, biometric databases are often created to aggregate useful data and to minimize data redundancy. While many such databases are currently available [6], researchers typically encounter several problems when using these databases. First, available biometric databases typically focus on just one or sometimes two concerns or biometric modalities. Biometric concerns or modalities typically include face, iris, fingerprint, vein, etc. These modalities have further characteristics that are important to track. For example the face modality captures face images however, the face can be captured in several wavelengths from the typical visible spectrum to that of near infrared (NIF) to far infrared (IR). Each spectrum has different qualities that are important to face recognition, however, algorithms do not exists that can compare face images between the different spectrums. Further, BIOBase can organize other attributes of the face, eye location, eye color, marks, scars, tattoos, etc.

The fact that datasets are typically tested separately presents challenges in the use of biometric databases. Frequently, tests are run in which a particular set of data needs to be organized in order to run the experiments. Our research group has seen a gradual collection of redundant biometric data and clusters of files on our server. This makes running experiments and tests very challenging for a large experiment. Some experiments is comprised of face images that exceed 10,000 or 100,000. It is impossible for a human to review all image collection to ensure that duplicate face images are not accidentally enrolled—possibly under a different subject identifier

In general, face recognition research conducted by academics will involve the use of multiple datasets to address a set of questions. The datasets are typically collected by academic researchers to address very specific questions within the confines of the biometric modality. Typically researchers will compose there experiments by combining several datasets with their desperate acquisition standards. BioBase was developed to address the needs of the Face Aging Group at University of North Carolina Wilmington (www.FaceAgingGroup.com) to conduct experiments over a set of 22 face databases with more than 1.7 Million face images.

To accomplish these goals and address the aforementioned problems, we designed BIOBase, a biometric database and tool. Our goal was to create an application to access a central database, fetch required files and build experimental datasets. These datasets can include one modality or multiple modalities. We also wanted to make an easily maintainable and easily adaptable system that may be used in multiple ways. For example, BIOBase can be used to create experiments or BIOBase can be used to modify data in the database.

The BIOBase tool is designed to be used by persons unfamiliar with querying databases. The tool allows users to narrow down the generated data into a subset that is more desirable for their experiment. Searching through images from a database is likely to be faster than reading in flat files [13]. BIOBase as a whole, transforms the prior flat-file work environment into a more robust domain.

The tool-driven building of datatsets for experiments has many benefits. One benefit includes the decrease in the amount of time it takes to build a particular dataset for a given experiment. The BIOBase tool has a visual user friendly design that allows the average computer user to choose a desired dataset, thus taking away the complexity from the actual process. Once a desired dataset has been created it can be saved in the database and/or on the user's computer. This alleviates clutter on a server, or personal computer, by eliminating the need to have multiple directories that hold data that the user chose not to use. Another benefit of BIOBase is that it can make longitudinal experimental datasets by cutting across multiple datasets to make experiment sets vary. This allows for more rigorous testing of algorithms. This can help biometric research labs around the world and create a more efficient International Biometric Society [7].

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

The rest of the paper is organized as follows. Section II describes project planning and software estimation activities for the BIOBase project. Section III describes software design and section IV describes implementation and testing. Section V describes discussion, lessons learned, and future work, and section VI summarizes the results.

II. PROJECT PLANNING

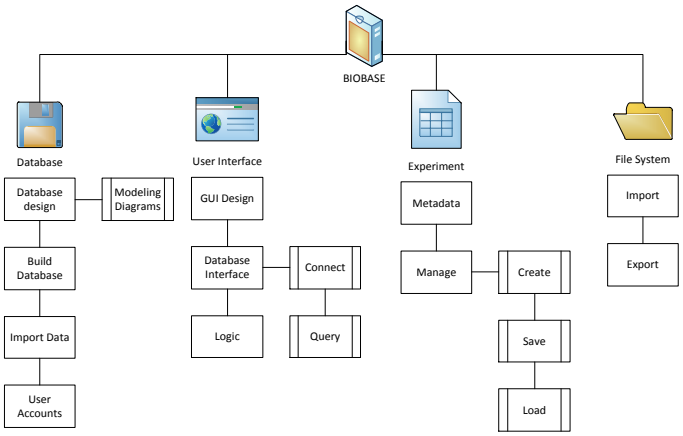


Figure 1. Work Breakdown Structure Diagram

Project planning began with requirements analysis and the development of a work breakdown structure as illustrated in Figure 1. The figure identifies the core tasks that were undertaken for the project and that were used as a basis for computing estimates of time, cost and human effort for the project. Given the immense size of BIOBase, the project was broken down into four main subtasks each of which was further refined.

A UML use case diagram that expresses the scope of the software requirements is shown in Figure 2.

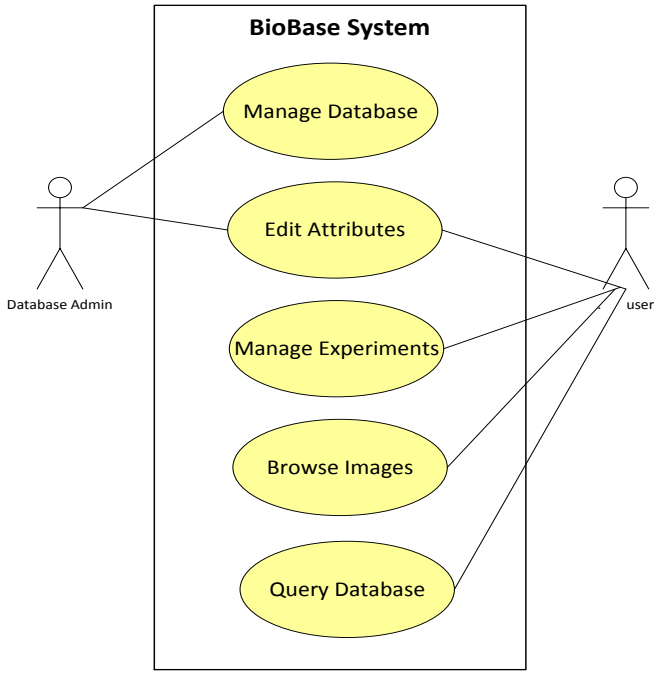


Figure 2. Use Case Diagram for BIOBase

The primary features of the software as reflected in Figure 2 include loading and saving experiments, querying the database and editing and updating attributes for entities stored in the database. A typical user query could for example involve finding all Hispanic and White male subjects between the ages of 30 and 35.

II.A Software Estimation

In addition to task analysis and scheduling, software estimation is a central activity in software project management. Estimation reflects an awareness of the need to manage uncertainties and recognition of the inevitability of risks in complex projects. Three techniques were used in our estimation effort: lines of code (LOC), function points (FP), and the COCOMO II model.

Lines of Code Estimation

Function	Optimistic	Likely	Pessimistic	Estimated LOC
Database Interface	100	550	700	500
Graphical User Interface	100	250	300	350
Experimental Management System	70	100	130	100
File Input	80	100	120	100
File Output	80	100	120	100
File Management System	150	200	250	200
User Accounts	100	150	200	150
			Total LOC for BIOBase ➔	**1500**

Table 1. Estimated Lines of Code *– These estimates were based on the programmers knowledge of the language used and experience with the functions.*

The LOC estimates for the BIOBase software was computed using software features listed in Table 1. Each estimate was computed as the mean of the optimistic, likely and pessimistic values following a beta distribution [14]. Each item in The "Estimated LOC" column is calculated using the formula:

$$Estimated\ LOC = \frac{Optimistic + (4 \times Likely) + Pessimistic}{6}$$

We assumed an average productivity rate of 500 LOC/person-month. This resulted in an estimated product duration of 3 person months (1500LOC/500LOC/PM). To compute the cost we assumed a starting median salary of $56,600 for computer science undergraduates as reported by payscale.com

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

(http://www.payscale.com/best-colleges/degrees.asp). This resulted in an estimated cost of building this application as:

$$3\ months \times 56{,}600\$/month = \$\mathbf{169{,}800.}$$

We postulated that the database interface, graphical user interface, file input, and file output estimates were more accurate than the estimates for the other features because the developers were more familiar with these features. We also computed the proportion of the total LOC that each feature represented. This was done to give us a sense of the relative effort that would be required for each feature. The feature proprotionality results are shown in Figure 4.

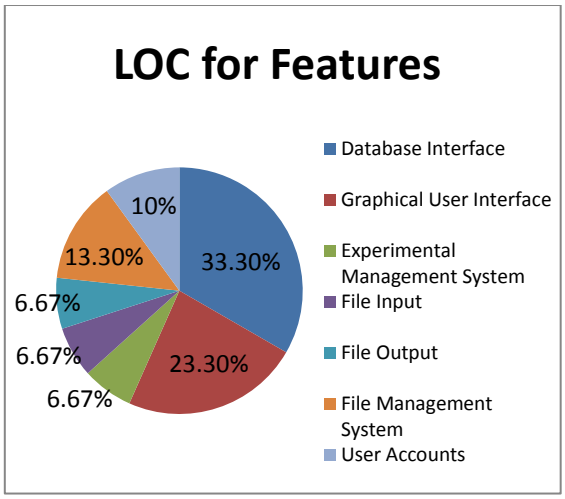


Figure 4. Lines of Code – Visual representation of code proportionality.

Since BIOBase is a a biometric database, it seems appropriate that the database feature of the application consumed the most lines of code (as seen in Figure 4), with the graphical user inferface requiring the second most lines of code. The rest of the functionality represented slightly less than 50% of the estimated LOC.

Function Point Estimation

		Weighting Factor			
Factors	**Count**	**S**	**A**	**C**	**Total**
External Inputs	**6**	**2**	**5**	**8**	**30**
External Outputs	**7**	**2**	**5**	**8**	**35**
External Inquires	**4**	**3**	**4**	**6**	**16**
Internal Files	**7**	**5**	**6**	**8**	**42**
External Interface Files	**5**	**3**	**5**	**6**	**25**
Count Total					**148**

Table 2: Information Domain Values (IDV)

KEY: S = Simple, A = Average, C = Complex

Function points [13] are a way to estimate software functionality using "functional" units rather than traditional lines of code. The intuition is that during the early stages of development it may be easier to conceptualize what the software should do (i.e., the software "functions"), than the number of lines of code of the final system.

The Function Points estimate for BIOBase was calculated using the data shown in Table 2 and Table 3. The computation is done in two steps. In the first step an information domain value (IDV) is computed as a function of the five information domain factors shown in Table 2. In the second step a value adjustment factor (VAF) is computed as a function of the fourteen factors shown in Table 3. Finally, the number of function points for the project is computed using the formula:

$$Function\ Points = IDV * [0.65 + (.01 \times VAF)]$$
$$= 148 \times [0.65 + (.01 \times 38)] = \mathbf{152.44}$$

Factor	Value
Backup and Recovery	2
Data Communications	4
Distributed Processing	2
Performance Critical	3
Heavily utilized operating environment	4
Online data entry	1
Input transaction over multiple screens	1
Master files updated online	1
Information domain values complex	3
Internal processing complex	4
Code designed for reuse	6
Conversion/installation in design	3
Multiple installations	4
Application designed for change	3
Total:	**41**

Table 3. Complexity Factors in VAF Computation

Assuming these 152.44 function points would be completed in the three month computed for the LOC estimation, results in an estimate of 50.81 function points per person month. This implies a cost per function point of $1,113.95.

II.B Risk Analysis

A sample of the risks considered for the project along with the mitigation, monitoring and management (RMMM) plan used for the project is listed in Table 4. It was our opinion that database failure and scope/task creep had the greatest likelihood of occurring, while project failure due to poor planning was least likely. On the other hand we estimated that server errors and insufficient time would have the greatest impact if those risks materialized.

Risk	C	P	I	RMMM
Programming Language Inexperience	ST	50%	3	Research and select best programming language(s) for project.
Server Errors	TE	40%	4	Test server integrity. In case of error, consult systems administrator.
Database Failure	TE	60%	3	Review database schema and implement queries

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

				which are optimal for the design.
Project failure due to poor planning	BU	20%	2	Plan weekly meetings and set goals to stay in scope and on task.
Network failure	TE	30%	2	Identify alternate networks that may be used.
Lack of Time	BU	35%	4	Schedule goals to meet requirements and deadlines to meet designated due dates.
Scope & task creep	BU	60%	2	Avoid dwelling on irrelevant tasks and make sure task is in scope.

Table 4. Risk Management Plan

KEY: C=Category, P=Probability, I=Impact
BU - Business risk, PS = Project size risk, TE = Technology risk, ST = Staff inexperience
Impact values: 1 = catastrophic, 2 = critical, 3 = marginal, 4 = negligible

III. SOFTWARE DESIGN

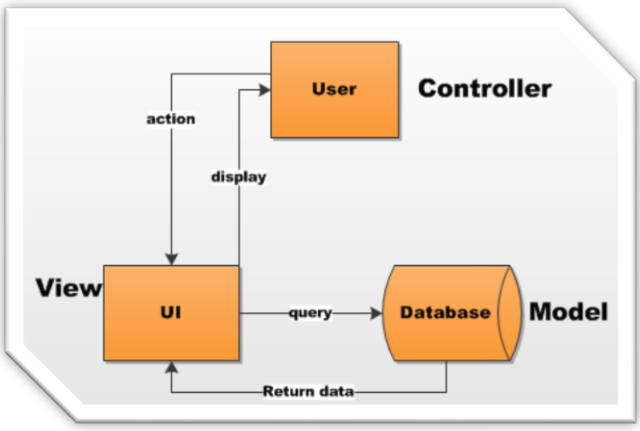


Figure 5. MVC Architectural Model

Software design began with the specification of the model view controller (MVC) software architecture shown in Figure 5. Information processing in the MVC architecture typically begins with a user-initiated request. This request is serviced by the controller and converted into a command that the model understands. Conversely, the view may query the model after which the model returns data to the view and awaits further user action. One of the benefits of the MVC design is that it has an organizational structure that supports scalability. In addition, the clean separation of tasks, makes it is easier to maintain and modify MVC architectures.

Subsystem Design

The activity diagram shown in Figure 6, illustrates how a user navigates through the system. Once logged in, the user can either load an experiment or create a new experiment. In order to create a new experiment the user must enter a query which BIOBase then populates. The user may then selects items from the query results to add to the experiment. Once the user has completed adding items to the experiment, the user can choose to save the experiment or close the program.

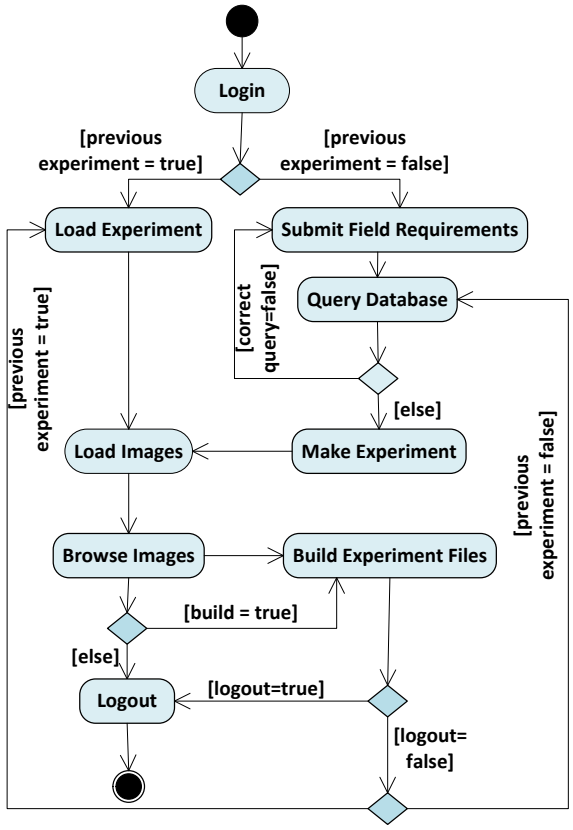


Figure 6. BIOBase Activity Diagram

Users may also choose to browse items in the database, for example, photos. In this instance, once the user submits a query, an array of images is loaded. The user can either accept a photo for the experiment or move on to the next photo. A class diagram for the overall BIOBase application is shown in Figure 7. The diagram shows the subsystem-level structural concepts required for the application to satisfy its software requirements.

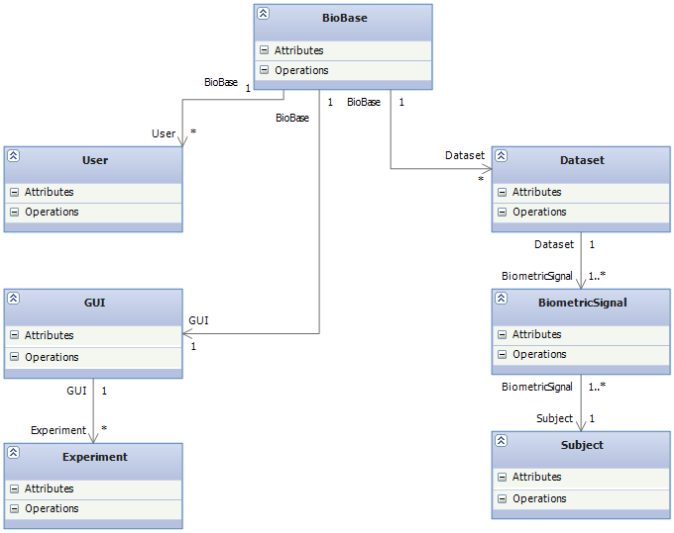


Figure 7. BIOBase Class Model

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

IV. IMPLEMENTATION & TESTING

Following early analysis, project planning, software and database design activities, and data set conversion, coding began iteratively with the development of the user interface. This user interface is shown in Figure 8. Once the basic functionality of the user interface was up and running, we began incrementally adding functionality, such as: clicking the arrows to switch pictures, giving the search button functionality, and allowing users to add photos to an experiment.

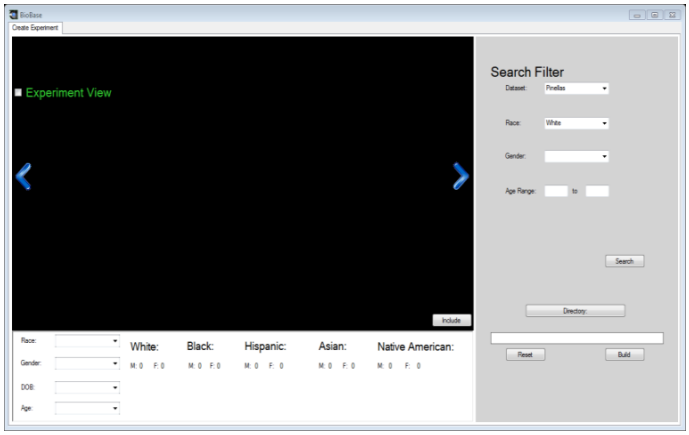


Figure 8. User Interface Design

The previous flat file system had to be converted into the current BIOBase database before work on the front end of BIOBase began. This data conversion included importing each CSV file into Microsoft SQL Server. Each data set had a separate CSV file and each record within the CSV file contained an image name. The images from the data sets were moved to the server and this path was used to set the image path for each image. BIOBase accesses the images using the image path and file name rather than storing the images in the database. This is most helpful for wide range experiments that could yield a plethora of search results. After all of the data sets were converted into the database a copy of the database was made for testing purposes.

The software was tested using a black-box testing approach focusing on the functionality and accuracy of the system. Accuracy is important for this application because the software will be interacting with a database with sensitive information. Test cases included:

1. Checks for invalid or missing functions.
2. Check to see the program can be run from machine to machine.
3. Database structure tests, ensuring data accuracy and stability.
4. Data and meta-data tests checking data integrity
5. A performance test checking for efficiency and speed.
6. Testing connections, compatibility and server response.

A sample of test cases is presented in Table 5. Each test case was executed (and corrections effected where necessary) until the test passed.

Description	Input	Expected Output	Test Status
Log in attempt	Initiate connection	Connection success	Passed
Connection with database	Request connection with database	Connected with database	Passed
Network connection terminated	Disconnect from database and network	Entered Offline mode	Passed
Load Experiment	Request for experiment file	File uploaded	Passed
Query images	Request images from database	Return image requests	Passed
Show images on interface	Load images to picture box	Images displayed for user search	Passed
Select/Deselect image	Select image for dataset	Image added to experiment set	Passed
Save experiment set	Save images to experiment set in file	A successful save of images and organized in files	Passed
Proper log out	User logged out	A proper close of the application displaying user login screen	Passed
Disconnection from database when logged out	A termination of connection to database	A closed connection to database	Passed

Table 5: Sample Test Cases

After the unit testing was complete, end-to-end testing began. The end-to-end testing was done by following the activity diagram in Figure 6. There were multiple iterations of the end-to-end testing following the different paths of the activity diagram. For example, in one iteration the tester would (1) log into the system, select search filter requirements, and click search. When the results are returned the tester would (2) scroll through the results, making sure that the results match the search requirements. The tester would add multiple records to an experiment and then save the experiment. The tester would then (3) browse through the experiment to make sure the records were added correctly. Finally, the tester would (4) log out of the system and move on to a different end-to-end test situation.

V. DISCUSSION, LESSONS LEARNED AND FUTURE WORK

	Requirements	Design	Actual Size
LOC	2955 LOC	2080 LOC	1500 LOC

Table 6. Original Estimates vs. Actual Values

The lessons learned from this project are many. For starters, software estimation is an evolving project discourse in

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

which we "fake rationality" as David Parnas calls it [16]. Table 4 shows requirements and design LOC estimates for the project along with the actual size. One reason for the overestimation, was a lack of comprehension of the original project scope. As a result, the scope of the project narrowed as the project evolved. Listed below are some recommendations resulting from this experience that other practitioners may find beneficial:

1. Where possible, include a domain expert to help with scoping.
2. Estimation should take into account the occurrence of unexpected events and their impact on the project.
3. The complexity of the project as a whole and in its constituent parts should be carefully considered when estimating time, cost, and effort.

BIOBase is an adaptable biometric database that allows users the benefits of accessing just a single database to store and manipulate multiple biometric modalities. The software provides an easy way to test across different datasets and different modalities. The application also encourages learning for people that are new to biometrics by simplifying their data collection.

Future work for BIOBase includes the ability to compare saved experiments and giving administrators the option to delete saved experiments after some designated time period has elapsed. This will help reduce redundant data and keep the database clean. In the future, BIOBase will also address security issues by enabling an administrator to add and modify user rights.

IV. CONCLUSION

Many research facilities have a problem with redundant data or unorganized data. Using BIOBase to keep multiple data sets organized can save tremendous amounts of human effort in planning large scale experiments across a set of biometric datasets. It makes it easy for researchers to create experiments, it saves researchers time when creating experiments, it can prevent redundant experiments, and it allows researchers to correct obvious mistakes about records being kept in the database. Unorganized data can also have an effect on the outcome of experiments. BIOBase offers a solution to this problem by keeping experiments and data on the BIOBase database. Keeping experiments on the database will allow users to easily query and compare previous experiments. This will allow for more thorough testing of algorithms and a better understanding of the results.

REFERENCES

[1]. Ababneh, Mohommad F. "Face Recognition Eigenface and Fisherface Performance." *European Journal of Scientific Research* July 21.1 (2008): 6-12. Print.

[2]. Al-Hijaili, Shoa'a J., and Manal AbdulAziz. "BIOMETRICS IN HEALTH CARE SECURITY SYSTEM, IRIS-FACE FUSION SYSTEM." *International Journal Of Academic Research* January 3.1 (2011). Print.

[3]. Argyropoulos, Savvas, Dimitrios Tzovaras, Dimosthenis Ioannidis, Yannis Damousis, Michael. G. Strintzis,, Martin Braun, and Serge Boverie. "Biometric Template Protection in Multimodal Authentication Systems Based on Error Correcting Codes." *Journal of Computer Security* 18.1 (2010): 161-85. Print.

[4]. Chen, Xuerong, and Zhongliang Jing. "INFRARED FACE RECOGNITION BASED ON LOG-GABOR WAVELETS." *International Journal of Pattern Recognition and Artificial Intelligence* 20.3 (2006): 351-61. Print.

[5]. ".CSV File Extension." *CSV File Extension*. Web. 23 Feb. 2012. <http://www.fileinfo.com/extension/csv>.

[6]. Helland, Pat. "If You Have Too Much Data, Then 'Good Enough' Is Good Enough." *Communications of the ACM* June 54.6 (2011): 40-47. Print.

[7]. "The International Biometric Society Â» Definition of Biometrics." *The International Biometric Society Â» Definition of Biometrics*. Web. 29 Mar. 2012. <http://www.biometricsociety.org/about/definition-of-biometrics/>.

[8]. "Introduction to Biometrics." *Biometrics.gov* -. Web. 23 Feb. 2012. <http://www.biometrics.gov/ReferenceRoom/Introduction.aspx>.

[9]. Koltzch, Gregor. "BIOMETRICS – MARKET SEGMENTS AND APPLICATIONS." *Journal of Business Economics and Management* VIII.2 (2007): 119-22. Print.

[10]. Lyon, David. "Biometrics, Identification And Surveillance." *Bioethics* 22.9 (2008): 499-508. Print.

[11]. Plaga, Rainer. "Biometric Keys: Suitable Use Cases and Achievable Information Content." *International Journal of Information Security* 8.6 (2009): 447-54. Print.

[12]. Pressman, Roger S. Software Engineering: A Practitioner's Approach. 7th ed. New York: McGraw-Hill Higher Education, 2010. Print.

[13]. Storey, Veda C., Roger H. Chiang, Debabrata Dey, Robet C. Goldstein, and Shankar Sundaresan. "Database Design with Common Sense Business Reasoning and Learning." *ACM Transactions on Database Systems* 22.4 (1997): 471-512. Print.

[14]. Roger S. Pressman. Software Engineering: A Practitioner's Approach 7th Ed. Page 700, McGraw-Hill 2010.

[15]. [13] Vipin Saxena and Manish Shrivastava. 2009. Performance of function point analysis through UML modeling. *SIGSOFT Softw. Eng. Notes* 34, 2 (February 2009).

[16]. D. L. Parnas and P C Clements, "A rational design process: How and why to fake it," IEEE Transactuions on Software Engineering, 12(2):251-257, IEEE Press 1986.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Prediction and Rule Extraction of Major Histocompatibility Complex Class II Epitopes by Logic Minimization

C. Aguilar-Bonavides[1, §,*], **R. Cruz-Cano**[2, §], **and C. Lanzas**[1, 3]
[1]National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN, US 37996
[2]Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD, US 20742
[3]Department of Biomedical and Diagnostic Sciences, University of Tennessee, Knoxville, TN US 37996
[*]Corresponding author
[§]These authors contributed equally to this work.

Abstract—*Helper T cells recognize pathogen peptides, known as epitopes that are displayed on the surface of professional antigen presenting cells (APCs) via major histocompatibility complex class II (MHC-II). The identification of epitopes capable of creating an immune response is essential for understanding the functioning of the immune system, and very important step toward vaccine and immunotherapy development. Computational prediction of epitopes can reduce the number of peptide candidates for further experimental confirmation. Here we present a method based on logic minimization for the prediction of epitopes. Logic Minimization Method (LMM) uses the axioms of Boolean algebra to minimize a set of digital variables. Our proposed method has been trained on peptide-MHC-II binding data, allowing for the generation of rules to identify MHC-II epitopes for the human allele HLA-DRB1*0101. We have identified a set of 33 rules that describe our dataset with predictive accuracy comparable to machine learning methods.*

Keywords: MHC class II, immunoinformatics, logic minimization, epitope prediction, machine learning.

1. Introduction

$CD4^+$ T helper lymphocytes recognize peptides in pathogens known as epitopes bounded to major histocompatibility complex class II (MHC-II). The MHC-II – epitope complex is displayed on the surface of professional antigen presenting cells (APCs). Once the T-cell recognizes and binds to the MHC-II – epitope complex, the APC sends out an additional co-stimulatory signal to activate the T-cell, initiating an immune response. The identification of epitopes capable of creating a response is essential for understanding the functioning of the immune system, and very important step toward synthetic vaccine and immunotherapy development. Identification and characterization of epitopes is a complex process, it involves the use of methods that are technically challenging, time-consuming and expensive, such as X-ray crystallography [1], peptide-microarray-based identification [2], and proteolytic fragmentation [3]. Therefore, their application is limited to a small number of proteins.

Computational prediction of epitopes can screen large number of proteins and reduce the number of peptide candidates for further experimental testing. There are two main challenges in predicting MHC-II epitopes. First, although the binding core of the MHC-II molecule is usually nine amino acids long, the open binding cleft allows a significant length variation (from 8 to 25 amino acids) [4]. Second, MHC molecules are the most polymorphic of mammalian proteins, and therefore they have ability to bind to very different sets of peptides [5].

Techniques in the identification of relevant motifs capable of binding to MHC molecules have been based on the use of experimentally derived quantitative matrices. Such techniques include position specific scoring matrix (PSSM) [6], Gibbs sampling approach, [7], and average relative binding (ARB) matrix methods [8]. These methods assume that each residue in the epitope independently contributes to peptide-MHC binding. However, the performance of matrix-based methods relies on the quality of the matrix and high quality matrices are not often available in many alleles. Methods that do not require peptide pre-processing are better predicting methods. Machine learning techniques such as Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are superior methods that have acquired high accuracies, in the range of 70-90% [9] and do not require peptide pre-processing. These machine learning techniques are "black-box classifiers". Understandable rules of motifs cannot be generated with these methods. In addition to the limited explanation capability, SVMs and ANNs can also be sensitive to noise [10]. Logic Minimization Method (LMM) uses the axioms of Boolean algebra to minimize a set of digital variables. The method proposed here is trained on peptide-MHC-II binding data, allowing for the generation of rules to identify MHC-II epitopes for allele *HLA-DRB1*0101*. Allelic variants of *HLA-DR* molecules display high structural and functional similarity [11], then, it is expected that

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

this approach could be generally applicable for predicting peptides that bind to other *HLA-DR* molecules. It is important for the treatment of autoimmune diseases to determine which peptides bind to *HLA-DR* variants that will help in treatment of these diseases [12].

1.1 Logic minimization method

LMM is the application of algebraic axioms to a binary dataset with the purpose of reducing the number of digital variables and/or rules needed to express it [13]. LMM is mostly used in the optimization of digital circuits. The purpose of LMM is to find a simplified representation for a given function requiring the minimum number of logic operations and variables for its implementation. By keeping only this relatively small number of variables for the rule extraction, it is easier to detect all of them at once to discover patterns or trends. Electronic devices are composed of a diverse number of digital circuits or gates that allow the device to perform its function. A fully functional electronic device has a complex combination of several gates, with some having the same inputs. A logic gate is an elementary building block of a digital circuit implementing a Boolean function (0 or 1). In order to optimize the number of digital circuits, LMM is performed, transforming and minimizing the representation of gates and achieving a technology-independent set of logical expressions. If an equivalent circuit can be formed with fewer gates or fewer inputs, the cost of the circuit is reduced and its reliability is enhanced. This process is called reduction or simplification of combinational logic circuits and is performed by using the laws and rules of Boolean algebra. For an introduction to LMM techniques, see [13].

Table 1 Binary function before LMM

Pattern	Inputs				Output
	A	B	C	D	
1	0	0	0	0	1
2	0	0	0	1	0
3	0	0	1	0	1
4	0	0	1	1	1
5	0	1	0	0	0
6	0	1	0	1	0
7	0	1	1	0	-
8	0	1	1	1	-
9	1	0	0	0	1
10	1	0	0	1	0
11	1	0	1	0	1
12	1	0	1	1	1
13	1	1	0	0	0
14	1	1	0	1	0
15	1	1	1	0	-
16	1	1	1	1	-

A Boolean function can be described using a truth table that defines the input-output relationship of binary variables in a Boolean function, and show whether a propositional expression is logically valid (Table 1). The truth table can be further reduced as shown in Table 2. Both tables have the same interpretation, if ($B = 0$ and $D = 0$) or ($C = 1$) then the output is 1, while in any other cases, the output is 0. In an input pattern, the symbol "-" means that the value of a particular input variable is irrelevant to determine the output. In an output pattern, we used "-" to denote that the output for a given input pattern is unknown or of no interest (don't care).

Table 2 Binary function after LMM

Pattern/Rule	Inputs				Output
	A	B	C	D	
1	-	0	-	0	1
2	-	-	1	-	1

Techniques used on the simplification of Boolean functions include Karnaugh maps, Quine and McCluskey algorithm [14], and ESPRESSO [13]. However, the software in which the most sophisticated algorithms are programmed has a limit of 22 input variables [15]. Their limitation resides in the use of the Read Only Memories (ROM) in which the algorithm is implemented. A ROM implementation of an n-input function requires 2^n memory cells which limits the effectiveness in terms of silicon area. Our problem has a minimum of 315 input variables, therefore we have written out own version, suitable for epitope prediction.

2. Methods

Peptide sequences and their binding affinities for the allele *HLA-DRB1*0101* were collected from the immune epitope database (IEDB) [16]IEDB contains data related to antibody and T cell epitopes for humans, non-human primates, rodents, and other animal species. *HLA-DRB1*0101* has been previously tested using computational methods [17]–[19]. *HLA-DRB1*0101* is associated with rheumatoid arthritis [20], foliaceous pemphigus [21], in delayed mediated hemolytic transfusion reactions [22], and in lyme disease arthritis, *HLA-DRB1*0101* appears to play a role in presentation of triggering microbial antigens [23].

We removed duplicated epitopes and unnatural peptides with more than 75% alanine [24]. A total of 5348 experimentally confirmed unique MHC-II epitopes were obtained, with a length distribution between 8 and 35 amino acids. To make the data suitable for LMM it is necessary to represent each amino acid using the binary encoding scheme represented by a 21-bit binary vector, where 20 bits are set to zero and 1 bit is set to 1. This scheme accounts for the 20 essential amino acids and unknown residues. When binary encoded, an epitope of length 15 will have 15*21=315 variables, and an epitope of length 35 will have 735 variables. Therefore, in order to reduce the number of input variables we selected epitopes with the most frequent lengths, and used SVM feature selection (FS) capabilities for selecting the top variables. In our

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

dataset, 94.5% of the peptides (5053) have lengths on the range of 12 to 15 amino acids, the majority being of length 15 (92.3%). From the 5053 epitopes selected, we randomly chose 3744 (74%) as a training set and 1309 (26%) as a testing set. Every epitope is represented by a vector of 315 binary variables, for epitopes with less than 15 amino acids long we added at the end of the sequence "unknown" amino acids to have binary vectors of equal lengths. The half maximal inhibitory concentration (IC_{50}) is the concentration of an inhibitor at which the response (or binding) is reduced by half. Using the training set, we categorized as binders those peptides with an IC_{50} value less than 300 and non-binders those peptides with an IC_{50} value greater than 300. With this scheme we obtained 3179 binders and 566 non-binders. We used SVM's feature selection capabilities with "SVM and Kernel Methods Matlab Toolbox" [25] to rank all the variables and select the most relevant (70 variables).

To apply LMM, the information must be encoded in binary fashion, therefore a series of contradictions can occur in the dataset when feature selection is applied, that is, the same pattern of input values can produce different output values. For example, both of the following 10-input variables and their corresponding outputs are different:

Input Pattern	Output
0 1 0 0 0 1 0 0 0 1	0
0 1 1 0 0 1 0 1 0 1	1

However, suppose that variables 1, 2 and 10 are the most representative of those input patterns; then, after selecting them, we have the same input pattern with different output, hence a contradiction:

Input Pattern	Output
0 1 1	0
0 1 1	1

The fewer variables are used, the more contradictions are obtained. Table 3 shows the percentage of contradictions after running SVM-FS with our data. The axioms of digital logic would not work when contradictions are present, therefore, is very important to erase them before applying LMM.

Table 3 Number of variables and percentage of contradictions after SVM-FS

Number of variables	Percentage of contradictions
20	84%
30	76%
40	58%
50	44%
60	23%
70	14%
80	12%
90	7%
100	3%
315	0%

We selected the top 70 variables based on Cruz-Cano et al. (2012). In their study the top selected variables contained 14 percent of contradictions, but were able to describe the entire dataset. After our variable selection we removed the contradictions erasing those patterns with an output value of 0 (non-binder) and leaving only output patterns of 1 (binder). The more times an input pattern repeats, the more relevant to describe the data, our top input pattern appeared 65 times with an output value of 1, and 3 times with an output value of 0. We selected the top 76 input patterns that represent 24.3 percent of the data (Table 4).

Given that the existing software for LMM can only handle a limited number of variables, we developed our own version. First, we decoded our chosen 70 variables from binary code back into amino acid form. Every position in the epitope can contain any one of the 20 amino acids plus an unknown amino acid. Since we used a 21-bit binary vector to represent every amino acid; variables from 1 to 21 are at position 1, variables 22 to 42 are at position 2, and so on. Below is the order in which amino acids can appear in every position.

A R N D C E Q G H I L K M F P S T W Y V –

For example, our first top five variables are 5, 4, 6, 1, and 309. The first four variables are at position 1 and the fifth variable is at position 15. If the variable acquires the value of 0 in any position, it is interpreted as not being certain amino acid, and vice versa. In Table 5 we show a decoding example with the top 5 variables.

We selected the variables that could be found in every pattern, for example, if position 11 is NOT D in every pattern, we removed that position from the rest and recorded this rule for later use. Position 1 was the position with more variation, as every amino acid could be found in such position of the epitope; therefore, we selected unique patterns considering common amino acid appearance in position 1. After removing all common variables at every position and recording the rules, we selected unique patterns and removed the rest.

Table 4 Top input patterns and percentage of data description

Number of binary variables	Percentage of samples without contradictions
20	15.79%
30	17.80%
40	19.47%
50	21.04%
60	22.30%
70	23.56%
76	24.32%
80	24.69%
90	25.64%
100	26.58%
500	48.88%
1000	64.61%
1500	80.34%
2000	96.07%
2126	100.00%

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

We compared the results of our LMM algorithm with two classification techniques, SVM and ℓ_1-minimization [26]. Both have been used on epitope prediction. Standard SVM takes a set of input data and predicts for each given input which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier [27]. Here we use the implementation of SVM available in MATLAB as part of the Statistics Toolbox. ℓ_1-minimization techniques provide a satisfactory method to solve sparse representation problems

Table 5 Binary encoding and amino acid decoding of top 5 variables

V5	V4	V6	V1	V309
0	0	0	0	0
0	0	0	0	0
0	0	0	1	0
0	0	0	0	0
0	0	0	0	0
0	0	1	0	0
0	0	0	0	0

P1	P1	P1	P1	P15
NOT C	NOT D	NOT E	NOT A	NOT P
NOT C	NOT D	NOT E	NOT A	NOT P
NOT C	NOT D	NOT E	A	NOT P
NOT C	NOT D	NOT E	NOT A	NOT P
NOT C	NOT D	NOT E	NOT A	NOT P
NOT C	NOT D	E	NOT A	NOT P
NOT C	NOT D	NOT E	NOT A	NOT P

V = variable, P = position

3. Results and discussion

Since we used a 21-bit binary vector to represent every amino acid and the maximum length of the epitopes selected was 15 amino acids, we obtained 315 input variables. After applying SVM-FS we selected the 70 top variables. With LMM, our selected variables were minimized to 33 input patterns, representing the set of rules for allele *HLA-DRB1*0101* (Table 7). Taking rule 1 as example, its interpretation will be the following:

If "amino acid at P1 is L" and "amino acid at P2 is not C or G" and "amino acid at P3 is not D or R" and "amino acid at P4 is not P, Q, D, G or E" and "amino acid at P5 is not E or E" and "amino acid at P6 is not R or E" and "amino acid at P7 is not E, D or S" and "amino acid at P8 is not D or R and amino acid at P9 is not P" and "amino acid at P10 is not D or F" and " amino acid at P11 is not D or I" and "amino acid at P12 is not D or K" and "amino acid at P13 is not T" and "amino acid at P14 is not P or E" and "amino acid at P15 is not T or amino acid is D" then the sequence is a binding epitope.

A great variety of residues are well tolerated at P1, leucine is the preferred amino acid in this position, but it can accommodate any amino acid in rule 20. Several positions disfavor most of the time acidic polar (negative) amino acids. Aspartic acid is not tolerated at positions 3, 4, 5, 7, 8, 10, 11, 12, and 15; whereas glutamic acid is avoided at positions 4, 5,7,10 and 14. Proline is deleterious at positions 4, 9 and 14; similarly, arginine is not tolerated at positions 3, 6 and 8. Rule 31 applies for epitopes of length 12, whereas rules 32 and 33 apply for epitopes of length 13 and 15; no rules where generated for epitopes of length 14.

We developed a Perl script to test the set of rules obtained after LMM and we applied this script to our testing set. We measured sensitivity, specificity, accuracy and Mathew's Correlation Coefficient as shown in table 6. Notice that LMM achieve a performance similar to that of machine learning techniques, when such techniques use binary vector representation to encode amino acids. In other studies [28], encoding schemes that make use of physicochemical properties of amino acids yield better results. However, the axioms of digital logic require an encoding scheme that avoids the largest number of contradictions, and only the binary representation produces the least amount.

Table 6 Prediction accuracy

	LMM	SVM	ℓ_1
Sensitivity (%)	90	100	97
Specificity (%)	69	2	9
Accuracy (%)	81	83	82
MCC	0.6	0	1.9

4.Conclusions

We have illustrated the use of LMM in the identification of potential MHC-II epitopes. Our method should be valid for predicting MHC-I epitopes and other applications where motif identification is necessary. The task of rule extraction from LMM is to devise rules directly from the data rather than from the model itself. The main feature of our model is that it is capable to generalize from its experience by assigning output values to previously unseen input patterns given a training dataset; in this sense, our algorithm is similar to machine learning methods. The main advantage of our proposed method is the generation of a set of rules expressed in a language that can be understood by any researcher interested in studying MHC-II epitope binding.

5. Acknowledgments

This work was conducted while a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Awards #EF-0832858 and #DBI-1300426, with additional support from The University of Tennessee, Knoxville.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 7 Rules for allele *HLA-DRB1*0101*

Rules	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
1	L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(T),D
2	S,K	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P),E	~(D,T)
3	E,L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D),K	~T	~(P,E)	~(D,T)
4	L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(K),D	~T	~(P,E)	~(D,T)
5	L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D),I	~(D,K)	~T	~(P,E)	~(D,T)
6	D	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(I),D	~(D,K)	~T	~(P,E)	~(D,T)
7	G	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D),F	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
8	S,E	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(F),D	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
9	D,E	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(R),D	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
10	A,D	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,S),D	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
11	R	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(E),R	~(D,S),E	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
12	A,L	~(C,G)	~(D,R)	~(P,Q,G,E),D	~(E),D	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
13	G,T	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(D),E	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
14	S	~(C,G)	~(D,R)	~(P,Q,D,G),E	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
15	L	~(C,G)	~(D,R)	~(P,Q,D,E),G	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
16	V,L	~(C,G)	~(D,R)	~(P,Q,G,E),D	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
17	L,S,V	~(C,G)	~(R),D	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
18	K	~(C),G	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
19	L	~(G),C	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
20	**	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
21	I,G	~(C,G)	~(D),R	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
22	E	~(C,G)	~(D,R)	~(Q,D,G,E),P	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
23	L	~(C,G)	~(D,R)	~(P,D,G,E),Q	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
24	D,L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(E),R	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
25	V,L,D,E,K,I,T	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D),S	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
26	G,S.K	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D),R	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
27	S	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D,T)
28	I,S,L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	T	~(P,E)	~(D,T)
29	L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(E),P	~(D,T)
30	N,L	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	~T	~(P,E)	~(D),T
31	K	~(G),C	~(D,R)	~(P,Q,D,G),E	~(E,D)	~(R),E	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	-	-	-
32	P,S	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D,S)	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	T	~-	-
33	Q	~(C,G)	~(D,R)	~(P,Q,D,G,E)	~(E,D)	~(R,E)	~(E,D),S	~(D,R)	~P	~(D,F)	~(D,I)	~(D,K)	T	~-	~(P),E

**Any amino acid
~Negation

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

6. References

[1] J. S. Murray, S. D. S. Fois, T. Schountz, S. R. Ford, M. D. Tawde, J. C. Brown, and T. J. Siahaan, "Modeling alternative binding registers of a minimal immunogenic peptide on two class II major histocompatibility complex (MHC II) molecules predicts polarized T-cell receptor (TCR) contact positions.," *J. Pept. Res.*, vol. 59, no. 3, pp. 115–22, Mar. 2002.

[2] S. Gaseitsiwe, D. Valentini, S. Mahdavifar, M. Reilly, A. Ehrnst, and M. Maeurer, "Peptide microarray-based identification of Mycobacterium tuberculosis epitope binding to HLA-DRB1*0101, DRB1*1501, and DRB1*0401.," *Clin. Vaccine Immunol.*, vol. 17, no. 1, pp. 168–75, Jan. 2010.

[3] D. Suckau, J. Köhl, G. Karwath, K. Schneider, M. Casaretto, D. Bitter-Suermann, and M. Przybylski, "Molecular epitope identification by limited proteolysis of an immobilized antigen-antibody complex and mass spectrometric peptide mapping.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 24, pp. 9848–52, Dec. 1990.

[4] A. Patronov, I. Dimitrov, D. R. Flower, and I. Doytchinova, "Peptide binding prediction for the human class II MHC allele HLA-DP2: a molecular docking approach.," *BMC Struct. Biol.*, vol. 11, no. 1, p. 32, Jan. 2011.

[5] M. Nielsen, S. Justesen, O. Lund, C. Lundegaard, and S. Buus, "NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure.," *Immunome Res.*, vol. 6, no. 1, p. 9, Jan. 2010.

[6] E. Petsalaki, A. Stark, E. García-Urdiales, and R. B. Russell, "Accurate prediction of peptide binding sites on protein surfaces.," *PLoS Comput. Biol.*, vol. 5, no. 3, p. e1000335, Mar. 2009.

[7] M. Nielsen, C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund, "Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach.," *Bioinformatics*, vol. 20, no. 9, pp. 1388–97, Jun. 2004.

[8] H. Bui, J. Sidney, B. Peters, M. Sathiamurthy, a Sinichi, K. Purton, B. Mothe, F. Chisari, D. Watkins, and a Sette, "Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications," *Immunogenetics*, vol. 57, no. 5, pp. 304–314, Jun. 2005.

[9] C. W. Tung, M. Ziehm, A. Kämper, O. Kohlbacher, and S.-Y. Ho, "POPISK: T-cell reactivity prediction using support vector machines and string kernels.," *BMC Bioinformatics*, vol. 12, no. 1, p. 446, Jan. 2011.

[10] N. Barakat and J. Diederich, "Eclectic Rule-Extraction from Support Vector Machines," vol. 2, no. 1, pp. 59–62, 2005.

[11] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, and L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network," *Bioinformatics*, vol. 14, no. 2, pp. 121–130, 1998.

[12] M. Bhasin and G. P. S. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," *Bioinformatics*, vol. 20, no. 3, pp. 421–423, Feb. 2004.

[13] K. A. Bartlett, R. K. Brayton, G. D. Hachtel, and A. R. Wang, "Multilevel Logic Minimization Using Implicit Don ' t Cares," vol. 7, no. 6, 1988.

[14] T. K. Jain, D. S. Kushwaha, and a. K. Misra, "Optimization of the Quine-McCluskey Method for the Minimization of the Boolean Expressions," *Fourth Int. Conf. Auton. Auton. Syst.*, pp. 165–168, Mar. 2008.

[15] R. K. Brayton, Gary D. Hachtel, C. McMullen, *Logic Minimization Algorithms for VLSI Synthesis*. Kluwer Academic Publishers, 1984.

[16] Y. Kim, J. Ponomarenko, Z. Zhu, D. Tamang, P. Wang, J. Greenbaum, C. Lundegaard, A. Sette, O. Lund, P. E. Bourne, M. Nielsen, and B. Peters, "Immune epitope database analysis resource.," *Nucleic Acids Res.*, vol. 40, no. Web Server issue, pp. W525–30, Jul. 2012.

[17] J. Hammertt, C. Belunist, D. Bolin, J. Papadopoulost, R. Walskyt, J. Higelin, W. Danho, F. Sinigagliat, and Z. A. Nagyt, "High-affinity binding of short peptides to major histocompatibility complex class II molecules by anchor combinations," vol. 91, no. May, pp. 4456–4460, 1994.

[18] K. W. Jørgensen, S. Buus, and M. Nielsen, "Structural properties of MHC class II ligands, implications for the prediction of MHC class II epitopes.," *PLoS One*, vol. 5, no. 12, p. e15877, Jan. 2010.

[19] H. Noguchi, R. Kato, T. Hanai, Y. Matsubara, H. Honda, V. Brusic, and T. Kobayashi, "Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules.," *J. Biosci. Bioeng.*, vol. 94, no. 3, pp. 264–70, Jan. 2002.

[20] A. Kapitány, E. Zilahi, S. Szántó, G. Szücs, Z. Szabó, A. Végvári, P. Rass, S. Sipka, G. Szegedi, and Z. Szekanecz, "Association of rheumatoid arthritis with HLA-DR1 and HLA-DR4 in Hungary.," *Ann. N. Y. Acad. Sci.*, vol. 1051, pp. 263–70, Jun. 2005.

[21] G. J. del Mar Sáez-de-Ocariz M, Vega-Memije M, Zúñiga J, Salgado N, Ruíz J, Balbuena A,

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Domínguez-Soto L, "HLA-DRB1*0101 is associated with foliaceous pemphigus in Mexicans," *Int J Dermatol*, vol. 44, no. 4, p. 350, 2005.

[22] D. Reviron, I. Dettori, V. Ferrera, D. Legrand, M. Touinssi, P. Mercier, P. de Micco, and J. Chiaroni, "HLA-DRB1 alleles and Jk(a) immunization.," *Transfusion*, vol. 45, no. 6, pp. 956–9, Jun. 2005.

[23] A. C. Steere, W. Klitz, E. E. Drouin, B. a Falk, W. W. Kwok, G. T. Nepom, and L. A. Baxter-Lowe, "Antibiotic-refractory Lyme arthritis is associated with HLA-DR molecules that bind a Borrelia burgdorferi peptide.," *J. Exp. Med.*, vol. 203, no. 4, pp. 961–71, Apr. 2006.

[24] M. Rajapakse, B. Schmidt, L. Feng, and V. Brusic, "Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms.," *BMC Bioinformatics*, vol. 8, p. 459, Jan. 2007.

[25] C. Stéphane, "SVM and kernel methods matlab toolbox," *Percept. Syst. Inf.*, vol. 21, 2005.

[26] R. Sanchez, M. Argaez, and P. Guillen, "Sparse representation via 1-minimization for underdetermined systems in classification of tumors with gene expression data," *2011 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2011, pp. 3362–3366, 2011.

[27] L. Wang, *Support Vector Machines: Theory and Applications*. Springer Verlag Berlin, 2005.

[28] F. Tian, L. Yang, F. Lv, Q. Yang, and P. Zhou, "In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach," *Amino Acids*, vol. 36, no. 3, pp. 535–554, Mar. 2009.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Automatic EEG bad epoch and artifact removal using clustering

Elizabeth Hames and Mary Baker
liz.hames@ttu.edu and mary.baker@ttu.edu
Department of Electrical Engineering
Texas Tech University
Lubbock, TX, U.S.A.
Submitted for consideration for the 2014 BIOCOMP Conference

Abstract— **This paper presents a method for automatically identifying and cleaning artifactual independent components from EEG data. Artifactual components are identified through first clustering components and then labeling artifactual clusters. Clustering allows for identification of many types of artifactual components, including eye movements, electrocardiogram signals, electromyogram signals, movement, and bad channels. The proposed artifact detection method is analyzed using EEG data acquired with an EGI 65-channel net. The benefits of cleaning components compared to removing components are tested with a simulated dataset. Results indicate that cleaning components is beneficial when potential cerebral signals are falsely identified as artifactual components or are mixed into artifactual components.**

Keywords—EEG, artifact removal, ICA, clustering, components

I. INTRODUCTION

EEG data undergo several steps of preprocessing before analysis to remove non-cerebral, or artifactual, contributions to the data. Artifact removal can be difficult because the spectrum, topography, and waveform shape of many physiological artifacts are identical to cerebral EEG activity [1]. Independent component analysis (ICA) separates the EEG data from the artifactual signal, leaving the data analyst with the task of identifying and removing artifactual components [2]-[4]. Identifying artifactual components is not a simple process. It involves viewing component time courses, topographies, and spectra. This entire process can be time consuming since the number of components is usually equal to the number of channels. Some artifacts are very ambiguous and difficult to label, making component selection subjective and prone to error. Automated component selection solves the issues of manual component selection by providing consistency and speed. This work proposes a method for automatically identifying artifactual components using clustering implemented through an eight step process called ABEAR (automatic EEG bad epoch and artifact removal).

A. Previous Artifact Removal Methods

Many other automated methods exist for artifact removal, but this method combines qualities of each to meet the following goals: 1) Detect multiple types of artifacts; 2) Automate independent component selection; 3) Eliminate the need for a supplemental signal; 4) Clean components.

Many previous methods for EEG artifact removal focus on the removal of only one or two types of artifacts. For example, filtering only removes artifacts outside of the EEG frequency band, such as DC drift or power line noise. Wavelet analysis is only capable of detecting artifacts that resemble a mother wavelet. We propose clustering as a solution to the problem of limited artifact detection. Clustering separates cerebral data from artifactual data regardless of artifact type. Through clustering, we automatically identify five types of artifactual components: eye movements (EOG), electromyogram signals (EMG), electrocardiogram signals (ECG), artifacts caused by movement (MOV), and artifacts caused by bad channels (BAD).

If not automated, artifact removal methods such as wavelet analysis or ICA require manual component or coefficient selection [2]-[6]. Manual artifact selection is time-consuming, inconsistent, and prone to human error. In this work, artifactual clusters are labeled through an outlier criterion to reduce human interaction, similar to [7]. It dramatically reduces the time spent on manually selecting artifactual components.

Many automated artifact removal methods, such as adaptive filtering or time/frequency regression, require supplemental signals, such as a measured ECG or EMG signal or even a synthetic signal [8]-[11]. Automated methods for component selection, such as correlation with a reference signal or supervised learning, also need training signals [12]-[14]. Additional signals are not always available or accurate for all participants. The artifact detection method presented in this work does not need any additional reference signals. The use of clustering for artifactual component identification is an unsupervised learning method that groups the data without training signals. Using an outlier criterion to label artifactual clusters eliminates the need for a reference signal, similar to [15].

Removing data, removing artifactual ICA components, or even wavelet thresholding, can result in throwing away useful information [16]-[18]. This work presents cleaning components as an alternative to completely removing them. The benefit to cleaning components is demonstrated in the Results section using a simulated dataset.

In summary, the proposed artifact removal method, ABEAR, separates EEG data into ICA components and classifies those components using an unsupervised clustering algorithm. Clustering allows the detection of five types of artifactual components: EOG, EMG, ECG, MOV, or BAD. Identified artifactual components are cleaned to reduce data loss.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

B. EEG Artifactual Components

Each type of artifactual component has unique properties, such as frequency spectrum, topography, or waveform shape, that distinguish it from EEG data or other artifacts. This section describes properties of five types of artifactual components.

EOG artifacts are the result of blinks, horizontal eye movements, or vertical eye movements. They have various shapes depending on the movement of the eyes. Blinks are characterized by high amplitude impulses while horizontal and vertical eye movements have rectangular-like shapes. The simultaneous occurrence of blinks and eye movements results in unpredictable waveform shapes [19]. All EOG components are located in the frontal channels near the eyes. In a topography map of EOG components, blinks have a single pole centered between the eyes and while eye movements usually have two opposite poles. The spectrum of EOG artifacts is dominated by low frequencies, below 5 Hz. Many physiological actions can result in EMG artifacts: clenching muscles, movement of the head, neck, or shoulders, or squinting and twitching of the eyes. EMG artifacts can take two forms: white-noise-like pattern and "railroad-cross-tie" pattern [20]. On topography maps, EMG artifacts are usually located in the peripheral channels (around the neck, ears, and eyes). EMG artifacts have a uniform spectrum [20]. ECG artifacts have fairly periodic QRS spikes [21]. The ECG waveforms repeat about once per second. ECG topographies show a bi-polar pattern across the entire scalp and their spectrum has mainly low-frequency dominance. MOV artifacts refer to any artifact caused by movement that is not an EMG waveform, i.e., shoulder or neck movement or repositioning. MOV artifacts tend to have very low-frequency dominance similar to eye movements and their topographies show peripheral channel localization. BAD artifacts refer to artifactual components whose topographies indicate localization to a single channel. BAD artifacts can result from high impedance electrodes, electrodes disturbed by touch, or misplaced electrodes. BAD components do not have a common shape or spectrum.

II. METHODS

We employed an eight step process to detect and remove EEG artifactual components. , as seen in Fig. 1. The following sections describe the eight steps in more detail.

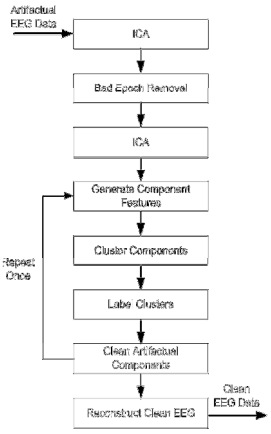


Fig. 1. Steps of ABEAR.

A. Bad Epoch Removal

"Bad" epochs are time-segments with large amounts of artifacts across several components. Bad epochs are instances where an artifact was not separable into just one component. Removing time-segments with inseparable artifacts can improve ICA and artifactual component identification. ICA must be computed before bad epoch removal since independent components are used to identify bad epochs. Variance and spectrum of components are used in identifying bad epochs.

The steps for bad epoch removal are as follows. The variance, *var*, and the spectrum, *spect*, between .5 and 50 Hz are computed for each epoch of each component. Next, the mean spectrum, *Savg*, and mean variance, *Vavg*, for each component are determined by (1) and (2). The spectrum error, *dS*, and variance error, *dV*, given by (3) and (4), are calculated for each epoch and each component using *Savg* and *Vavg*. Within each epoch, components are labeled "bad" if *dS* or *dV* are greater than a threshold, *Error*. Finally, an epoch is labeled as "bad" if it contains more bad components than a threshold, *perbad*ncomp*. To improve the detection of bad epochs, the selection process is repeated with adjusted means that exclude previously labeled bad epochs. ICA is recomputed on the EEG dataset after all bad epochs are removed. The newly generated components are the input to the component features step. The thresholds used for bad epoch detection were selected based on a performance comparison using manually labeled bad epochs.

$$Savg(i) = \frac{\sum_{j=1}^{nepoch} spect(i,j)}{nepoch}, i = 1..ncomp \quad (1)$$

$$Vavg(i) = \frac{\sum_{j=1}^{nepoch} var(i,j)}{nepoch}, i = 1..ncomp \quad (2)$$

$$dS(i,j) = \frac{|spect(i,j) - Savg(i)|}{Savg(i)}, i = 1..ncomp, j = 1..nepoch \quad (3)$$

$$dV(i,j) = \frac{|var(i,j) - Vavg(i)|}{Vavg(i)}, i = 1..ncomp, j = 1..nepoch \quad (4)$$

B. Component Features

Component features are measurements taken from each component used for clustering. Eight 1-dimensional features were generated for each component: topography histogram, spectrum fit 1, spectrum fit 2, frontal location score, peripheral location score, average auto-correlation, Lorentz threshold, and sym3 repeatability. The eight features were normalized and then combined into one 8-dimensional feature.

The eight features are similar to features tested by [15]. The features were chosen as a representative of component properties commonly used to manually identify artifactual components. The eight features are chosen over single features such as spectrum, topography, etc. for two reasons. First, they provide consistency across participants for the five types of artifactual components. Time courses, spectrums, and even topographies can vary across participants within a given

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

category of artifactual component. Representing the components with multiple features reduces dependency on a single measurement, providing more consistency across participants. Second, the single dimensional features allow clustering of multiple types of artifactual components. No one feature provides good separation of all five types of artifactual components. For example, spectrum would be excellent for separating EOG and EMG components, but not necessarily BAD or MOV components since their spectrums are inconsistent. The eight one-dimensional features address the properties of each type of artifactual component.

The histogram of a component's topography explains its scalp power distribution. The first number of the histogram array is the number of channels for which the component is least distributed. If the first number of the histogram is large (>.9*nchan), the component is localized to one channel. This is useful for detecting BAD or MOV artifacts.

The shape of a spectrum can be described using a second order polynomial, given by (5). The first coefficient, C_1, describes concavity and the second coefficient, C_2, describes the slope. The first and second coefficients of the second order polynomial spectrum fit are obtained as features, spectrum fit 1 and spectrum fit 2. The concavity and slope are very large for EOG artifacts and very small for EMG artifacts.

$$S = C_1x^2+C_2x+C_3 \tag{5}$$

The frontal location score describes the amount of component power located in the frontal channels. Given a component's topography, the frontal location score is the fraction of channels with a relative power greater than 40% of the maximum relative power that are located in frontal channels. Frontal channels are designated as channels within -55 to 55 degrees from center and with a radius greater than .4 from Cz. Similarly, given a component's topography, the peripheral location score is the fraction of channels with a relative power greater than 40% of the maximum relative power that are located in peripheral channels. Peripheral channels are designated as channels that have a radius greater than .4 from Cz.

The Lorentz threshold is used to quantify the amount of component signal described by a sym3 wavelet. It has previously been used to despike signals, meaning it would be largest for EOG or ECG components [22]. The signal, C, used for computing the threshold, is the approximate wavelet coefficient for the sym3 wavelet, shown in (6).

$$L = \sqrt{\frac{\sum_1^N {C_i}^2}{N}} \tag{6}$$

The sym3 repeatability feature uses the Lorentz threshold to estimate the number of spikes in a component. The repeatability of a component is the number of values within the sym3 approximation coefficient that are greater than 2*L.

C. Clustering

Common clustering algorithms, such as k-means, require a user specified number of clusters. It is not always clear how many groups exist in a set of components and this number can change across datasets. The Isodata clustering algorithm iteratively finds the optimal number of clusters by splitting and merging clusters [23]. This is beneficial for clustering EEG components since the number of artifactual components changes across datasets. In Isodata, clusters are split based on a standard deviation threshold and merged based on a separation distance threshold. Although Isodata can automatically determine the number of clusters, there are many user defined thresholds that rely upon the properties of the data. A modified Isodata algorithm eliminates the need for these thresholds by using fuzzy membership functions [24]. The modified Isodata algorithm is used to cluster the component features. The details of the modified Isodata algorithm will not be discussed in this paper to conserve space. However, details can be obtained through the author.

Three parameters determine the initial conditions for clustering: initial cluster centers, maximum number of iterations, and the number of repetitions. The results presented in this paper are generated from the following initial conditions. Initial cluster centers are randomly selected from the feature vectors. The maximum number of iterations is set to 100 and final clusters are selected from the best of three repetitions of clustering.

D. Labeling Clusters

Isodata segments component features into unlabeled groups. The artifactual clusters must be distinguished from non-artifactual clusters. Artifactual clusters have outlying average 1-dimensional features, so we implement an outlier criterion to determine which clusters are artifactual. The 1-dimensional features are averaged for members of each cluster. The average 1-dimensional features, *f*, are used to compute the percent difference of each cluster from the average of all clusters, given by (7). Clusters whose percent difference is greater than .75 are labeled as artifactual.

$$Percent\ Difference = \frac{|f_i - \sum_{i=1}^{m} f_i|}{standard\ deviation(\boldsymbol{f})}, i = 1..m \tag{7}$$

E. Component Cleaning

Rather than removing artifactual components, we implement a cleaning process to prevent a loss of cerebral EEG data. All artifactual components undergo despiking and denoising as part of the cleaning process.

1) Despiking: Despiking removes large spikes within a signal, such as blinks or heart beats. Despiking can be accomplished through wavelet thresholding [22],[24]. Sym3 wavelet coefficients for four levels are computed for overlapping segments of length one second. All values of the coefficients are thresholded using the level-dependent threshold in (8) [26]-[27], where MAD is the median absolute deviation of the wavelet coefficient, x. Dividing by .6745

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

normalizes the threshold by the standard deviation for Gaussian white noise [28]. All values that exceed the threshold are replaced using a cubic interpolation. The thresholding process is repeated until all values in the coefficients are below the threshold.

$$Threshold = \frac{4 * MAD(|x|)}{.6745} \tag{8}$$

2) Denoising: Denoising is performed on the same segments from despiking. Denoising is also accomplished through wavelet thresholding. Db8 wavelet coefficients are computed for each segment using four levels. All values of the coefficients are thresholded using the level-dependent threshold in (9) [28]-[29]. All values below the threshold are replaced using a soft thresholding.

$$Threshold = \frac{MAD(|x|)}{.6745} \tag{9}$$

III. EEG DATA

Both recorded data and simulated data are utilized to evaluate the artifact detection and removal process. The recorded data provides ground truth for clustering and component labeling. The simulated data provides ground truth for cleaning components. Both types of data are described in the following sections.

A. Recorded Data

In a previous, unrelated study, we acquired EEG data from 62 participants using a 65-channel net. The session consisted of 42 cognitive tasks. Each participant's dataset was epoched at the start of each task for six seconds. All 62 datasets were filtered with a 1 Hz high-pass filter and a 50Hz low-pass filter and epoched around 23 event markers for six seconds (for a total of 138 seconds). Each dataset was decomposed into 65 components using Infomax ICA. An artifactual version and a manually cleaned version of the recorded data was saved for use in this work. The manually cleaned version was acquired from removing bad epochs followed by removing artifactual ICA components. EEGLab was used to perform ICA, view properties of the components, and remove artifactual epochs and components [30]. Manually selected artifactual components were used to test component classification rates. It is important to note that only one expert provided labeled components. In order to reduce bias in component labeling, further experts need to be consulted in the future. Also, only data acquired from an EGI net is used to verify the proposed artifact removal algorithm. Generalization cannot be made for other EEG systems until further testing is done.

B. Synthetic Data

Simulated EEG data are created by forward projecting six artifactual sources onto a scalp and adding background EEG noise to the resulting data. A blink source is created by convolving a single blink signal, extracted from a recorded EEG time series, with a series of randomly generated impulses. The impulses are spaced between .5 and 1 second apart. Similarly, an ECG source is created by convolving an ECG signal, simulated using MATLAB's *ecg* function, with a series of randomly generated impulses [31]. The impulses are spaced between 1 and 1.7 seconds apart. Three EMG sources are created from white Gaussian noise. Lastly, a cross-tie EMG pattern is created with a train of Gaussian pulses [20]. A forward model of the sources is created using BESA EEG simulation software, available at *www.besa.de*. Dipoles are positioned at the front of the head, middle of the head, and temporal and parietal locations corresponding to the blink, ECG, and EMG signals respectively. A thirty-three channel artifactual dataset is generated from the forward model. EEG background noise is added to each channel of the artifactual data. The background noise is generated to mimic the frequency spectrum and amplitudes of EEG data [32]. A non-artifactual dataset is also created from the EEG background noise to test component cleaning.

IV. RESULTS AND DISCUSSION

A. Bad Epoch Removal

The selection of bad epochs is dependent upon two user specified thresholds: *Error* and *perbad*. Values for *Error* and *perbad* are selected as .15 and .8, respectively, to maximize the precision and recall as calculated by (10) and (11). The manually labeled bad epochs from the 62 recorded datasets are used to determine the true positives, false positives, and false negatives generated by the automatic labeling of bad epochs. The chosen thresholds result in precision and recall rates of about .7. While classification is not perfect, precision and recall rates are acceptable. Manual classification is very subjective, leading to inconsistency and possible mislabeling. Automatic selection of bad epochs provides consistency.

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{10}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{11}$$

B. Component Features

Three datasets are randomly selected from the 62 recorded datasets to display examples of the 1-dimensional features, shown in Fig.2. The eight features in order are topography histogram, spectrum fit 1 and 2, frontal and peripheral location scores, average auto-correlation, Lorentz threshold, and sym3 repeatability. The features are averaged for each of the five types of artifactual components, as determined by manual labeling: EOG, EMG, ECG, BAD, and MOV. The 1-dimensional features demonstrate consistency across participants. The EOG components are best distinguished by their large spectrum fit 1 and 2 values and high EOG scores.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Conversely, the EMG components are best distinguished by their very low fit1 and 2 values. The ECG components have the highest sym 3 repeatability as well as a very high EMG score and low topography histogram value. The BAD components all have a very high histogram topography value due to their localized nature and a very low Lorentz threshold. The MOV components all have a high EMG score.

C. Clustering

Initial cluster centers are randomly selected from the feature vectors. The maximum number of iterations is set to 100 and final clusters are selected from the best of three repetitions of clustering. Clustering was performed for three initial values of *m*, the number of clusters: 13, 16, and 19. The final number of clusters for these three values of *m*, averaged across all 62 datasets, are 21, 23, and 26 respectively.

Precision and recall, as measured by (12) and (13), are used to evaluate how well Isodata clusters the components into artifactual and non-artifactual clusters. The manually labeled components from the 62 recorded datasets are used to determine which components are artifactual and non-artifactual. High precision means that a cluster contains mostly artifactual components, indicating a good separation of artifactual and non-artifactual components. High recall means that the artifactual components are highly concentrated in a cluster, indicating that the artifactual components are not spread across many clusters.

$$Precision = \frac{number\ of\ artifactual\ components\ in\ a\ cluster}{total\ number\ of\ components\ in\ a\ cluster} \quad (12)$$

$$Recall = \frac{number\ of\ artifactual\ components\ in\ a\ cluster}{total\ number\ of\ artifactual\ components} \quad (13)$$

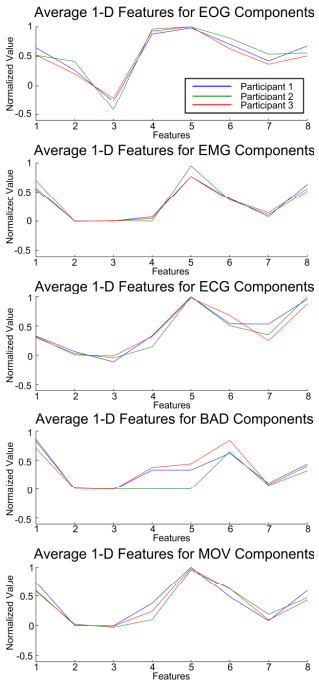


Fig. 2. 1-dimensional features averages across five types of artifactual components for three participants

Table 1 shows the percentage of clusters in all 62 datasets with a precision greater than .72 for each value of *m*. A large percentage of clusters having a precision greater than .72 does not necessarily indicate improved clustering; it could result from many single member clusters. Recall can provide additional information to evaluate clustering. Table 4 also shows the average recall for the clusters with a precision greater than .72 for each initial value of m. The modified Isodata results in a large percentage of clusters with a high precision as well as high recall rates for those clusters, and this outcome is not overly sensitive to the initial choice of *m*.

TABLE I. CLUSTERING RESULTS

	***m*=13**	***m*=16**	***m*=19**
Percentage of Clusters with Prec>.72	61%	68%	70%
Recall for clusters with Prec>.72	.63	.65	.68

The recall and precision verify that clustering adequately separates artifactual components from non-artifactual components, with no indication of how it separates individual types of artifactual components such as EOG, EMG, ECG, MOV, or BAD. A closer look at the composition of clusters shows how individual types of artifactual components are grouped. A survey of the types of components belonging to clusters that contain either non-artifactual, EOG, EMG, ECG, MOV, or BAD components was conducted for each of the 62 datasets. Table 2 displays the average results of the survey across the 62 datasets. The results in Table 2 are computed for ***m***=19. Artifacts can vary widely across participants so it is interesting to also observe the standard deviation of the cluster member survey. Table 3 presents the standard deviations for the corresponding averages from Table 2.

TABLE II. AVERAGE OF CLUSTER MEMBER SURVEY

Types of Clusters	**Types of components within each type of cluster**					
	Non-artifactual	***EOG***	***EMG***	***ECG***	***MOV***	***BAD***
Non-artifactual	44.0	2.0	3.7	0.3	0.6	1.8
EOG	3.2	7.1	1.1	0.0	0.1	0.4
EMG	6.0	0.9	7.8	0.0	0.2	1.1
ECG	0.6	0.0	0.1	1.1	0.0	0.2
MOV	0.7	0.1	0.2	0.0	1.5	0.1
BAD	3.4	0.7	1.3	0.2	0.1	3.6

TABLE III. STANDARD DEVIATION OF CLUSTER MEMBER SURVEY

Types of Clusters	**Types of components within each type of cluster**					
	Non-artifactual	***EOG***	***EMG***	***ECG***	***MOV***	***BAD***
Non-artifactual	5.9	2.5	3.1	0.5	1.0	1.7
EOG	3.2	3.0	1.5	0.3	0.7	0.6
EMG	5.5	1.0	4.7	0.2	0.5	1.2
ECG	1.3	0.3	0.3	0.6	0.0	0.5
MOV	1.4	0.5	0.5	0.0	1.4	0.5
BAD	3.4	1.2	1.6	0.4	0.4	2.1

The numbers along the diagonal of Table 2 represent the average number of each type of artifactual, or non-artifactual,

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

component. The off diagonal elements indicate the number of mixed components. For example, the top right number is the number of BAD components located within clusters that contain non-artifactual components. It would be ideal for all the numbers in the top row and far left column, except the top left cell, to be zero. This would indicate that all artifactual components are separated from non-artifactual components. This however, is not the case. The top row indicates that very few artifactual components are mixed into non-artifactual clusters. The left column indicates that there are large numbers of non-artifactual components mixed into EMG clusters. It is not necessarily ideal for all other off-diagonal elements to be zero. For instance, if some EMG components have similar properties of BAD components, then the corresponding EMG components might be assigned to BAD clusters. In this case, however, most of the non-diagonal elements are near zero.

In Table 3, most off-diagonal standard deviations are very low. The highest standard deviations occur for the average number of non-artifactual and EMG components as well as the number of non-artifactual components located in EMG clusters. Some standard deviations are higher than their averages, indicating that one or two datasets may have extreme outliers, with most datasets having similar values.

D. Cluster Labeling

Equations (10) and (11) are used to evaluate the results of cluster labeling. Following cluster labeling, all components are labeled as artifactual or non-artifactual depending on what type of cluster they belong to. The manually labeled components from the 62 recorded datasets are used to determine the true positives, false positives, and false negatives generated by cluster labeling. Precision and recall are computed from all datasets. For example, the number of true positives is the total number of true positives across all 62 datasets. Recall is calculated for each type of artifact: EOG, ECG, EMG, MOV, or BAD. Final classification results are given in Table 4. The EMG and BAD components result in the lowest recall rates while MOV and EOG components result in the highest recall rates. Recall can be increased by lowering the outlier threshold. However, this threshold adjustment results in a lower precision. Overall, the clustering and cluster labeling processes result in very good recall rates with a decent precision. Since the components are cleaned as opposed to removed, a lower precision is acceptable.

TABLE IV. COMPONENT CLASSIFICATION RESULTS

Average Precision	0.594
Average Recall	0.816
EOG Recall	0.891
EMG Recall	0.729
ECG Recall	0.815
MOV Recall	0.934
BAD Recall	0.728

E. Component Cleaning

1) Simulated Data: Seven components are manually identified as artifactual after the simulated dataset is decomposed using ICA. Fig. 3 shows a subset of the components after components cleaning. The EOG and ECG spikes are no longer present. EMG noise is reduced.

Both component removal and component cleaning are performed on the simulated data to observe the benefits provided by cleaning components. Channel spectra of reconstructed data from cleaning and removing components, as well as the original artifactual data, are compared to channel spectra of non-artifactual data using a distance measure, *dS*, given by (14). Fig. 4 shows the spectra comparisons for three frequencies. In (14), the "cleaned data spectra" is replaced by the removed component spectra or original data spectra to compare them to the non-artifactual data. Two sets of components are labeled as artifactual for the calculation of (14). The first set consists of seven truly artifactual components. The second set of components consists of the seven artifactual components with an additional ten randomly chosen components.

$$dS = |Nonartifactual\ Data\ Spectra - Cleaned\ Data\ Spectra| \quad (14)$$

Very little difference can be seen in spectra between cleaning and removing only the seven artifactual components. However, when more components are selected than necessary, a large difference between the data resulting from component removal and the non-artifactual data is observed across all frequencies. Component cleaning results in a smaller difference between spectra, indicating that it is beneficial for instances when too many components, or unnecessary components, are identified as artifactual.

2) Recorded Data: Two datasets are randomly selected from the 62 recorded datasets to view the results of component cleaning. Fig. 5 shows a subset of components for the two datasets before and after cleaning. All large spikes caused by EOG or ECG artifacts are completely removed. EMG components present more of a challenge for cleaning than EOG or ECG artifacts. The EMG components are reduced, while not completely removed.

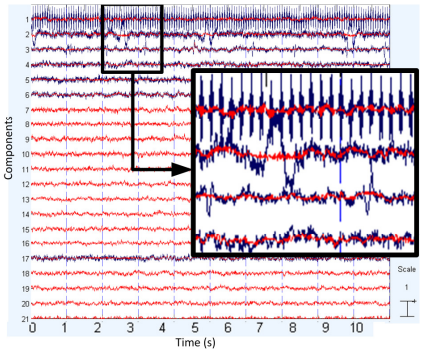


Fig. 3. Simulated data components before (blue) and after(red) cleaning.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

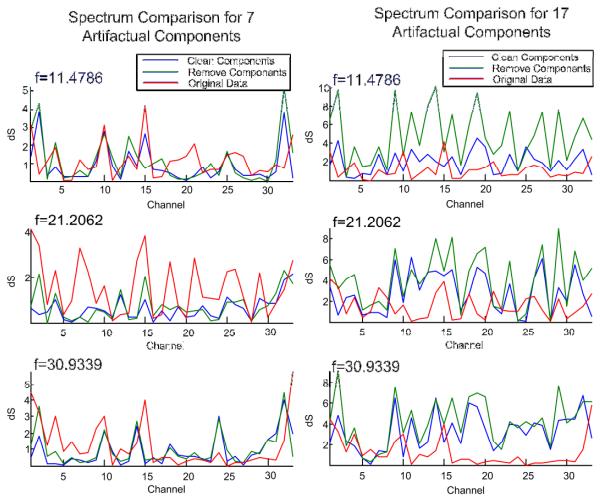


Fig. 4. Simulated data difference in spectra, *dS*, across all channels for three frequencies. The value for dS represents the difference between reconstructed data spectra, by either component removal, component cleaning, or the original data, and the non-artifactual data.

V. Conclusions and Future Work

This work evaluates the effectiveness of clustering for automated identification of artifactual EEG components as well as the benefits of cleaning components compared to removing components. The generation of components is improved through the automatic removal of bad epochs prior to ICA. The 8 1-dimensional features used for clustering demonstrate consistency across datasets. They also improve clustering speed through their low dimensionality. The 1-dimensional features were selected to represent features commonly used in manual classification. They are only a subset of possible features. A number of new features could be generated and tested using an algorithm like Sequential Forward Floating Search (SFFS) [33]. SFFS searches for the best combination of features to improve classification rates.

In this paper, an outlier criterion was used to label artifactual clusters. While the outlier detection resulted in acceptable recall and precision rates, it is not perfect and is dependent upon a threshold. Further exploration into possible methods for labeling clusters could improve classification rates and remove the dependency on a threshold.

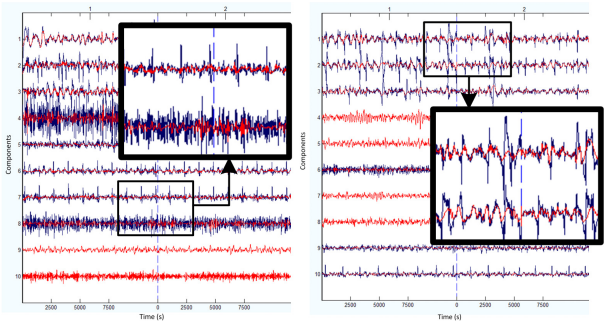


Fig. 5. Original (blue) and cleaned (red) components for two example datasets. The left figure magnifies an ECG and an EMG component while the right figure maginifies EOG components.

A simulated dataset demonstrated the benefits of cleaning versus removing components. When a large number of components are labeled as artifactual, as is common in practice, the channel spectrum resulting from cleaning components is more similar to the non-artifactual channel spectrum as compared to removing components. A possible explanation for the benefit of cleaning components is that even artifactual components contain some cerebral signal. ICA cannot separate noisy signals such as EMG, as well as PCA. For this reason, artifactual signals can be spread across several components that contain cerebral signals.

In general, the artifact removal method presented in this work successfully identified and removed artifactual contributions to EEG data acquired with a 65-channel EGI net according to classification rates obtained by one expert. In order to generalize this method to EEG data acquired on other systems, further testing must be completed and additional experts should provide components labels. The concept of clustering components is not limited to EEG data. It could be extended to other neuroimaging data such as fMRI or DTI. Features can also be customized to detect cerebral signals as well as artifactual, extending ABEAR to the use of detecting event related responses.

References

[1] D. W. Klass, "The continuing challenge of artifacts in the EEG," *Amer. J. EEG Technol.*, vol. 35, pp. 239-269, 1995.

[2] S. Makeig *et al.*, "Independent Component Analysis of Electroencephalographic Data," in *Advances in Neural Inform. Process. Syst.*, D. Touretzky, M. Mozer, M. Hasselmo, 8th ed. Cambridge, MA: MIT Press, 1996, pp.145-151.

[3] A.J. Bell, T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129-1159, 1995.

[4] T.P. Jung *et al.*, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp.163-178, 2000.

[5] V. Samar *et al.*, "Wavelet Analysis of Neuroelectric Waveforms: a conceptual tutorial," *Brain and Language*, vol. 66, pp. 7-60, 1999.

[6] S.J. Schiff *et al.*, "Fast wavelet transformation of EEG," *Electroencephalography and Clinical Neurophysiology*, vol. 91, pp. 442–455, 1994.

[7] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, pp. 229-240, 2010.

[8] P. He, G. Wilson, C. Russell, "Removal of ocular artifacts from ectro-enecphalogram by adaptive filtering," *Medical and Biological Eng. and Computing*, vol. 42, no. 3, pp. 407-412, 2004.

[9] A.G. Correa *et al.*, "Artifact removal from EEG signals using adaptive filters in cascade," *J. Physics: Conf. Series*, vol. 90, no. 012081, 2007.

[10] S.A. Hillyard, R. Galambos, "Eye movement artifact in the CNV," *Electroencephalography and Clinical Neurophysiology*, vol. 28, pp. 173-182, 1970.

[11] J.C. Woestenburg, M.N. Verbaten, J.L. Slangen, "The removal of the eye-movement artifact from the EEG by regression analysis in the frequency domain," *Biological Psychology*, vol. 16, pp. 127-147, 1983.

[12] S. Halder *et al.*, "Online artifact removal for brain-computer interfaces using suppor vector machines and blind source separation," *Computational Intelligence and Neuroscience*, no. 82069, 2007.

[13] P. LeVan, E. Urrestarazu, J. Gotman, "A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification," *Clinical Neurophysiology*, vol. 117, pp. 912-927, 2006.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[14] I. Winkler, S. Haufe, M. Tangermann, "Automatic classification of artifactual ICA-Components for artifact removal in EEG signals," *Behavioral and Brain Functions*, vol. 7, no. 30, 2011.

[15] N. Mammone, F.L. Foresta, and F.C. Morabito, "Automatic artifact rejection from multichannel scalp EEG by wavelet ICA," *IEEE Sensory J.*, vol. 12, no. 3, pp. 533-542, 2012.

[16] M.T. Akhtar, C.J. James, "Focal artifact removal from ongoing EEG- a hybrid approach based on spatially-constrained ica and wavelet de-noising," in *31st Annu. Int. Conf. IEEE EMBS*, Minneapolis, Minnesota, 2009, pp. 4027-4030.

[17] M.T. Akhtar, C.J. James, W. Mitsuhashi, "Modifying the spatially-constrained ICA for efficient removal of artifacts from EEG data," in *4th Int. Conf. Bioinformatics and Biomed. Eng.*, 2010, pp. 1-4.

[18] G. Inuso *et al.*, "Wavelet-ICA methodology for efficient artifact removal from Electroenecphalograhic recordings," in *Proc. Int. Joint Conf. Neural Networks*, Orlando, Florida, 2007, pp. 1524-1529.

[19] M. Iwasaki *et al.,* "Effects of eyelid closure, blinks, and eye movements on the electroencephalogram," *Clinical Neurophysiology*, vol. 116, pp. 878-885, 2005.

[20] I. Goncharova *et al.,* "EMG contamination of EEG: spectral and topographical characterisitics," *Clinical Neurophysiology*, vol. 114, pp. 1580-1593, 2003.

[21] M. Potter, N. Gadhok, W. Kinsner. "Separation performance of ICA on simulated EEG and ECG signals contaminated by noise," in *2002 IEEE Canadian Conf. Elect. and Comput. Eng.*, vol. 2, pp. 1099-1104.

[22] G. Katul, B. Vidakovic, "Identification of low-dimensional energy containing/flux transporting eddy motion in the atmospheric suraface layer using wavelet thresholding methods." *Discussion Papers of ISDS*, vol. 96, no. 23, pp. 1-24, 1998.

[23] N. Memarsadeghi *et al.*, "A Fast Implementation of the ISODATA Clustering Algorithm," *Int. J. Computational Geometry and Applicat.*, vol. 17, pp. 71-103, 2007.

[24] W. Liu *et al.*, "An adaptive clustering algorithm based on the possibility clustering and isodata for multispectral image classification." *Int. Archives of the Photogrammetry, Remote Sensing and Spatial Inform. Sci.*, vol. XXXVII, Part B7, pp. 565-568, 2008.

[25] D.G. Goring, V.I. Nikora, "Despiking acoustic doppler velocimeter data." *J. Hydraulic Eng.*, vol. 128, no.1, pp. 117-126, 2002.

[26] R.Q. Quiroga, Z. Nadasdy, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural Computation*, vol. 16, pp. 1661-1687, 2004.

[27] M.Z. Othman, M.M. Shaker, M.F. Abdullah, "EEG spikes detectin, sorting, and localization," *World Academy of Sci., Eng., and Technol.*, vol. 9, pp. 205-208, 2005.

[28] M. Mamun, M. Al-Kadi, M. Marufuzzaman, "Effectiveness of wavelet denoising on electroenecphalogram signals," *J. Applied Research and Technol.*, vol. 11, pp. 156-160, 2013.

[29] R.E. Herrera *et al.*, "Removal of non-white noise from single trial event related EEG signals usnig soft thresholding," in *IEEE Proc. 22nd Annu. EMBS Int. Conf.*, Chicago, IL, 2000, pp. 793-795.

[30] A. Delorme, S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[31] R. Karthik (2003). "ECG simulation using MATLAB," www. mathworks com/matlabcentral/fileexchange/10858.

[32] N. Yeung *et al.*, "Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods," *Psychophysiology*, vol. 41, pp. 822-832, 2004.

[33] M.C. Baker et al., "An SFFS technique for EEG feature classification to identify sub-groups," in 25th Int. Symp. Computer-Based Medical Syst., Rome, June 2012, pp.1-4.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

A Master-Slave MPI Approach for NGS Data Mining

R. Zanella
Mathematics and Computer Science Department, and
Laboratory for Technologies of Advanced Therapies (LTTA),
University of Ferrara, Ferrara, Italy

Abstract— *Next-Generation Sequencing (NGS) platforms can sequence a full human genome in previously unimagined speed. High-dimensional data produced must be filtered and annotated in order to be effective and of functional comprehension to end-users. Among public software packages, GAMES [1] aims to reduce the complexity in large scale DNA projects by filtering and annotation procedures. Developed in Perl language and available as open-source, the tool can be fruitfully adopted for small input data. We developed a C++ MPI version, based on Master/Slave pattern, in order to exploit the embarrassingly parallel nature of the NGS filtering problem, thus providing an efficient tool to the bioinformatic community.*

Keywords: MPI, NGS Sequencing, Master/Slave pattern

1. Introduction

Commercially available Next-Generation Sequencing (NGS) platforms have the ability to read exponentially more DNA base pairs per sequencing run than mature methods like Sanger sequencing. These techniques offer the possibility of examine large regions of the genome much more rapidly and with a less cost per Megabase of DNA Sequence [2].

As a result, NGS based methodologies are progressively introduced in clinical diagnosis, particularly for complex disorders that may be caused by a combination of several genes. NGS methods are becoming a standard approach in genomics: they are fruitfully applied in whole-genome sequencing, discovery of transcription factor binding sites or non-coding RNA expression profiles. NGS can improve clinical diagnostics, for example by detecting single nucleotide polymorphisms (SNP), insertions and deletions (InDels), and other rearrangements to identify disease-associated variants in clinical samples.

The advent of NGS platforms poses several problems in data analysis: in this paper we address the efficient filtering and annotation of the impressive amount of data a consolidated sequencing instrument can produce.

The project focuses on developing a parallel pipeline for the analysis and annotation of NGS biological data. Tools for the analysis of short/long DNA reads data sets have become available in the past few years. Among them, GAMES (Genomic Analysis of Mutations Extracted by Sequencing)[1], proposes a pipeline for high level mining of NGS results. Given the increasing capabilities of consolidated instruments (Illumina Genome Analyzer, ABI SOLiD, Ion Torrent, etc), the large amount of data available cannot be efficiently processed with the serial approach adopted in GAMES. A careful selection of parallelization approaches suitable for handling very large amount of sequence data has been carried out, leading to C++ GAMES, a parallel version of the GAMES pipeline, well suited to run on a standard cluster with MPI resources. On the computational side, the Master/Slave architectural pattern is adopted for acquiring a dynamic load balancing: a portion of DNA is sent by the Master to a working node, that collects the statistics and returns the results.

This paper is organised as follows. Section 2 combines a short description of the NGS data and SAM/BAM file format with the description of available tools for handling the contained information. Section 3 describes Master/Slave adopted approach. Section 4 compares the performances of the proposed tool with Perl GAMES implementation in terms of computational time and speedup achieved in three different test set. Section 5 presents conclusions.

2. Overview of NGS Data Access

DNA contains genetic instructions used in the development and functioning of all living organisms, coded as a sequence of four types of molecules called nucleotide bases, or bases, coded with G,A,T,C characters.

Sequencing architectures are able to record and collect nucleobases, grouping them in reads: these are a batch of consecutive bases. Their length is, for NGS architectures, $25-300$ bases: the collection of the extracted reads, aligned to a reference genome, forms the input data considered in this work. Thanks to the advance in the extraction technique, we can collect roughly 10^9 bases/day: one third of the whole genome length.

The extraction approach may result in discrepancies between overlapping DNA fragments: in this case the Phred quality score, an information generated in the sequencing phase, helps to figure out a guess of the base molecule present in the sample.

The alignment of each read to the reference genome aims to identify regions of the genome the read belongs to, by evaluating metrics that express the similarity between the sequence and the chosen genome reference. The read is aligned to the portion of genome where the chosen similarity metric is the higher.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

After the aligning step, each read is associated to a portion of genome, and the quality score of the alignment is retained. A poor score value may refer to an error on the extraction phase that degraded the information, or the probability of correct alignment is too low: for these reasons the read is usually not considered int the statistics.

Alignment algorithms may generate information during the process: gaps are inserted between the read bases so that identical or similar characters are aligned in successive columns. In the same manner, characters present/not present in the read, are flagged as insertion/deletions. Extracted reads and the related aligning information (insertions/deletions) are then stored in ASCII file, with SAM (Sequence Alignment/Map)[3] format, or more generally, in its binary compressed version, BAM (Binary Alignment/Map). The reference genome is stored in fasta format.

Many libraries offer function to manipulate fasta or SAM/BAM files. For this project we use the libbam library, available in the open source samtools project[4]. Unfortunately, these functions are not thread safe: this suggests the use of the library in a multiprocess environment, where more than a process can collect data from a portion of the input files, and prohibited a multithreading approach.

The adopted library, developed in ANSI C language, provide efficient utilities on manipulating alignments in the SAM format. In particular, a callback-based set of functions is present for parsing user-defined regions of the provided alignment.

In order to perform a statistic assessment, the programmer is required to provide two functions with a specified task, then the general region parsing algorithm is started by calling the function `bam_parse_region`.

In the following, we adopt a callback naming similar to the one present in on-line examples of samtools[4] and we briefly describe the implementation required for the GAMES filtering. For a biological based description of the full procedure, we refer the reader to GAMES paper[1].

2.1 Pileup Parsing

For each position in the region, the pileup parsing function `pileup_func` is called: it can perform a statistic evaluation of all the reads covering the position by accessing the data stored in a list, called read pileup. In this phase, position-based metrics are evaluated, such as consensus, Phred-like consensus quality, number of hits of reads covering the position, the second best call and the respective quality score and counts.

User defined parameters can be tailored to the particular input file and affect the filtering procedure. The user sets mapping quality, readsâĂŹ length (according to the experiment), minimal quality threshold, minimum coverage, minor allele frequency and maximal repetitivity. The parameters used to define the goodness of mutation detection are the minimum mapping quality, a measure of the confidence that a read is correctly aligned and the minimum Phred's base-specific quality score.

2.2 Read Fetching

The construction of the list parsed by `pileup_func` is partially performed by libbam: whenever the library starts to parse a position, all the reads covering that genomic coordinate and not alredy present in the pileup are filtered through the callback function `fetch_func`.

Users interested in reads with determined characteristics, (e.g. high alignment quality) can filter the reads at this level and decide whether a read under evaluation fulfils the requirements and can be stored in the read list.

We make note that a read is parsed through `fetch_func` only once, and, given the SAM/BAM file format requirements, the reads are filtered in genomic coordinate order. For this reason, at this level we filter out the reads by looking at read-wide metrics, such as alignment quality.

3. Master/Slave MPI approach

Naive parallelization of the pileup scanning process involves splitting of the input file in chunks (e.g one chunk for each chromosome), serial scanning, and reordering of the results following genomic coordinates.

The time spent by the filter algorithm to retrieve significant positions depends on the coverage: the number of reads that are aligned to the genome portion under analysis. As pointed out in Subsection 2.2, a read denoted by, for example, a low alignment quality, is discarded, it's per position-information is no longer taken into account during the pileup parsing phase, thus leveraging the total amount of work due for that chunk.

For this reason, the time spent on manipulating the information carried by a chunk is not easily calculated. While the coverage information can be obtained by a pre-scan of the input file, an in deep investigation aiming at equally subdividing the workload among peers would require a time comparable to the full filtering and investigation.

Our approach is based on the Master/Slave, or Manager/Workers[5] pattern, and offers:

- runtime splitting of the input data,
- fair load balancing,
- in-order collection of results.

Given N MPI nodes, the implementation pattern adopted involves the definition of a unique *Master* node, usually node acquiring MPI rank 0, and $N-1$ *Slaves*. In the following, we summarise the operations performed by Master and Slave nodes.

3.1 Master Node

Master node is responsible to define a job chunk, by setting start and end region coordinates, and a progressive id number. Moreover, it keeps track of the busy nodes (the

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

ALGORITHM 1 Master Node

Let $activesl = 0$, $totalsl$ the total number of Slaves,

1. Startup phase
   FOR each unprocessed job
   Send a job to free Slave
   $activesl \leftarrow activesl + 1$
   IF $activesl == totalsl$, THEN
   GOTO: 2. Main loop
   ENDIF
   ENDFOR

   IF $totalsl > activesl$, THEN
   Send cleanup message to idle Slaves
   GOTO: 3. Cleanup phase
   ENDIF
2. Main loop
   FOR each unprocessed job
   Wait for a result message
   Read the message from Slave
   Send a new job to Slave
   ENDFOR
3. Cleanup phase
   WHILE $activesl > 0$, THEN
   Wait for a result message
   Read the message from Slave
   Send cleanup message
   $activesl \leftarrow activesl - 1$
   ENDWHILE
4. Exit

Slaves already processing a job), and sends a new job to the idle ones.

As summarised in Algorithm 1, the workflow can be divided in three phases. During startup phase, Master node sends a different job to the available Slaves, then, in the main loop, it waits for a result message from any of the working nodes. After result retrieval, a new job is sent to the idle node, until all the available job are performed.

Finally, during cleanup phase, last results are received, and a termination job is sent to all Slaves. During both main loop and cleanup phases, this node performs a reordering of the received results and writes the output file containing the genome annotations. Master node reads genome information, located in the header section of fasta reference file, and divides the overall genome length in chunks, thus it does not need the information stored in SAM file.

3.2 Slave Node

For each received job, Slave node performs statistical selection of relevant positions on the assigned section of the genome, and sends the results to Master, until a cleanup message is received, as sketched in Algorithm 2.

ALGORITHM 2 Slave Node

1. Wait for a job
   IF cleanup, THEN
   GOTO step 2
   ELSE
   run job
   send results to Master
   GOTO step 1
   ENDIF
2. Exit

Slave node receives only start and end coordinates of a region, and make use of libbam functions described in Section 2 to access read information. We stress that, thanks to the callback-oriented implementation of libbam, the request of filtering a region not covered by any read results in not calling `pileup_func` and `fetch_func` at all. In this case, the required time to run the assigned job is minimized, and an empty result message is sent back to Master node.

Since Slave nodes require to read both BAM and fasta files, the processes must reside on a I/O capable nodes.

3.3 Data Splitting and Result Retrieval

The window length (i.e. the number of bases residing on a job chunk) is a user-provided parameter. We stress that this implementation detail can be easily substituted by a more complex schedule procedure, involving a runtime estimation of the window length.

Results are generally collected not in order, as sketched in Figure 1. In this example, results owning to job 1 and 2 are gathered first, then Master node receives results related to job 0. The result acquisition is aimed at minimizing memory resources: Master node writes the ordered statistics directly to output file, and keeps in memory only the ones belonging to jobs whose predecessors are not already accomplished.

4. Numerical Results

4.1 Small datasets

In this section we present the results of filtering three different input files: since both Perl and C++ GAMES obtain same output files, we focus only on the speedups of C++ GAMES in comparison to Perl GAMES implementation.

The results are obtained by running the codes on a twin CPU architecture Intel Xeon X5690, equipped with 188 GB of RAM; input files are stored on RAID 0 (Striping) partition, thus providing improved performance on disc access.

Table 1 shows the characteristics of the selected input tests: the coverage is far away from the whole human

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

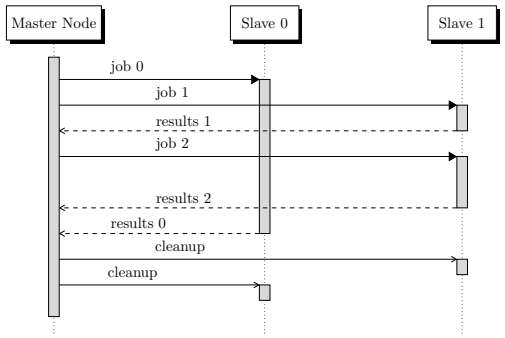


Fig. 1: Communication pattern example: the total work is divided into 3 jobs.

Table 1: Characterisation of the proposed tests: coverage (expressed in 10^6 bases), total number of reads present, file size.

Test name	coverage [mb]	aligned reads	file size [MB]
test01	10.0	432549	59
test02	11.7	904862	114
test03	49.5	2533304	328

genome length (approximately $3 * 10^9$ bases). For example, test03 represents a collection of reads whose coverage is $49.5 * 10^6$ bases, approximately $1/60$ of the bases in the human genome. Nevertheless, the use of Perl GAMES would require a considerable computational time: for the previously discussed test, Perl GAMES took roughly 15 hours, as stated in Table 2. The choice of the implementation language seems to play a not negligible role in the development of GAMES: C++ serial version took 85.5 minutes for the same input file.

As a first note, Perl GAMES does not implement the early discard feature exposed in Subsection 2.2. Moreover, the cause of this time improvement has to be addressed to the way Perl GAMES acquire data from SAM files: at low level, Perl exploits libbam features through the module Bio::DB:Sam [6], but considerable amount of time is spent in "translating" the information acquired by libbam (which is C code) to native Perl structures, suited to be handled to Perl GAMES implementation.

Not only the choice of C++ language implementation can impact to the overall performance, but also allow one to use MPI libraries for parallelization. Table 3 summarises the time spent for each input file, when varying the number of Slaves. The use of up to 23 Slaves reaches the maximum number of simultaneous threads (24) sustained by the available architecture, when Hyper-Threading Intel feature is chosen, and reaches the minimum of the required time for all the test files considered.

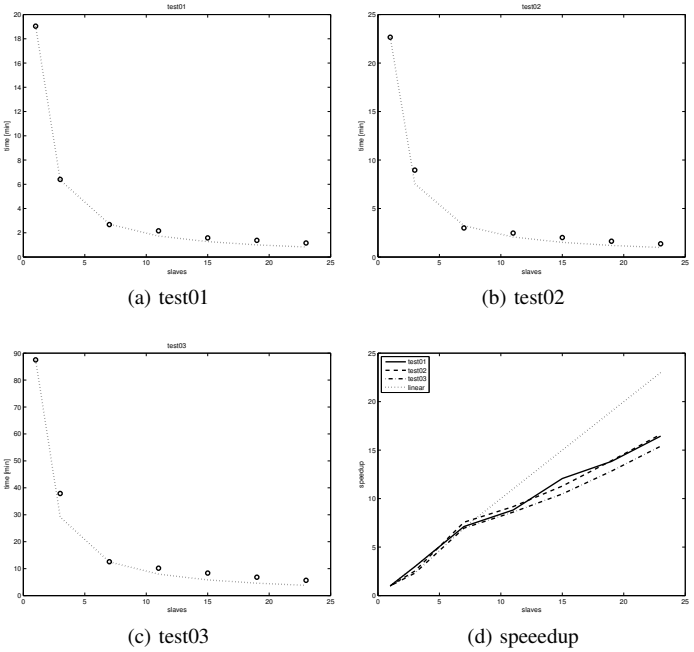
In Figure 2, charts a-b-c show the time required for each input file, denoted by a circle, compared to the expected time (dotted line), when varying the number of Slaves.

Table 2: Performance of the implementations, time expressed in minutes.

		C++ GAMES	
Test name	perl GAMES [min.]	serial [min.]	parallel (1 Slave) [min.]
test01	150.3	20.3	19.0
test02	339.8	23.8	22.7
test03	934.0	85.5	87.5

Table 3: Performance of the parallel C++ GAMES, time expressed in minutes.

	C++ parallel GAMES [min.]						
Test name	1 Sl.	3 Sl.	7 Sl.	11 Sl.	15 Sl.	19 Sl.	23 Sl.
test01	19.0	6.4	2.7	2.6	1.6	1.4	1.2
test02	22.7	9.0	3.0	2.5	2.0	1.6	1.4
test03	87.5	38.0	12.6	10.1	8.3	6.8	5.7



(a) test01 (b) test02

(c) test03 (d) speeedup

Fig. 2: Computational time required for the test files, when varying the number of Slaves, dotted line is the expected time. In bottom right figure, for each test is sketched the relative speedup, dotted line is the linear speedup.

As a consequence of the implementation pattern chosen, when running C++ GAMES with 1 Slave, only one node is active on the statistic assessment of the provided reads: for this reason we took this computational time as a reference. The dotted line is obtained by dividing the computational time required by running the code with one Slave, by the effective Slave nodes number.

The proposed implementation behaves well when comparing the expected time with the measured one in all the

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 4: Characterisation of the proposed tests: coverage (expressed in 10^6 bases), total number of reads present, file size.

Test name	coverage [mb]	aligned reads	file size [MB]
test01	1081	191976384	12153
test02	1078	182240964	11635
test03	1102	207441950	13156
test04	1087	190022968	12299
test05	1091	197392806	12468

considered test.

Panel (d) of Figure 2 sketches the relative speedup: this is obtained by dividing the time obtained with one Slaves by the computational time with s Slaves, dotted line is the ideal speedup.

In this case, the chart shows a relatively good speedup, since no saturation point is reached when varying the node number up to all the available ones. Moreover, the behaviour is not far from the ideal dotted line. However, we can note that speedup degrades around 7: this is due to disk bandwidth saturation: although BAM/SAM reside in a mirrored (2 disks) filesystem, this seems no longer fullfill data requests from more than 7 slaves (8 MPI processors). Same behavior can be observed in Table 3: for each line, a substantial decrease in running time is still found when swithing from 3 to 7 slaves. This is no more true from 7 to 11 slaves.

4.2 Large size datasets

Speedup results of Section 4.1 are related to small size datasets and are the result of DNA analysis performed on commodity equpment. In this subsection we present scalability results, when running MPI C++ GAMES on a large cluster: 274 Compute nodes, each one equipped by 2 esa-core Intel(R) Xeon(R) CPU E5645 @2.40GHz and 48 GB RAM per Compute node. Node interconnections are provided by a Qlogic QDR (40Gb/s) Infiniband high-performance network, high bandwidth network filesystem is available at each node.

4.2.1 Group 01 datasets

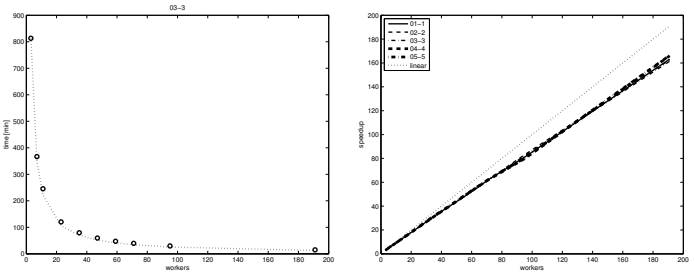
This dataset is used for testing the efficiency of the proposed approach, when very large input files are involved. As per table 4, each file of this group has a coverage which is 20-100 times higher than former tests, moreover also total number of reads and file size are increased.

4.2.2 Group 02 datasets

In this section we are intersted in a medium size coverage data, and check if also in this case we can achieve good scalability of our approach also with around 200 nodes.

5. Conclusions

In this paper we presented a Master/Slave parallel tool for the analysis of Next Generation Sequencing Data. The

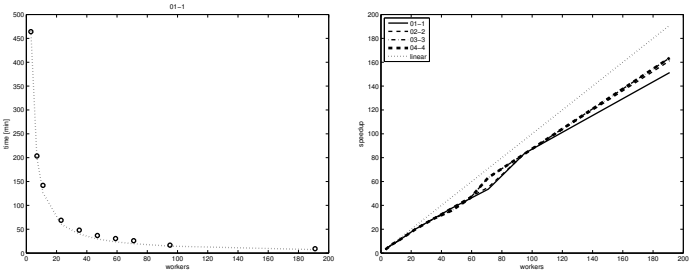


(a) test03 (b) speedup

Fig. 3: Panel (a) sketches the computational time required for test03, of big dataset, dotted line is the expected time. In Panel (b), for each test the relative speedup is sketched, dotted line is the linear speedup.

Table 5: Huge-size datasets, computational time expressed in minutes.

Slaves	Test01	Test02	Test03	Test04	Test05
4	806.7	806.1	813.2	807.7	819.0
8	363.7	356.8	366.8	367.6	361.2
12	242.0	240.1	245.5	242.2	244.5
24	117.7	118.3	120.4	117.7	117.7
36	77.2	77.5	79.4	77.7	77.5
48	58.6	58.3	59.5	58.6	59.0
60	47.2	46.7	47.4	46.7	47.0
72	39.0	38.9	39.6	39.1	39.4
96	29.9	29.1	29.7	30.4	30.6
192	14.8	14.9	15.1	14.6	14.8



(a) test01 (b) speedup

Fig. 4: Panel (a): computational time required test01 file, when varying the number of Slaves, dotted line is the expected time. In bottom right figure, for each test is sketched the relative speedup, dotted line is the linear speedup.

Table 6: Characterisation of the proposed tests: coverage (expressed in 10^6 bases), total number of reads present, file size.

Test name	coverage [mb]	aligned reads	file size [MB]
test01	641	72344195	4839
test02	781	89756630	6009
test03	654	79523593	5324
test04	654	79523593	8529

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Table 7: Medium-size datasets., computational time expressed in minutes

Slaves	Test01	Test02	Test03	Test04
4	464.1	562.1	469.4	469.4
8	203.5	250.0	198.5	198.3
12	142.0	169.0	142.0	142.0
24	68.8	82.7	70.0	70.0
36	48.1	57.4	48.8	48.8
48	36.9	44.1	39.3	39.3
60	30.5	35.9	30.6	30.5
72	26.0	30.6	22.3	22.3
96	16.7	20.2	17.0	17.0
192	9.2	10.5	8.6	8.6

proposed C++ MPI implementation shows a relatively good speedup when varying the number of nodes, and achieves a remarkable speedup compared to the original Perl GAMES.

References

[1] Sana, M.E., Iascone, M., Marchetti, D., Palatini, J., Galasso, M., Volinia, S.: Games identifies and annotates mutations in next-generation sequencing projects. Bioinformatics 27(1), 9–13 (2011), http://bioinformatics.oxfordjournals.org/content/27/1/9.abstract

[2] National Human Genome Research Institute: Dna sequencing costs, http://www.genome.gov/sequencingcosts/

[3] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, .G.P.D.P.: The sequence alignment/map (sam) format and samtools. Bioinformatics (2009)

[4] Samtools, http://samtools.sourceforge.net

[5] Ortega-Arjona, J.L., Roberts, G.: Architectural patterns for parallel programming. In: EuroPLoP. pp. 225–260 (1998)

[6] Stein, L.D.: Bio::db::sam, http://search.cpan.org/ lds/Bio-SamTools/lib/Bio/DB/Sam.pm

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Using Composite Digital Data to Improve the Interpretation of Analog Displays

Robert A. Warner, MD

Tigard Research Institute

Tigard, Oregon, USA

Abstract

Analog displays of data are used to detect a wide variety of conditions of interest, but the analog patterns associated with nominally similar conditions can be variable. Therefore, it is often difficult to develop optimal rules for identifying the conditions by reviewing only individual examples of them. In the present study, I describe a method for producing a "visual average" of analog tracings of similar conditions that addresses the problem of the heterogeneity of analog patterns. In addition, I describe a method to determine quickly and intuitively whether any portion of an analog tracing deviates statistically significantly from the baseline condition or from a normal standard. Applying the "visual averaging" method to electrocardiographic (ECG) data statistically significantly improved the detection of prior inferior myocardial infarction (IMI). Applying the statistical significance method to the same ECG data further improved the detection of prior IMI.

Keywords: digital data, analog display, patterns

1 Introduction

Analog displays are used to analyze data for many purposes. Analog displays of data help identify or predict important conditions like severe illnesses, meteorological changes and seismic events. A major reason why analog displays of data are so often useful is that humans are highly adept at recognizing both simple and complex patterns.[1] However, since nominally similar conditions of interest often have protean manifestations, the analog patterns that depict each case of the condition may vary widely. This heterogeneity may make it difficult to clearly identify patterns that are most likely to be associated with the condition of interest. In the case of analog patterns that seem to exclude the condition of interest, the heterogeneity of individual patterns may also prevent one from distinguishing between patterns that are merely atypical of the condition from those that do not reflect the condition at all.

Besides helping to identify conditions of interest, analog displays can also facilitate the understanding of the mechanisms of many phenomena. For example, the study of analog recordings of the heart's electrical activity greatly helped to elucidate the mechanisms responsible for both lethal and non-lethal cardiac arrhythmias. This raises the question of which general type of analog pattern most reliably reflects the nature of the condition to be investigated.

In the present study, I have described and tested two methods for optimizing the ability of analog displays of data to identify conditions of interest. One of the methods accounts mathematically for the variability of the patterns that individual cases of the conditions frequently exhibit.

2 Materials and Methods

2.1 Selection of patients

I studied the QRS complexes of the electrocardiograms (ECGs) of 862 patients who had undergone cardiac catheterization with left ventriculography and coronary angiography for suspected coronary artery disease. The QRS complex is the portion of the ECG that shows the changes in potential difference generated by the heart's cells during electrical depolarization of the ventricles.[2] Of these patients, the Normal Group consists of 497 (58%) with no evidence of coronary artery disease by cardiac catheterization. The IMI Group consists of the remaining 365 patients with evidence of prior inferior myocardial infarction (IMI) by cardiac catheterization. This evidence consists of either akinesia or dyskinesia of the inferior portion of the left ventricle shown in

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

the left anterior oblique projection accompanied by a ≥75% narrowing of either the right coronary artery or a dominant circumflex coronary artery.[3] The ECG of each patient had been acquired within 24 hours of the patient's cardiac catheterization.

2.2 ECG data from each patient

The digital ECG data used in the analyses had been downloaded from commercial ECG machines (GE/Marquette®) to a personal computer and imported into a spreadsheet. The digital ECG data from each patient had been stored on the ECG in sequential 4 ms. sampling intervals. To plot the ECG for each patient, I used the spreadsheet software to plot a line graph of these data for the entire duration of each patient's QRS complex. The linear plots of these data for each patient constitute the analog representations of the patient's ECGs and are illustrated in Figures 1 and 2. In each case, the QRS complexes analyzed were those recorded by standard ECG Lead aVF. Lead aVF was chosen because previous studies have demonstrated its usefulness for discriminating between normal patients and patients with IMI.[4]

2.3 Composite ECGs

Besides plotting line graphs for each of the patients, I constructed a normal composite ECG by calculating the mean voltages of all 497 patients in the Normal Group at each 4 ms. sampling interval and then plotting line graphs of these mean voltages sequentially (4 ms., 8 ms., 12ms., etc.) for the entire duration of the composite QRS complex. By using the mean voltages of the Normal Group at each sampling interval, the normal composite linear graph reveals the "average" pattern of the Lead aVF QRS complex for all 497 normal patients. I then repeated this process for the IMI Group at each 4 ms. interval of sampling to produce the "average" pattern of the Lead aVF QRS complex for all 365 IMI patients. Figure 3 shows the composite displays of the normal and of the IMI composite ECGs.

2.4 Z score plot

To further analyze any possible differences between the data of the Normal vs. the IMI groups at each sampling interval, I constructed a Z score plot of the IMI Group data by using the following calculated values[5]:

- The mean and standard deviation (SD) of the voltages for the 497 Normal Group patients at each 4 ms. sampling interval
- The mean of the voltages for the 365 IMI Group patients at each 4 ms. sampling interval
- The Z scores of the IMI Group data at each 4 ms. sampling using the formula**:**

$$Z = \frac{(\text{Mean of IMI Group} - \text{Mean of Normal Group})}{\text{SD of Normal Group}}$$

As the formula shows, if the mean of the IMI Group data at a given sampling interval exceeds the mean of the corresponding Normal Group data, the resulting Z score is positive. If the mean of IMI Group data is less than the mean of the corresponding Normal Group data, the resulting Z score is negative. The principal advantage of computing the Z scores is that the inclusion of the SDs in the calculations accounts for the variability of the data at each sampling interval. Consequently, each value of a Z score has a corresponding statistical P value. For example, for $Z \geq 1.65$ or $Z \leq -1.65$, the corresponding one-tailed P value is <0.05.[6] This relationship between Z scores and P values permits one to test the null hypothesis concerning apparent differences between the Normal and the IMI Group data at each 4 ms. interval of sampling.

Figure 4 shows the linear plot of the calculated Z scores of the IMI Group data at each interval of sampling.

2.5 Diagnostic performances

Using receiver-operating characteristic (ROC) curves, I measured the diagnostic sensitivities at 98% specificity for detecting prior IMI for each of the following sets of data:

- The digital ECG QRS complex data from all the 4 ms. sampling intervals
- The digital ECG data from only the 4-to 32 ms. sampling intervals. These are the sampling intervals at which the composite graphs in Figure 3 show directional divergence between the data of the Normal Group vs. the IMI Group. This divergence consists of the data of the Normal Group being positive while the corresponding data of the IMI group are negative.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

- The digital data from only the 20-to 28 ms. sampling intervals. These are the sampling intervals at which the Z score plot in Figure 4 shows statistically significant differences between the data of the Normal vs. the IMI Groups.

I used chi square analysis to test the null hypothesis regarding any apparent differences in diagnostic performance yielded by data of the 20 to 28 ms. sampling intervals vs. the data from all the sampling intervals and also vs. the data from the 4 to 32 ms. sampling intervals. For these comparisons, I a priori selected a P value <0.01 to avoid a Type 1 error associated with multiple comparisons.

3 Results

3.1 Individual ECGs

Figure 1 demonstrates the heterogeneity of the analog patterns exhibited by individual Lead aVF QRS complexes in the Normal Group. Each of the four recordings is from a different patient in that group. The individual recordings exhibit diverse patterns of positive and negative deflection of variable duration and amplitude in the initial, mid- and terminal portions of the tracings.

Figure 2 shows similar heterogeneity of the analog patterns exhibited by individual Lead aVF QRS complexes in the IMI Group.

Comparing the Lead aVF QRS patterns shown in Figures 1 and 2 suggests that it would be difficult to derive reliable rules for discriminating between the Normal and IMI Groups by examining only the analog displays obtained from individual patients.

3.2 Composite tracings

Figure 3 shows the respective analog patterns of the composite tracings of the Lead aVF QRS complexes of all the patients in the Normal Group and of all the patients in the IMI Group. In contrast to the individual patients' QRS complexes shown in Figures 1 and 2, the composite tracings in Figure 3 show visual synopses of the ECG QRS patterns that incorporate all the recorded data of the entire Normal Group and of the entire IMI Group. The composite tracings in Figure 3 reveal marked differences between the Normal vs. the IMI Groups. In the Normal Group composite, the QRS complex is almost entirely positive. However, the IMI Group composite has a broad initial negative component followed by a broad and relatively low amplitude terminal positive component. Specifically, sampling intervals 4 through 32 of the IMI composite are negative and

Figure 1
Examples of Individual Tracings
In the Normal Group

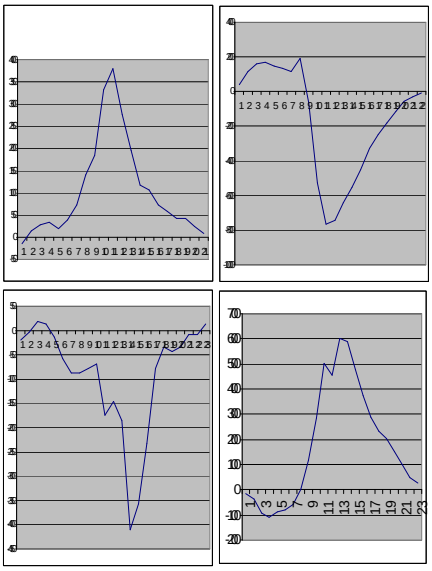
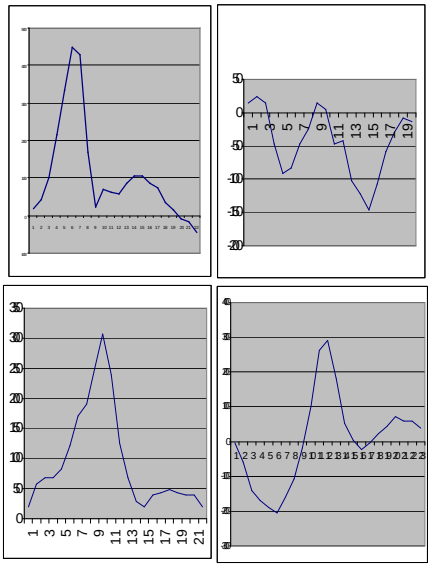


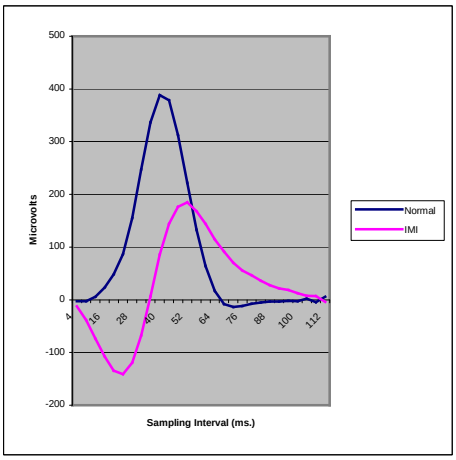
Figure 2
Examples of Individual Tracings
In the IMI Group



the corresponding sampling intervals of the WNL composite are positive.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Figure 3
Composite Lead aVF QRS Morphology in Normal vs. IMI Groups
Mean Voltages at Each Sampling Interval



3.3 Z Scores

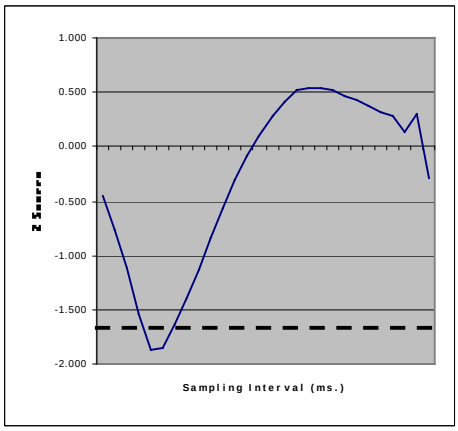
Figure 4 shows the Z score plot of the QRS data in Lead aVF for all the patients in the IMI Group. Similar to the linear display of the IMI Group's Lead aVF QRS voltage data shown in Figure 3, the linear display of the Z scores has a broad initial downward deflection followed by a relatively low amplitude upward deflection. The dashed horizontal line indicates the level below which the data of the IMI Group differ from the data of the Normal Group at the P<0.05 level of confidence. Only the 20 ms., 24 ms. and the 28-ms. samples are below this dashed line. Therefore, it is at only these three sampling intervals that the data of the IMI Group differ statistically significantly from the data of the Normal Group.

3.4 Diagnostic performances for detecting prior IMI

The ROC curves revealed that at 98% specificity, the respective diagnostic sensitivities for prior IMI are:

- 5% using the data from all the sampling intervals
- 34% using the data only from the 4 to 32 ms. sampling intervals as guided by the composite voltage plots in Figure 3
- 52% using the data from only the 20 to 28 ms. sampling intervals as guided by the Z score plot in Figure 4.

Figure 4
Z Score Plot of Lead aVF QRS in IMI Group



The diagnostic performance yielded by using only the 20 to 28 ms. sampling interval data was statistically significantly superior to that yielded by the data from all the sampling intervals (chi square = 199, $P<4x10^{-45}$) and also to that yielded by using only the 4 to 32 ms. data (chi square = 24.3, $P<9x10^{-7}$).

4 Discussion

The analysis of analog patterns is important for identifying and understanding conditions of interest in such diverse fields as engineering, medicine, statistics, economics and the social and physical sciences. Relevant to the present study, a cardiologist might be presented with arrays of digital data that can be used to generate a standard analog ECG waveform. However, it is likely that he will be able to interpret the analog waveform more efficiently than would be possible by examining only the digital data. This is largely because the human brain is highly adept at recognizing visual patterns.[7] This skill at pattern recognition is exemplified by our ability to quickly recognize a particular person's face even when that person is in a large crowd of other people.[8]

In studying analog displays of data, however, the viewer must have reliable answers to these questions:

1. "What types of patterns in the displays should I be looking for?"
2. "How much deviation from an expected pattern should be considered abnormal."
3. "Which, if any, portions of the analog display meet that criterion for abnormality."

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Together, the composite graphs and the Z score plots provide answers to all three of these questions.

In contrast, as shown by the heterogeneity of the patterns of the tracings in Figures 1 and 2, it is difficult to infer general rules for distinguishing between the Normal Group vs. the IMI Group by examining only individual examples from each of these groups. However, by using the mean voltages at each sampling interval, the composite tracings provide a more comprehensive picture of the patterns associated with the Normal vs. the IMI condition. The composite tracings provide visually "averaged" analog patterns that exhibit clear differences between the Normal and the IMI Groups.

By accurately revealing the general types of visual patterns associated with phenomena of interest, composite analog displays can also improve the scientific understanding of those phenomena. For example, statisticians often recommend that researchers generate histograms of data to reveal whether the data have gaussian or non-gaussian distributions.

The composite tracings shown in Figure 3 suggest that the initial portions of the QRS complexes in Lead aVF are more likely to discriminate between Normal and IMI than are the subsequent parts of the QRS complex. This is because the initial part of the Normal composite is positive and that of the IMI composite is negative.

The Z score plot in Figure 4 further refines this observation. Figure 4 demonstrates that it is specifically at the 20, 24 and 28 ms. intervals of sampling that the data of the Normal and the IMI Groups differ statistically significantly from each other. Using the ROC curves to compare the diagnostic performances for detecting IMI reveals the importance of the increased analytical precision provided by the Z score plots. Employing the data recorded from all the sampling intervals yielded a diagnostic sensitivity of 5.0%. Using only the data from the 4 ms. to the 32-ms. sampling intervals as guided by the composite tracings in Figure 3 increased the sensitivity to 34%. Using only the data from the 20 to the 28-ms. sampling intervals as guided by the Z score plot further increased the sensitivity to 52%. All these diagnostic sensitivities were obtained at 98% specificity. Compared to using all the data in the Normal and IMI Groups, the improvement in diagnostic performance as suggested by the composite tracings is highly significant ($P<9x10^{-7}$). The incremental improvement produced by using the Z score plot is even more significant ($P<4x10^{-45}$).

The above observations can be used by those charged with the task of diagnosing prior IMI with the ECG. The present study's findings suggest that individual patients whose QRS complexes in Lead aVF are abnormally negative at the 20, 24 or 28 ms. sampling intervals are especially likely to have had a prior IMI.

There are additional ways in which the calculation and graphing of Z score values can improve the analysis of analog displays. First, the Z score can clearly show which, if any, portions of an analog tracing are greater or less than values that represent a statistically significant difference from a previously chosen normal standard or set of baseline values. The horizontal dashed line in Figure 4 demonstrates this capability. A person who is reviewing a series of analog tracings could quickly determine which, if any, portions of each tracing intersect lines that represent various levels of statistical significance, e.g. $P<0.05$ ($Z\geq1.65$), $P<0.01$ ($Z\geq2.33$) or $P<0.001$ ($Z\geq3.08$). This process would enable the reviewer to quickly and accurately determine the presence and location of all statistically significantly abnormal portions of each analog tracing.

Second, using Z scores facilitates the analog display of multiple parameters. It is often desirable to display simultaneously graphs of parameters whose raw data are measured using units of different scales. Producing a meaningful graph of a parameter that is measured in large numerical units may "drown out" the simultaneous display of a parameter that is measured in smaller units. However, this difficulty is avoided by graphing the Z scores instead of the raw data. This is because the Z scores of the data of all parameters are expressed in the same units – the SD. Therefore, changes in all parameters that are represented by Z scores will be meaningfully displayed on the same scale in analog displays.

Third, in the present study, the Z scores expressed the statistical significance of portions of data representing an abnormal condition (previous IMI) compared to corresponding data from a normal condition (the absence of previous IMI). An alternative use of Z scores would be to compare the data collected during a period of moni-

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

toring to the data obtained during a baseline period. This would reveal any statistically significant changes from the baseline condition. An example of such an application would be stress testing for the detection of ischemic heart disease. Baseline ECG data can be collected from the patient during the initial period of rest. The means and SDs of relevant ECG parameters during this resting, baseline period would be calculated. These means and SDs would be combined with corresponding data obtained during the exercise period to calculate Z scores of the exercise data. The plots of these Z scores would show whether there were any statistically significant ECG changes that occurred during exercise. The use of such a statistically based technique would eliminate much of the subjectivity in the interpretation of time series of data.

5 Conclusions

- The variability among individual analog displays of nominally similar conditions of interest can impair the identification of these conditions.
- Composite analog displays of relevant sets of data constitute "visual averages" of the data.
- The composite displays help reveal which parts of individual analog displays discriminate best between cases and non-cases of the condition interest.
- Z score plots further improve the accurate detection of cases of the condition of interest.

6 References

1. Milewski, R, Govindaraju, V. Binarization and cleanup of handwritten text and carbon copy medical form images. Pattern Recognition 41 (4):1308-1315. 2008.
2. Warner RA, Hill NE, Mookherjee S, Smulyan H. Improved electrocardiographic criteria for the diagnosis of left anterior hemiblock. Am. J. Cardiol. 51:723-726, 1983.
3. Warner RA, Hill N. Sheehe P, Mookherjee S, Fruehan. Improved criteria for the diagnosis of inferior myocardial infarction. Circulation 66:422-428, 1982.
4. Warner RA, Hill NE. Optimized electrocardiographic criteria for prior inferior and an anterior myocardial infarction. J. Electrocardiol. 45:209-213, 2012.
5. Warner RA, Olicker AL, Haisty WK, Hill NE, Selvester RH Wagner GS. The importance of accounting for the variability of electrocardiographic data among diagnostically similar patients. Amer. J. Cardiol. 86:1238-1240, 2000.
6. Warner RA. Color-coded z scores for the display and analysis of biomedical data. 2nd World Congress on Biomarkers and Clinical Research, J Mol Biomark Diagn. 2(4):33, 2011.
7. Richard O. Duda, Peter E. Hart, David G. Stork (2001) Pattern classification (2nd edition), Wiley, New York
8. Nelson CA. (March–June 2001). The development and neural bases of face recognition. Infant and Child Development 10 (1–2): 3–18.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

BRATUMASS Source Separation: Multi-fractal Analysis for Near-field Microwave Signal Feature

A. Limin Xiao[1], B. Zhifu Tao[2], C. Yizhou Yao[3], D. Meng Yao[1]*,
E. Blair Fleet[4], F. Erik D. Goodman[4], G. Jinyao Yan[5], and H. John R. Deller[5]
[1]School of Info Sci and Tech, East China Normal University, Shanghai, China
[2]Dept of Elec Info Engineering, Suzhou Vocational University, Suzhou, China
[3]College of Science, Shenyang University of Technology, Shenyang, China
[4]BEACON Center, Michigan State University, East Lansing, MI, U.S.
[5]ECE, Michigan State University, East Lansing, MI, U.S.
*Corresponding Author, e-mail: myao@ee.ecnu.edu.cn

Abstract - *This paper aims at the data acquired by BRATUMASS (Breast tumor microwave sensor system), discussing from multi-fractal perspective to find the difference among different population fractal spectrum parameters, and then proposes a signal multi-fractal calculation model. In a statistical sense, the model calculated results show that the presence of breast tissue lesions would break the original breast tissue eigenvalue distribution rules, manifested as α_0 decreased; moreover, lesions will cause some data of the original rule deletion, manifested as spectrum width $\Delta\alpha$ narrowed distinctly. With respect to several pathological experiment data, the most serious degree is caused by malignant lesions (breast cancer), manifested as $\Delta\alpha$ and α_0 reaching a minimum.*

Keywords: Multi-fractal; BRATUMASS; Fractal spectrum; Breast cancer

1 Introduction

In recent years, with the development of fractal theory, it has been noted that tissue carcinogenesis can be described by fractal theory. Studies have shown that [1] tumor and cancer seemingly complex process system, is actually described by repeatedly calculating simple recursive formulas. The main difference between cancer cells and normal cells is, the former development almost stopped, and splitting speed is extraordinarily rapid, irregular, inconsistent, vague, clutter and chaos The cancer cells and its diffusion formed with bifurcation structure. The cancerous area is a chaotic region and also there is a similar singular attractor of carcinogenesis fractal element. People also know from analysis on breast cancer tissue cell nucleus forma that breast cancer and breast fibroadenoma are with fractal characteristics, fractal dimension quantitatively describes the irregular extent of tumor cell nucleus shape. It is of significance for the identification of benign and malignant tumor on pathology, and can be used as a reference index in the differential diagnosis of breast cancer [2]. In generally, the higher degree of malignancy is the worse cells differentiation [3]. On morphologically It can be roughly identify whether the lump is malignant or not[4], malignant tissues such as "Coral shape", its boundary is not clear, there are strands of infiltrating between tumor and gland; for the benign such as "cobblestone shape", its boundary clear and regular.

Breast tumor microwave sensor system (BRATUMASS) is a breast tumor data acquisition system which is using the difference of dielectric constant and conductivity parameters between malignant tumor tissue and normal breast tissue, to localize the breast tumor by detected target dielectric properties [5]. On data processing, According to the targets distance distribution, transforming the sampling signal in time domain to the corresponding whole detection space electromagnetic property distribution, which is in order to achieve the malignant tissues detection. However, from the aspect of microwave signal feature, how to use the intrinsic feature to detect the target and find proper signal parameters to remark the characterization of malignant breast lesions, is a work of practical significance, and is also a challenge [6]. This paper is from the perspective of application of multifractal signals processing, based on BRATUMASS system experimental data, discussing the corresponding meaning of fractal parameters, putting forward with the multifractal calculating model steps, through analysing the different medical clinical cases experiment parameters, resulting in the difference among different breast tissues lesions in fractal spectrum parameter.

2 Multifractal spectrum

2.1 Model

The multifractal theory is applied for BRATUMASS system signal analysis in which we should transform time domain sampling signal data to the corresponding detection space electromagnetic property distribution(spectral sequence

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[x_i])[5], which are a function of the target distance. The multifractal spectrum [7][8] calculation steps are as follows:

①Normalize spectral sequence xi, and represented by

$$p_i = \frac{x_i}{\sum x_i} \quad (1)$$

And then, divide the normalized spectrum sequence to non-overlapped segmentation window, which range interval is L.

② Calculate each window (or box) spectrum peak probability $p_j(L)$, the spectrum peak probability equal to the sum of all normalized segmentation window of spectral sequences.

③Select the appropriate value q, through $p_j(L)$ calculating the q partition function

$$M_q = \sum_{j=1}^{n} p_j^q(L) \quad (2)$$

Where, n is the total window number of time and space distance which is equal to L; q is a real number which range from $-\infty$ to $+\infty$. Considering multifractal distribution, partition function with window length obeys the following scale relations [9][10]:

$$M_q \propto L^{t(q)} \quad (3)$$

④ Drawing the corresponding $\ln(L) \sim \ln(M_q(L))$ curve according to equation (3) , if there is a good linear relationship of $\ln(M_q(L))$ with the change $\ln(L)$, then the distribution obeys multifractal distribution [11][12]. The slope of the $\ln(L) \sim \ln(M_q(L))$ curve is $\tau(q)$, from which can calculate the multifractal spectrum[13][14]. The calculation formula is as follows:

$$\frac{d}{dq}[t(q)] = a(q) \quad (4)$$

$$t(q) = qa(q) - f(a) \quad (5)$$

Figure 1 is an example of the preprocessed BRATUMASS system signal $\ln(L) \sim \ln(M_q(L))$ diagram. We can see that signal preprocessed by BRATUMASS signals obviously comply with the multifractal rule.

2.2 The significance of model parameters

α describes the singular degree of spectral sequence in each interval, fractal spectrum width $\Delta\alpha = \alpha_{max} - \alpha_{min}$ characterizes the difference between minimum and maximum probability, and it shows the distribution uniformity of normalized feature parameters of the fractal structure[14][15][16]. We apply it to the preprocessing sequence in BRATUMASS system, then α reflects the characteristic value, α is smaller, the eigenvalue is higher, Vice versa. Here α_{min} represents the maximum eigenvalue (exponent), α_{max} represents the minimum eigenvalue. $\Delta\alpha$ reflects the fluctuation extent of eigenvalue. Greater $\Delta\alpha$ represents bigger change of regional eigenvalue; $\Delta\alpha = 0$ corresponds to completely uniform distribution. $f(\alpha)$ reflects the frequency of occurrence eigenvalue which is α corresponded. The larger of $f(\alpha)$, the more frequency of occurrence α corresponded. $f(\alpha_{min})$ and $f(\alpha_{max})$ correspond to the subset of the fractal dimension of α_{min} and α_{max}, respectively[16][17][18]. The maximum of $f(\alpha)$ is the peak of multifractal spectrum, corresponding to the peak α recorded as α_0, the mean of all the individual α_0 in same categories of clinical cases group recorded as $\overline{a_0}$; characteristic parameter $\Delta\alpha$ can be divided into $\Delta a_L = a_0 - a_{min}$ and $\Delta a_H = a_{max} - a_0$ represent the left and right side of the multi fractal spectrum range of α, respectively [18]

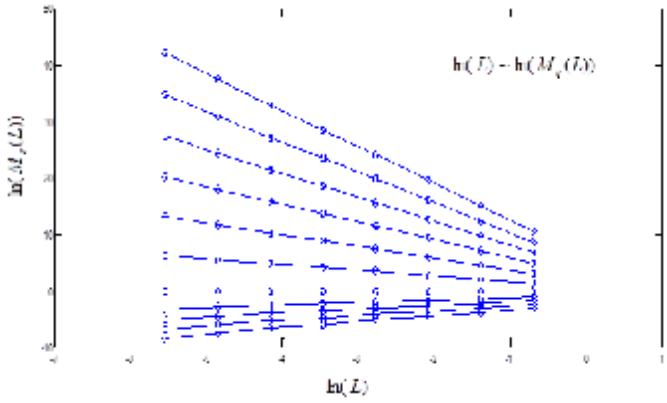


Figure1 an example of signal $\ln(L) - \ln(M_q(L))$ diagram

3 Statistics and analysis of experimental signals

In the present experiment, there were 13 cases (definite diagnosis after surgical operation) effective breast cancer patients' data, 22 fibroma lesions data, 13 other breast disease data and 8 normal cases' data. Figure 2 is the diagram of $f(\alpha) \sim \alpha$ multifractal spectrum calculated by above multifractal analysis method. At least, it can be seen from Figure 2 these several note points:

1) All the $f(\alpha) \sim \alpha$ of obtained data is with a single peak, which is an important feature of multifractal signal [19][20].

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2)13 cases of breast cancer data α distribution are between 0.4-1.6, but breast disease, such as fibroma, etc. and healthy people appeared over the spectral distribution of 0.4-1.6. Furthermore, we consider Δα distribution: breast cancer spectral width Δα < fibroma data spectrum width Δα < General breast disease spectrum width Δα< normal population breast data spectrum width Δα.

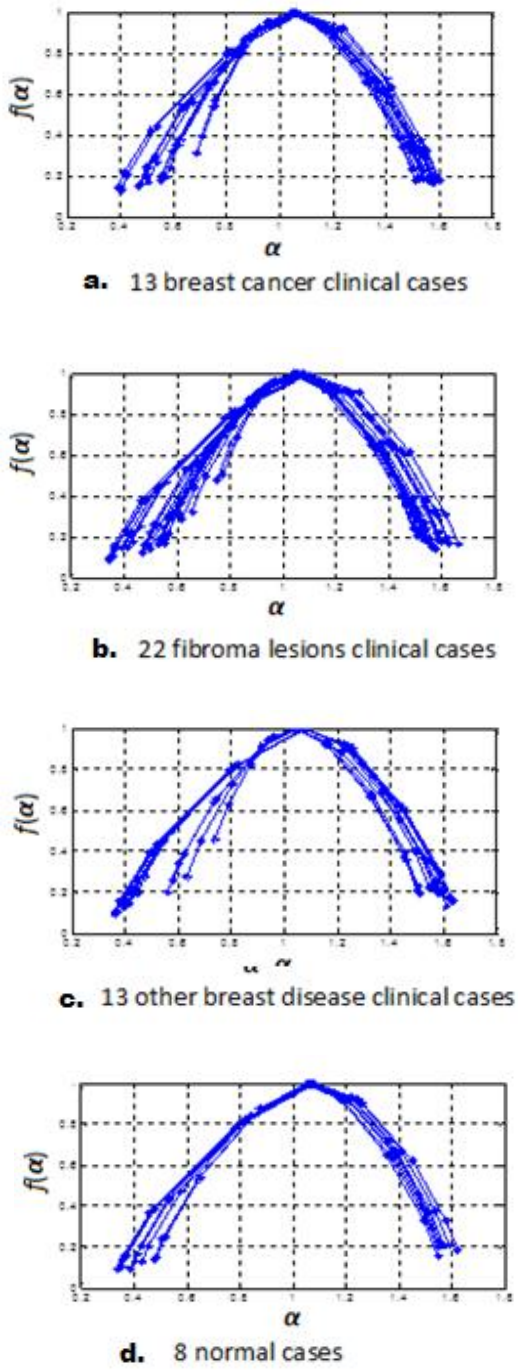


Figure 2 multifractal analysis results

3) The variation of the mean value of multifractal spectrum peak α_0 is very obvious, the breast cancer data spectrum peak point $\overline{a_0}$ < fibroma data spectrum peak point $\overline{a_0}$ < general breast disease $\overline{a_0}$ < normal population data spectrum peak point $\overline{a_0}$. Table 1 is the result of multifractal spectrum peak mean value $\overline{a_0}$ and spectrum width Δα. The difference α_0 between different groups within individuals is shown in Figure 3, different individuals value of α_0. The population number 1 is the data in breast cancer patients; 2 is fibroma population data; 3 is a group of data of patients with breast disease; 4 is a normal population data. As can be seen in the same population, there is little difference between the α_0 value.

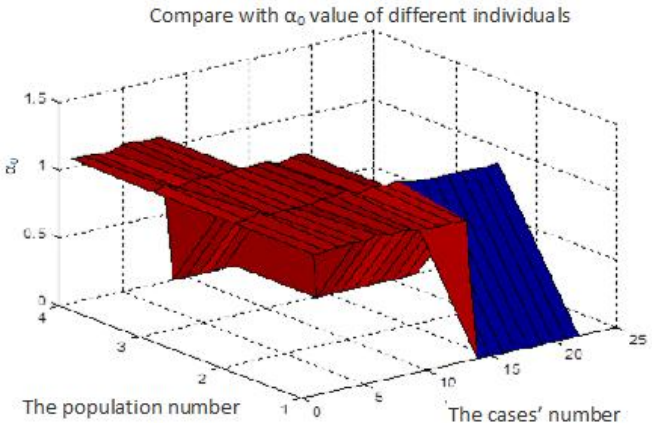


Figure 3 α_0 value of different individuals. The population number 1 is breast cancer population data, in which effective data is 13; 2 is fibroma population data, in which effective data is 22; 3 is mastopathy population data, in which effective data is 13; 4 is the normal population data, in which effective data is 8

Table 1. multifractal spectrum peak average $\overline{a_0}$ of received data

Population type	$\overline{a_0}$	Δα
breast cancer	1.0557527460646	1.0224925818808
fibroma	1.05906415779319	1.04652224615539
mastopathy	1.05992643184810	1.10510388810413
normal	1.06167884518363	1.15938913828987

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

We can conclude from the meaning of $f(\alpha) \sim \alpha$ spectrum that the experimental data is of multifractal properties obviously, and in the α spectrum distribution interval view, the distribution of breast cancer interval significantly reduced (limited between 0.4-1.6), which indicates the eigenvalue corresponding tissue species distribution decreased significantly, further clarify the differentiation degree decreased obviously. From health population to the breast cancer population, the α_0 value decreased, further explained the breast cancer causing local tissue differentiation weakened. As a result, the differentiation decreased degree of breast cancer is stronger, the corresponding $\overline{a_0}$ is smallest. Fibroma is smaller, mastopathy is small. The value of different populations of multifractal spectrum contrasts as shown in figure 4.

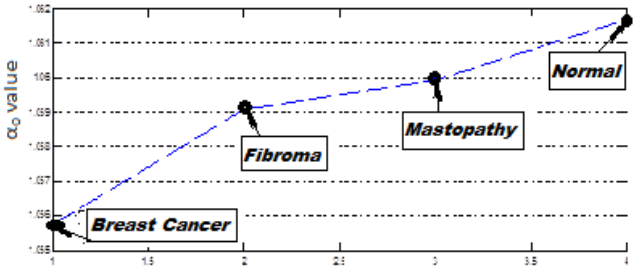


Figure 4 Comparison of different populations of multifractal $\overline{a_0}$ spectrum value

4 Conclusions

We can conclude from the current analyses of existing data results that the microwave back scattering signal obtained by BRATUMASS through preprocessing is with obvious multifractal characteristics. The multifractal rule of BRATUMASS signal actually corresponds to the actual breast tissue eigenvalue distribution, from a certain aspect, it has fractal characteristics. It can be drawn from the above results that the normal cases breast tissue distribution corresponded a spectral width and peak value is large, and lesions tissue cause the width and peak of multifractal spectrum decreased, it lead to the change of tissue eigenvalue distribution , above all , this change to maximum which is caused by malignancy lesions. We can indicate that normal breast tissue has intrinsic regularities of distribution in its fractal analysis space, and the above research shows that regular pattern are broken by lesions tissue (corresponding to α0 reduced), which lead to some part of regular pattern lost (corresponding to spectral width Δα smaller), above all, the malignant lesions caused loss is the most serious (Δα and α0 reach minimum). The study of fractal characteristics of the BRATUMASS signal, namely, which is corresponds to the pathological changes of local tissue differentiation degree of change. And it is also greatly important for cancer detection and classification.

5 Acknowledge

This work has been performed while Prof. Meng Yao was a visiting Professor in Beacon Center, 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 References

Number in square brackets (“[ ]”) should cite references to the literature in the main text. List the cited references in numerical order at the very end of your paper (under the heading `References'). Start each referenced paper on a new line (by its number in square brackets).

[1] Lin Hongyi, Li Yingxue “The fractal theory—singularity explore”. Beijing institute of technology press, 237-238, Sep 1992.

[2] Zhoghua Zhang, Yuan Ding Xu “Fractal analysis in the form of the nucleus of breast cancer”. Cancer, 63, Feb 1996. Ree Source Person. “Title of Research Paper”; name of journal (name of publisher of the journal), Vol. No., Issue No., Page numbers (eg.728—736), Month, and Year of publication (eg. Oct 2006).

[3] Gongxin Zhang etc. “Breast tumor cell nucleus in the form of fractal analysis” Journal of Zhengzhou University (Medical Sciences) Vol. 37 No. 1-433-04 Jul. 2002

[4] Hongtiao Yu, etc “Map of diagnosis and treatment of breast tumor” Henan science and technology press, 6-9, May 1996.

[5] Tao Zhi-fu,etc. “Biopsy Back Wave Preprocessing Rese2arch of BRATUMASS System based on Applications of Fractional Fourier Transform” Proceedings of The 2010 International Conference on Bioinformatics & Computational Biology Volume II , July 2010.

[6] Natalia K. Nikolova, “Microwave imaging for breast cancer”, IEEE microwave magazine. 78-94, Dec 2011.

[7] Ming Li, “Fractal Time Series — A Tutorial Review, Mathematical Problems in Engineering” Volume 2010, Article ID 157264, 26 pages, 2010.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[8] Ming Li, Massimo Scalia, and Cristian Toma, "Non-Linear Time Series: Computations and Applications, Mathematical Problems in Engineering" Volume 2010, 2010.

[9] K. Falconer, "Fractal Geometry: Mathematical Foundations and Applications" John Wiley and Sons, NewYork, 1990.

[10] Ming Li, S. C. Lim, and Huamin Feng, "A novel description of multifractal phenomenon of network traffic based on generalized Cauchy process", ICCS 2007, Eds., Shi, van Albada, Sloot, and Dongarra, Springer LNCS 4489, 1-9, May 2007.

[11] Kent, J. T., Wood, T. A, "Estimating the Fractal Dimension of a Locally Self-Similar Gaussian Process by Using Increments" J. R. Statit. Soc. B 59, 579-599, 1997.

[12] T.Tel.Fractals,"Multifractals and Thermodynamics". Zeitschrift der Natur-forschung A,43:1154–1174,1988

[13] PENTLAND A P. "Fractal based description of natural scenes" [J].IEEE PAM, 661-674, Jan 1984.

[14] Kantelhardt, Jan W, Zschiegner, Stephan A,-Koscielny-Bunde, Eva, Havlin, Shlomo, Bunde, Armin,Stanley, H.Eugene, "Multifractal detrended fluctuation analysis of nonstationary time series", Physica A: Statistical Mechanics and its Applications, 87, Dec 2002.

[15] C. J. G. Evertsz and B. B. Mandelbrot. "Multifractal measures". In H. O. Peitgen, H. Jürgens, and D. Saupe, editors, Chaos and Fractals, pages: 849–881. Springer-Verlag, New York, 1992

[16] Davies, S., Hall, P "Fractal Analysis of Surface Roughness by Using Spatial Data" Journ of the Royal Statistical Society Series B 61, 3-37, 1999.

[17] Hall, P., Roy, R "On the Relationship between Fractal Dimension and Fractal Index for Stationary Stochastic Processes" The Annals of Applied Probability 4, 241-253, 1994.

[18] P. Mannersalo, I.Norros, and R.Riedi. "Multifractal products of stochastic processes". COST257, 31, 1999.

[19] P.Gon¸calv`es."Existence test of moments: Application to Multifractal Analysis" In Proceedings of the International Conferenceon Telecommunications, Acapulco(Mexico), May 2000.

[20] P.Gon¸calv`es."Existence test of moments: Application to Multifractal Analysis" In Proceedings of the International Conferenceon Telecommunications, Acapulco(Mexico), May 2000.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

OPTIMAL SCALING VALUES FOR TIME-FREQUENCY DISTRIBUTIONS IN DOPPLER ULTRASOUND BLOOD FLOW MEASUREMENT

F. García Nocetti, J. Solano González, E. Rubio Acosta

Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Circuito Escolar S/N, Ciudad Universitaria, México D. F., 04510, México
Contacting email: fabian.garcia@iimas.unam.mx

Abstract: Time-frequency distributions (TFD) are an alternative for signal analysis associated to Doppler ultrasound blood flow measurement, since these do not suppose that the signal is stationary. TFD have a scaling factor with which an optimization problem to get spectral estimations can be proposed. The optimization problem can be solved analytically or experimentally considering the characteristics of the studied signals that correspond to three simulated Doppler ultrasound quasi-stationary signals represent a typical blood flow in the Carotid, Coronary and Femoral arteries. Modified-B, Choi Williams, Born Jordan and Bessel distributions are considered. In this work the optimal scaling factor values, the so-called optimal parameters have been determined experimentally for different conditions of SNR and window length to estimate pseudo instantaneous mean frequency (PIMF) and RMS bandwidth (RMSB). Modified-B distribution produces the best PIMF and RMSB spectral estimations.

Keywords: Time-Frequency Distributions, Signal Analysis, Doppler ultrasound blood flow.

1.- INTRODUCTION

The analysis of a signal $x(t)$ using the Fourier transform magnitude reveals which frequency components are present but it does not locate them temporarily. If a signal is multiplied by a sampling window $W(t)$, with support $-T/2 < t < T/2$, then the temporal location of the frequency components capability is artificially introduced. In this way, the spectrogram $S(t,\omega)$ is obtained:

$$S(t,\omega) = \left| \int_{-\infty}^{\infty} W(\tau - t) x(\tau) e^{-j\omega\tau} d\tau \right|^2 \qquad (1)$$

Nevertheless, as the support of $W(t)$ diminishes, the temporal resolution increases but the frequency resolution diminishes. In the opposite case, as the support of $W(t)$ increases, the temporal resolution diminishes but the frequency resolution increases. Further, the use of the Fourier transform supposes that the signal is stationary. For the case of the signals associated to Doppler ultrasound blood flow measurement, this supposition is fulfilled as the support of $W(t)$ diminishes, sacrificing the frequency resolution.

Time frequency distributions (TFD) (Cohen, 1989) are an alternative for signal analysis associated to Doppler ultrasound blood flow measurement, since they do not suppose that the signal is stationary. Further, the temporal location of the frequency components is done in an intrinsic way and they do not suffer from the commitment between the temporal and frequency resolutions explained previously.

The Time Frequency Distributions have a scaling factor with which an optimization problem to get spectral estimations can be proposed. Two spectral estimations are considered: pseudo instantaneous mean frequency and RMS bandwidth. This optimization problem can be solved analytically (Boashash, and Sucic, 2003) or experimentally considering characteristics of the studied signals. The considered signals are three simulated Doppler ultrasound quasi-stationary signals that represent a typical blood flow in the Carotid, Coronary and Femoral arteries. Modified-B, Choi Williams, Born Jordan and Bessel distributions are considered.

Previous works have suggested optimal scaling factor values experimentally calculated (García, *et. al*, 2002 b; Cardoso, *et al.*, 1996).

2. TIME FREQUENCY DISTRIBUTIONS

The time frequency distributions (TFD) of the Cohen class considered in this work are the Bessel, the Born Jordan, the Choi Williams and the Modified-B distributions.

The discrete TFD of a complex signal *x(n)* of length *L=2N-1*, whose elements are numbered from *1-N* to

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

N-1, when it is evaluated at discrete time *n=0,* and optimized (Boashash, and Black, 1987) is:

$$DTFD(0,k) = 4\,\mathrm{Re}\left[\sum_{\tau=0}^{N-1} W(\tau)W^*(-\tau)f_{GAF}(0,\tau)e^{-j\frac{2\pi k\tau}{N}}\right] - 2W(0)W^*(0)f_{GAF}(0,0) \quad (2)$$

where $f_{GAF}(n,\tau)$ is the generalized autocorrelation function, *W(n)* is a (Hanning) sampling window of length *L=2N-1*, and *k* is the discrete frequency taking integer values from *0* to *N-1*.

The $f_{GAF}(\bullet)$ for the discrete Bessel TFD (Guo, and Durand, 1994) is:

$$f_{GAF}(0,\tau) = \sum_{\mu=\max\{-2\alpha|\tau|,-N+1+|\tau|\}}^{\min\{2\alpha|\tau|,N-1-|\tau|\}} \left(\frac{1}{\pi\alpha|\tau|}\sqrt{1-\left(\frac{\mu}{2\alpha\tau}\right)^2}\right) x(\mu+\tau)x^*(\mu-\tau) \quad (3)$$

where α is a scaling factor taking the half of any natural value. Note that $f_{GAF}(0,0) = x(0)x^*(0)$.

The $f_{GAF}(\bullet)$ for the discrete Born Jordan TFD (Cohen, 1989) is:

$$f_{GAF}(0,\tau) = \sum_{\mu=\max\{-2\alpha|\tau|,-N+1+|\tau|\}}^{\min\{2\alpha|\tau|,N-1-|\tau|\}} \left(\frac{1}{4\alpha|\tau|}\right) x(\mu+\tau)x^*(\mu-\tau) \quad (4)$$

where α is a scaling factor taking the half of any natural value. Note that $f_{GAF}(0,0) = x(0)x^*(0)$.

The $f_{GAF}(\bullet)$ for the discrete Choi-Williams TFD (Choi, and Williams, 1989) is:

$$f_{GAF}(0,\tau) = \sum_{\mu=-N+1+|\tau|}^{N-1-|\tau|} \left(\sqrt{\frac{1}{4\pi\tau^2/\sigma}}e^{-\frac{\mu^2}{4\tau^2/\sigma}}\right) x(\mu+\tau)x^*(\mu-\tau) \quad (5)$$

where σ is a scaling factor taking any positive real value. Note that $f_{GAF}(0,0) = x(0)x^*(0)$.

The $f_{GAF}(\bullet)$ for the discrete Modified-B TFD (Hussain, and Boashash, 2002) (Boashash, et al, 2013) is:

$$f_{GAF}(0,\tau) = \sum_{\mu=-N+1+|\tau|}^{N-1-|\tau|} \frac{\Gamma(2\alpha)}{2^{2\alpha-1}\Gamma^2(\alpha)}\left(\frac{1}{\cosh^2(\mu)}\right)^{\alpha} x(\mu+\tau)x^*(\mu-\tau) \quad (6)$$

where α is a scaling factor taking any positive real value.

3.- OPTIMAL PARAMETER

An optimal parameter is that time frequency distribution scaling factor value that minimizes the RMS error of a spectral estimation. The value of the optimal parameter depends on the spectral estimation that is realized. Also it depends on the length of the sampling window (L) and noise to signal ratio (SNR). In this work it is considered the spectral estimations of the pseudo instantaneous mean frequency (PIMF) and RMS bandwidth (RMSB). In consequence, optimal parameters are calculated for each of them.

In this work an experimentally determination of optimal parameters is accomplished.

4. DOPPLER ULTRASOUND SIGNAL SIMULATION

In order to characterize the pseudo instantaneous mean frequency (PIMF) and the RMS bandwidth (RMSB) error estimations when the TFD are used, it has been proposed the utilization of three simulated Doppler ultrasound quasi-stationary signals that represent a typical blood flow in the Carotid, Coronary and Femoral arteries. Their characteristics are well documented (Evans, 2000) (De Lazzari, *et al*., 2006).

Briefly, the signals' duration is 30 cardiac cycles at 61.6 ppm; they have a constant RMSB of *100Hz* and their PIMF wave form are shown in figures 1a to 1c. The simulation procedure is accurate described in (Cardoso, *et al.*, 1996). In this work, a sampling rate f_o*=12800Hz* is considered. Note that the sampling rate must be four times the signal's maximum frequency when TFD are used.

A white noise is added to the whole signal before starting the signal analysis procedure, according to typically prescribed signal noise ratios (SNR). In this work, SNR of -30dB and noiseless cases are considered (the minus sign will be omitted).

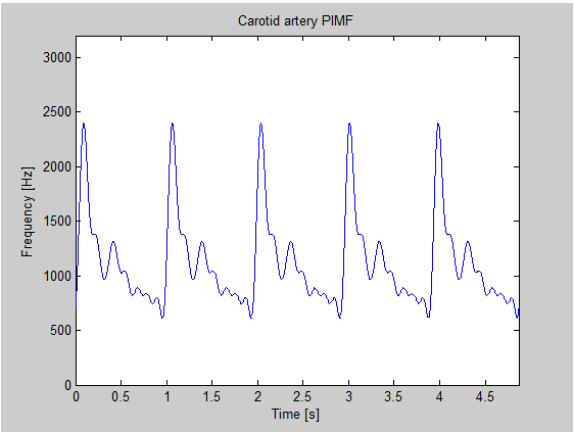


Fig 1a: Signal's pseudo instantaneous mean frequency (PIMF) wave form of the simulated Doppler ultrasound quasi-stationary signal that represents a typical blood flow in the Carotid artery.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

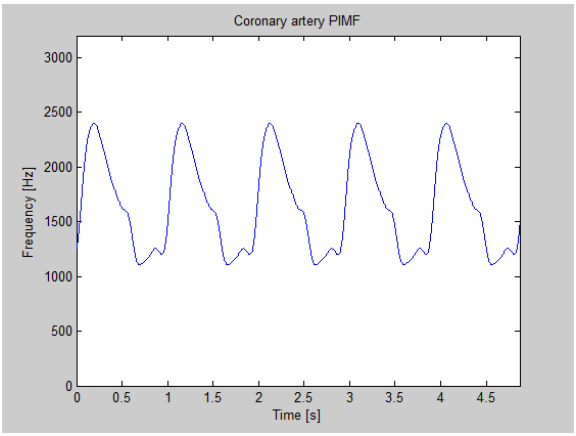


Fig 1b: Signal's pseudo instantaneous mean frequency (PIMF) wave form of the simulated Doppler ultrasound quasi-stationary signal that represents a typical blood flow in the Coronary artery.

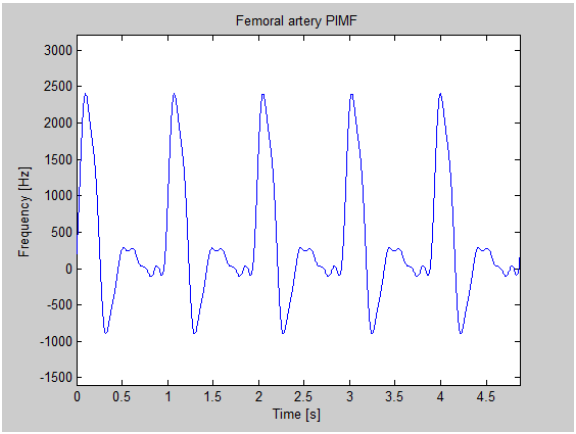


Fig 1c: Signal's pseudo instantaneous mean frequency (PIMF) wave form of the simulated Doppler ultrasound quasi-stationary signal that represents a typical blood flow in the Femoral artery.

5. SPECTRAL ESTIMATION

The spectral estimation of both the RMSB and the PIMF is worked out as in (Cardoso, *et al.*, 1996; Fan, and Evans, 1994). Their procedures have a common part. First, a signal piece of length L is taken from the n^{th} to the $(n+L-1)^{th}$ elements of the whole signal, it will be called the n^{th} signal window. In this work, L can be *127, 255, 511* and *1023*, and $L=2N-1$. The signal window's elements are numbered in the discrete time domain from $1-N$ to $N-1$. The quadrature signal's elements are also numbered in the discrete time domain from $1-N$ to $N-1$. Second, the TFD of this quadrature signal is calculated using equation (2) and (3), (4), (5) or (6) depending on the study case, considering prescribed scaling factors. The TFD's elements are numbered in the discrete frequency domain from $-N/2$ to $N/2-1$.

Finally, the pseudo instantaneous power distribution (PIPD) of this TFD is calculated. Its elements are also numbered in the discrete frequency domain from $-N/2$ to $N/2-1$. The PIPD is defined as:

$$PIPD(0,k) = \begin{cases} TFD(0,k) & TFD(0,k) \geq 0 \\ 0 & TFD(0,k) < 0 \end{cases} \quad (7)$$

In case of the PIMF calculation, the pseudo instantaneous mean frequency associated to the n^{th} window signal is stated by:

$$PIMF(n) = \frac{\sum_{k=-N/2}^{N/2-1} f_k \bullet PIPD(0,k)}{\sum_{k=-N/2}^{N/2-1} PIPD(0,k)} \quad (8)$$

where f_k is the real frequency associated to discrete frequency k. Observe that n can be considered as the whole signal's discrete time variable, running from *0* to $T-L$. Indeed, it represents the total amount of fully overlapped signal windows of length L in the whole signal (an overlapping of $L-1$ elements). That is, the PIMF(1) correspond to the 1st signal window; the PIMF(2), to the 2nd signal window; and so on. On the other hand, in case of the RMSB calculation, the RMS bandwidth associated to the n^{th} window signal is stated by:

$$RMSB(n) = \sqrt{\frac{\sum_{k=-N/2}^{N/2-1} \left(PIMF(n) - f_k\right)^2 \bullet PIPD(0,k)}{\sum_{k=-N/2}^{N/2-1} PIPD(0,k)}} \quad (9)$$

with the same considerations as in equation (8).

6. ERROR ESTIMATION

Typically, in any spectral estimation, the error has two independent components (Cardoso, *et al.*, 1996). The first component represents the mean of the errors of the estimated values respect to the theoretic values. That error will be called the bias. The second component represents the standard deviation of those errors. Then, the root mean square (RMS) error is estimated according to:

$$error_{RMS} = \sqrt{bias^2 + std^2} \quad (10)$$

In case of calculating the error estimation of the PIMF, it can be done with:

$$error(n) = PIMF_{estimated}(n) - PIMF_{theoretic}(n) \quad (11)$$

$$bias = \frac{1}{m}\sum_{n=1}^{m} error(n) \quad (12)$$

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

$$std^2 = \frac{1}{m}\sum_{n=1}^{m}\left(error(n) - bias\right)^2 \quad (13)$$

where m is the total amount of fully overlapped signal windows of length L in the whole signal of length T, in consequence, $m = T\text{-}L+1$. Whereas, in case of calculating the error estimation of the RMSB, it can be done with:

$$error(n) = RMSB_{estimated}(n) - RMSB_{theoretic}(n) \quad (14)$$

with the same considerations as in equations (11), (12) and (13).

7.- RESULTS

The following graphs show the spectral estimation errors which are obtained for different values of scaling factors. Note that the graphs have a minimum point defining the optimum value of the scaling factor, that is, the optimal parameter.

The analysis is done separately for PIMF and RMSB. Similar results are obtained when analyzing the Carotid, Femoral or Coronary blood flow simulated signals.

7.1.- Bessel Distribution

Figures 2a and 2b shows the results obtained using Bessel distribution for window length of L=511 and without noise.

Table 1 shows optimal parameters for different window lengths and noise levels for both, PIMF and RMSB estimations.

	PIMF		RMSB	
L	Noiseless	SNR=30	Noiseless	SNR=30
127	3	3	3	3
255	3	3	3	3
511	3	3	3	3
1023	3	3	3	3

Table 1: Optimal parameters for Bessel distribution.

7.2.- Born Jordan Distribution

Figures 3a and 3b shows the results obtained using Born Jordan distribution for window length of L=511 and without noise.

Table 2 shows optimal parameters for different window lengths and noise levels for both, PIMF and RMSB estimations.

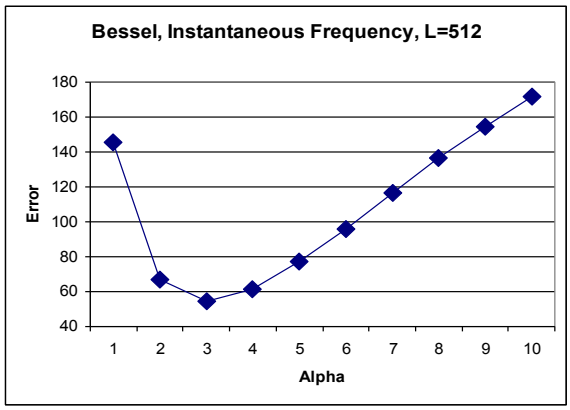


Fig 2a: PIMF estimation error as a function of scaling factor value using Bessel distribution. Optimal parameter for L=511 without noise is $\alpha = 3$.

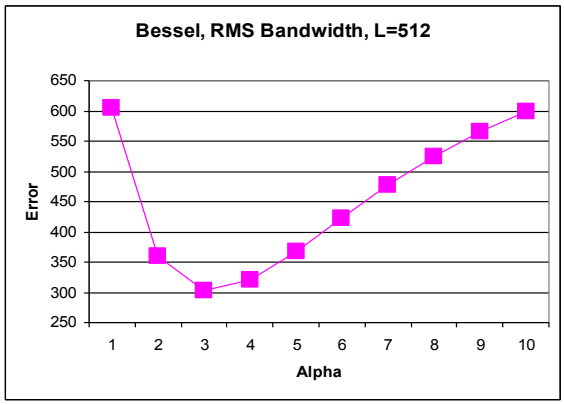


Fig 2b: RMSB estimation error as a function of scaling factor value using Bessel distribution. Optimal parameter for L=511 without noise is $\alpha = 3$

	PIMF		RMSB	
L	Noiseless	SNR=30	Noiseless	SNR=30
127	2	1	2	1
255	2	1	2	1
511	2	1	2	1
1023	2	1	2	1

Table 2: Optimal parameters for Bessel distribution.

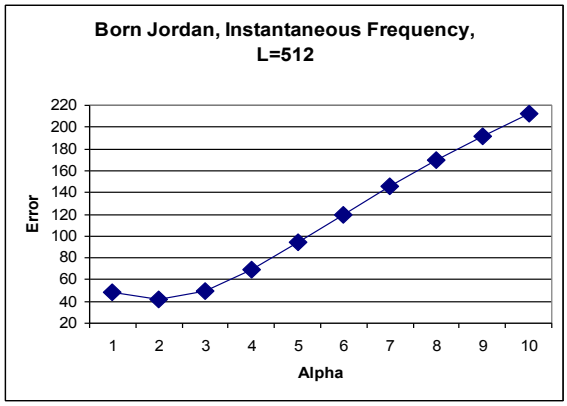


Fig 3a: PIMF estimation error as a function of scaling factor value using Born Jordan distribution. Optimal parameter for L=511 without noise is $\alpha = 2$.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

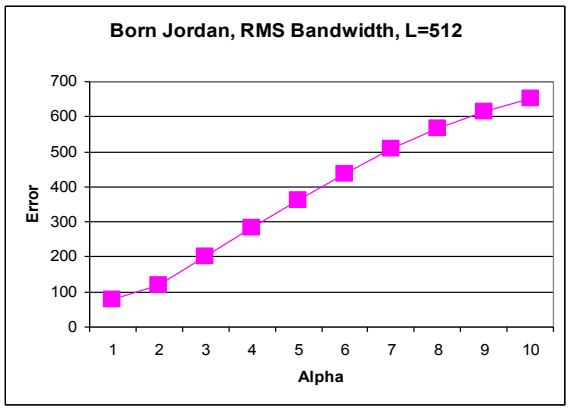


Fig 3b: RMSB estimation error as a function of scaling factor value using Born Jordan distribution. Optimal parameter for L=511 without noise is $\alpha = 1$.

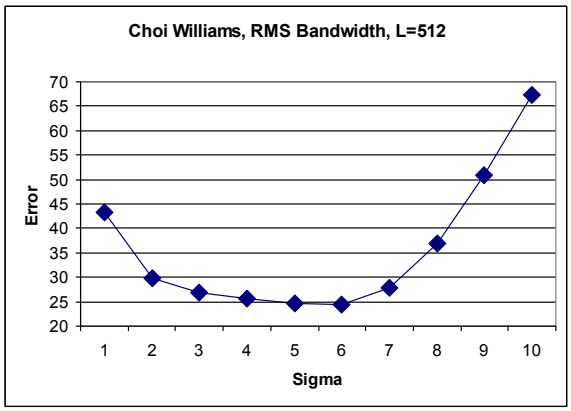


Fig 4b: RMSB estimation error as a function of scaling factor value using Choi Williams distribution. Optimal parameter for L=511 without noise is $\sigma = 6$.

7.3.- Choi Williams Distribution

Figures 4a and 4b shows the detailed results obtained using Choi Williams distribution for window length of L=511 and without noise.

Table 3 shows optimal parameters for different window lengths and noise levels for both, PIMF and RMSB estimations.

	PIMF		RMSB	
L	Noiseless	SNR=30	Noiseless	SNR=30
127	0.3	0.4	5	13
255	0.3	0.3	6	12
511	0.3	0.4	6	12
1023	0.4	0.6	6	12

Table 3: Optimal parameters for Choi Williams distribution.

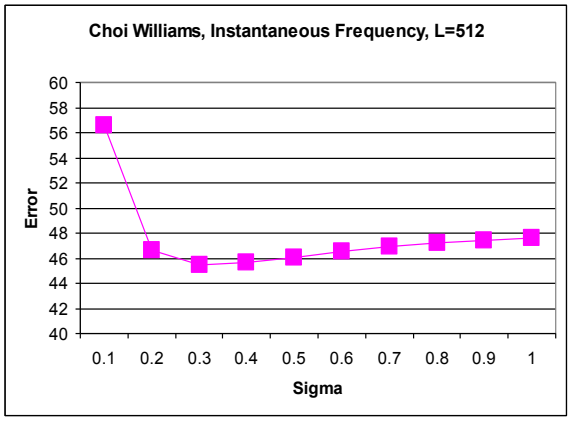


Fig 4a: PIMF estimation error as a function of scaling factor value using Choi Williams distribution. Optimal parameter for L=511 without noise is $\sigma = 0.3$.

7.4.- Modified-B Distribution

Figures 5a and 5b shows the results obtained using Modified-B distribution for window length of L=511 and with SNR=30dB.

Table 4 shows optimal parameters for different window lengths and noise levels for both, PIMF and RMSB estimations.

	PIMF		RMSB	
L	Noiseless	SNR=30	Noiseless	SNR=30
127	0.007	0.06	0.007	0.05
255	0.004	0.03	0.004	0.03
511	0.005	0.03	0.005	0.03
1023	0.007	0.03	0.007	0.03

Table 4: Optimal parameters for Modified-B distribution.

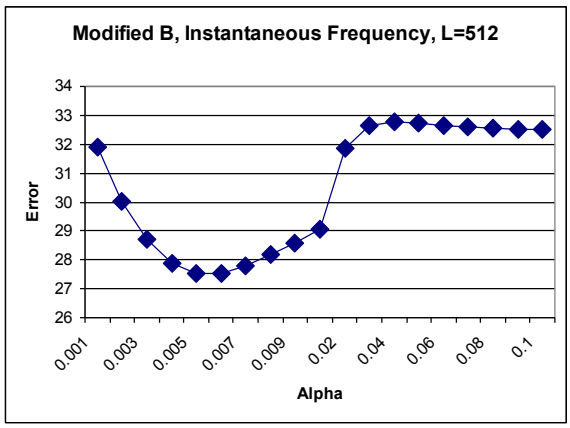


Fig 5a: PIMF estimation error as a function of scaling factor value using Modified-B distribution. Optimal parameter for L=511 with SNR=30dB is $\alpha = 0.005$.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

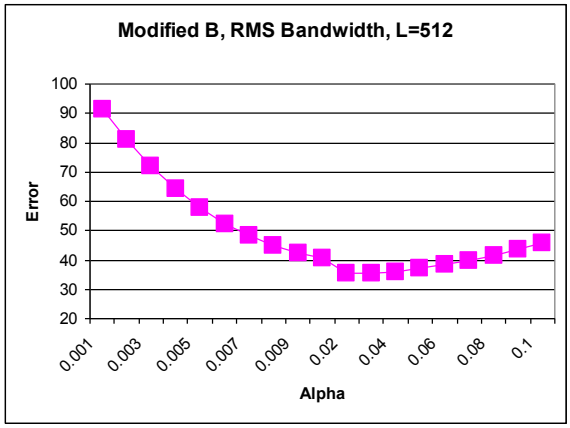


Fig 5b: RMSB estimation error as a function of scaling factor value using Modified-B distribution. The optimal parameter for L=511 with SNR=30 is $\alpha = 0.03$.

8.- RESULTS ANALYSIS

Tables 1, 2 3 and 4 show optimal parameters (optimal scaling factor values) experimentally obtained for Bessel, Born Jordan, Choi Williams and Modified-B respectively. Note that optimal parameters may depend on window length, SNR. Also, optimal parameters may be different for PIMF and RMSB spectral estimation.

Figure 6 compares the error in the PIMF estimation for the different TFD considered, using optimal parameters, without noise, and considering different window lengths (L=127, 255, 511 and 1023).

It is observed that the TFD that more accurately estimates the PIMF is the B-Modified, followed by Born Jordan and Choi Williams, estimating it nearly alike; finally Bessel distribution.

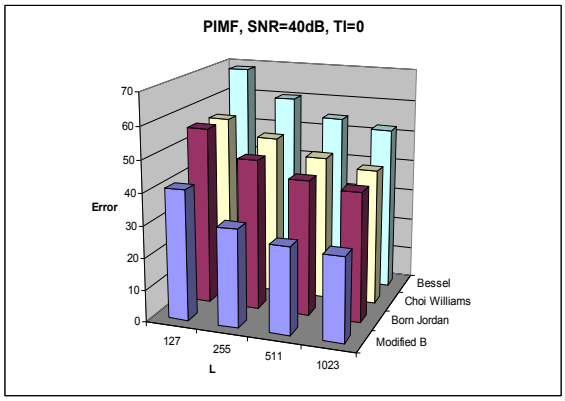


Fig. 6.- Error in estimating the PIMF with optimal parameters.

Figure 7 compares the error in the RMSB estimation for the different TFD considered, using optimal parameters, without noise, and considering different window lengths (L=127, 255, 511 and 1023).

It is observed that the TFD that more accurately estimates the PIMF are the B-Modified and Choi Williams, estimating it nearly alike, followed by Born Jordan and, finally, Bessel distribution.

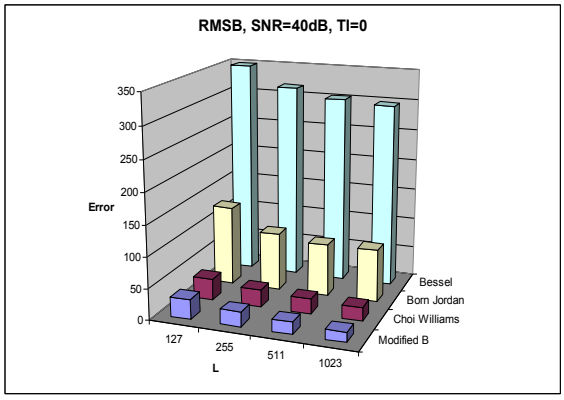


Fig 7.- Error in estimating the RMSB with optimal parameters.

9.- CONCLUSIONS

In this work the optimal scaling factor values, the so called optimal parameters, have been determined experimentally for different conditions of SNR and window length to estimate PIMF and RMSB. The results are shown in tables 1 to 4. The considered signals were three simulated Doppler ultrasound quasi-stationary signals that represent a typical blood flow in the Carotid, Coronary and Femoral arteries.

Bessel, Born Jordan, Choi Williams and Modified-B distributions were considered; tables 1, 2 3 and 4 show optimal parameters (optimal scaling factor values) experimentally obtained respectively. Note that optimal parameters may be different for estimating PIMF and RMSB. Also, the Modified-B distribution produces the best PIMF and RMSB spectral estimations.

10. ACKNOWLEDMENTS

The authors acknowledge project DGAPA-UNAM-PAPIIT (IN101213), project Consorciado CYTED (P506PIC0295) by the financial support. Also we want to acknowledge to M. Fuentes, J. Contreras, S. Padilla and M. Vazquez for their technical support in the development of this work.

REFERENCES

Boashash, B. and P. Black (1987). An Efficient Real-Time Implementation of the Wigner-Ville Distribution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. **ASSP-35**. 1611-1618.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Boashash B. and Sucic V. (2003) Resolution Measure Criteria for the Objective Assessment of the Performance of Quadratic Time- Frequency Distributions. *IEEE Transactions on Signal Processing*. **51**, 1253-1263.

Boashash B. Azemi G. and O'Toole J.M. (2013) Time-Frequency Processing of Nonstationary Signals. *IEEE Signal Processing Magazine*. **November 2013**, 108-119.

Cardoso, J. G. Ruano and P. Fish (1996). Nonstationary Broadening Reduction in Pulsed Doppler Spectrum Measurements Using Time-Frequency Estimators. *IEEE Transactions on Biomedical Engineering*. **43**. 1176-1186.

Choi, H. and W. Williams (1989). Improved Time-Frequency Representation of Multicomponent Signals Using Exponential Kernels. *IEEE Transactions on Acoustics, Speech and Signal Processing*. **37**. 862-871.

Cohen, L. (1989). Time-Frequency Distributions -A Review. *Proceedings of the IEEE*. **77**. 941-981.

De Lazzari C., et al. (2006) Coronary Blood Flow: Comparison between in Vivo and Numerical Simulation Data. Computers in Cardiology. 33 881−884

Evans D., McDickn N. (2000) Doppler Ultrasound. Physics, Instrumentation and Signal Processing. Ed. John Wiley & Sons, LTD. England. Second Edition.

Fan, L. and D. Evans (1994). Extracting Instantaneous Mean Frequency Information from Doppler Signals Using the Wigner Distribution Function. *Ultrasound in Med. & Biol*. **20**. 429-443.

García-Nocetti F., Solano J., Rubio E (2002) Precision enhancement of Doppler Ultrasound spectral estimation by finding TFD optimal parameters. *Forum Acusticum Sevilla. Special Issue of the Revista de Acústica.* **33**. September 2002. Sevilla, Spain.

García-Nocetti F., Solano J., Rubio E., Moreno, E. (2002) High Performance Computing of Time Frequency Distributions for Doppler Ulltrasound Signal Analysis. *Preprints of the 15th Triennial World Congress of the IFAC.* July 2002, Barcelona Spain**.**

Guo, Z., L. Durand and H. Lee (1994). The Time-Frequency Distributions of Nonstationary Signals Based on a Bessel Kernel. *IEEE Transactions on Signal Processing*. **42**. 1700-1707.

Hussain Z. And Boashash B. (2002). Adaptive Instantaneous Frequency Estimation of Multicomponent FM Signals Using Quadratic Time-Frequenc Distributions. *IEEE Transactions on Signal Processing*. **50**. 1866-1876.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

SESSION

POSTERS

Chair(s)

TBA

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

KBase: An Integrated Knowledgebase for Predictive Biology and Environmental Research

A. Palumbo[3], J. Baumohl[1], A. Best[2], J. Bischof[2], B. Bowen[1], T. Brettin[2], T. Brown[2], S. Canon[1], S. Chan[1], J.-M. Chandonia[1], D. Chivian[1], R. Colasanti[2], N. Conrad[2], B. Davison[3], M. DeJongh[6], P. Dehal[1], N. Desai[2], S. Devoid[2], T. Disz[2], M. Drake[3], J. Edirisinghe[2], G. Fang[7], J. P. Lopes Faria[2], M. Gerstein[7], E. Glass[2], A. Greiner[1], D. Gunter[1], J. Gurtowski[5], N. Harris[1], T. Harrison[2], F. He[4], M. Henderson[1], C. Henry[2], A. Howe[2], M. Joachimiak[1], K. Keegan[2], K. Keller[1], G. Kora[3], S. Kumari[5], M. Land[3], F. Meyer[2], S. Moulton[3], P. Novichkov[1], T. Oh[8], G. Olsen[9], R. Olson[2], D. Olson[2], R. Overbeek[2], T. Paczian[2], B. Parrello[2], S. Pasternak[5], S. Poon[1], G. Price[1], S. Ramakrishnan[5], P. Ranjan[3], W. Riehl[1], P. Ronald[8], M. Schatz[5], L. Schriml[10], S. Seaver[2], M. Sneddon[1], R. Sutormin[1], M. Syed[3], J. Thomason[5], N. Tintle[6], W. Trimble[2], D. Wang[7], D. Ware[5], D. Weston[3], A. Wilke[2], F. Xia[2], S. Yoo[4], D. Yu[4], R. Cottingham[3], S. Maslov[4], R. Stevens[2], A. Arkin[1].

[1] Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2] Argonne National Laboratory, Argonne, IL, USA
[3] Oak Ridge National Laboratory, Oak Ridge, TN, USA
[4] Brookhaven National Laboratory, Upton, NY, USA
[5] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[6] Hope College, Holland, MI, USA
[7] Yale University, New Haven, CT, USA
[8] University of California, Davis, CA, USA
[9] University of Illinois at Champaign-Urbana, Champaign, IL, USA
[10] University of Maryland, College Park, MD, USA

http://kbase.us

This paper was submitted as a poster.

Project Goals- *The KBase project aims to provide the computational capabilities needed to address the grand challenge of systems biology: to predict and ultimately design biological function. KBase enables users to collaboratively integrate the array of heterogeneous datasets, analysis tools and workflows needed to achieve a predictive understanding of biological systems. It incorporates functional genomic and metagenomic data for thousands of organisms, and diverse tools including (meta)genomic assembly, annotation, network inference and modeling, thereby allowing researchers to combine diverse lines of evidence to create increasingly accurate models of the physiology and community dynamics of microbes and plants. KBase will soon allow models to be compared to observations and dynamically revised. A new prototype Narrative interface lets users create a reproducible record of the data, computational steps and thought process leading from hypothesis to result in the form of interactive publications.*

Keywords: Bioinformatics, Metagenomics, Systems Biology

1 Introduction

The Department of Energy (DOE) Systems Biology Knowledgebase (KBase) is an emerging computational environment that enables researchers to bring together the diverse data, algorithms, analytical tools, and workflows needed to achieve a predictive understanding of biological

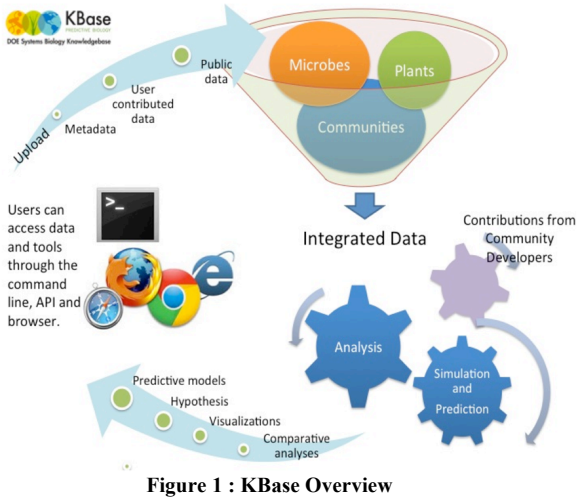


Figure 1 : KBase Overview

systems (see Fig.1). As a project supported by the Office of Biological and Environmental Research within the DOE Office of Science, KBase focuses on microbial and plant systems that support DOE missions in energy production and

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

environmental science. However, the KBase approach to analyzing and modeling DOE-relevant microbes and plants can be applied to organisms from across the tree of life.

2 KBase Overview

The overview diagram (Fig. 1) shows how users interact with KBase by uploading biological data, analyzing it with tools developed by both KBase and the community, and using analysis results to drive experiments and a better understanding of biological systems.

KBase is also an open and extensible development environment that invites and trains community members to contribute new tools and data. Tool developers can implement their methods as new KBase services, making their tool accessible to a wide user community and placing a world of biological data at their fingertips for tool validation. Data producers can integrate their data into the KBase data model, so that all of the analysis and visualization tools available in KBase may be applied to interpret the data. By enabling members of the community to integrate and use a wide spectrum of analysis tools and datasets, KBase will serve as a catalyst for biological research, accelerating discovery for DOE missions and providing insights and benefits that can ultimately serve numerous application areas.

Systems biology is driven by the ever-increasing wealth of data resulting from new generations of genomics-based technologies. With the success of genome sequencing, biology began to generate and accumulate data at an exponential rate. In addition to the massive stream of sequencing data, each type of technology that researchers use to analyze a sequenced organism adds another layer of complexity to the challenge of understanding how different biological components work together to form a functional living system. Achieving this systems-level understanding of biology will enable researchers to predict and ultimately design how the system will function under certain conditions. Gaining this predictive understanding, however, requires an unprecedented level of collaboration among researchers in different disciplines around the world. A new collaborative computational environment is needed to bring these researchers together so they can share and integrate large, heterogeneous datasets and readily use this information to develop predictive models that drive scientific discovery.

KBase is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Complex Networks Associated with Positive Selection and Drug Resistance in the Malaria Parasite

Timothy G. Lilburn[1], Hong Cai[2], and Yufeng Wang[2]
[1]328 Jamestown Road, Front Royal, Virginia, USA
[2] Department of Biology, South Texas Center for Emerging Infectious Diseases,
University of Texas at San Antonio, San Antonio, TX 78249, USA

Abstract –*Mu et al. have published an extensive dataset of SNPs collected in the context of the response of the malaria parasite to drug treatment. We looked for networks centered on proteins shown to be under positive selection in the face of drug treatments. Thesees proteins were often associated with subnetworks involved in invasion of the host cell by the parasite.*

Keywords: malaria, networks, evolutionary analysis

1 Introduction

Malaria is one of the most serious infectious diseases in the world, affecting 300-500 million people, and is responsible for nearly one million deaths every year. The rapid evolution and spread of drug resistance in parasites has led to an increase in morbidity and mortality rates in malaria-afflicted regions. Drugs impose positive selection on the malaria parasite *Plasmodium falciparum*, leaving genomic signatures. Recent genome-wide association studies (GWAS) of 189 parasite isolates revealed thousands of single nucleotide polymorphisms (SNPs) [1]. Here we explore the protein-protein associations of genes that were under positive selection for drug response. Our network analysis reveals previously uncharacterized proteins that are implicated in a wide variety of cellular processes, including transport, transcriptional regulation, translation, signal transduction, cell cycle, entry to the host, and metabolism. This systems-level analysis allowed us to identify the evolutionary signatures of networks that show genotype-phenotype association, providing new insights into parasite biology, pathogenesis

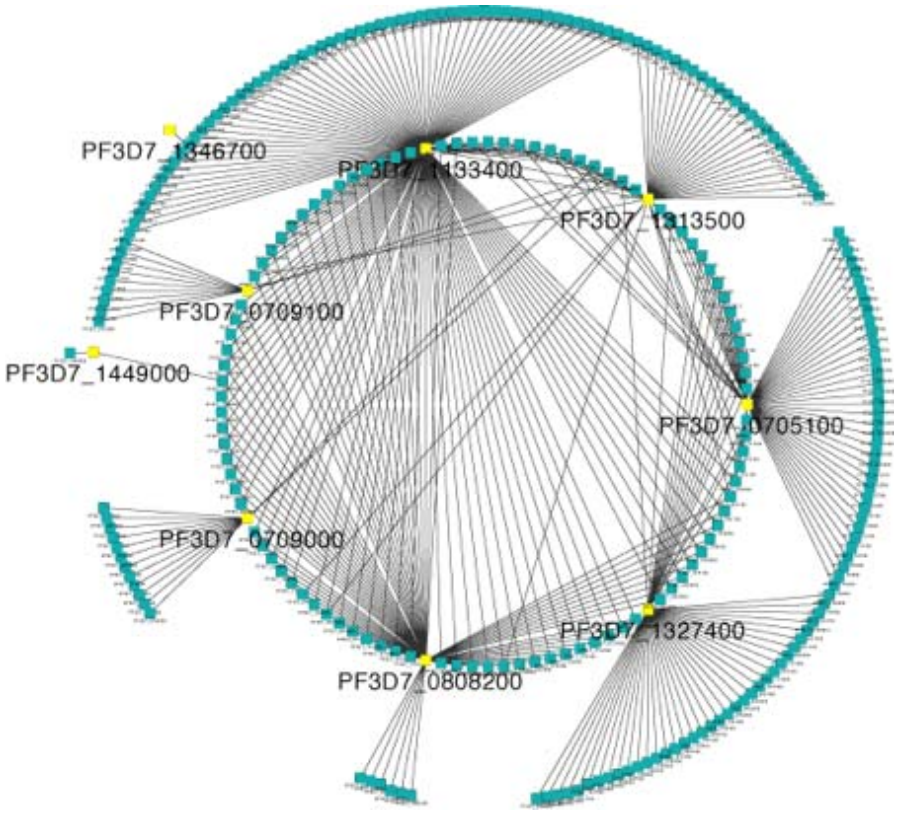


Figure 1. The largest subnetwork generated using the iHS proteins from Mu et al. The iHS proteins are shown in lighter gray with their gene IDs. Note that only the 390 edges between the high iHS proteins and their first neighbors in the subnetwork are shown.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-3, Printed in the United States of America

and virulence. The hub proteins with high connectivity showing a high degree of genetic variability can serve as potential drug targets.

2 Results and Discussion

We downloaded the set of high confidence protein-protein associations (defined as having an S_value ≥ 0.7) for *P. falciparum* from the STRING database. Mu et al. found 51 loci with integrated haplotype scores (iHS) that were unusually large (> |2.3|), indicating that these alleles were responding to selective pressures imposed by drug treatments. Twenty-four of these loci were in the high confidence association network. We extracted the subnetworks that involved the proteins encoded by these loci. They ranged in size from two to 288 nodes. Figure 1 shows the largest subnetwork. Although only 390 direct links are shown, in reality 8,809 edges connect this set of nodes, indicating that the subnetwork is modular, a result supported by its relatively high clustering coefficient (0.684). The largest subnetwork contains three of the iHS proteins from the Mu et al. paper. The two most highly connected of these nine proteins are plasmepsinX (PF3D7_0808200) and the apical membrane antigen 1 (AMA1) (PF3D7_1133400). The former is an aspartyl protease and is thought to be involved in host cell invasion and in growth and replication of the parasite inside the host cell. The latter is an adhesion and is also involved in host cell invasion. The next three most highly connected iHS proteins in this subnetwork are all of unknown function, although they are conserved across multiple species of *Plasmodium*. In fact, just over half of the nodes in this subnetwork represent proteins of unknown function, so it seemed worthwhile to do an enrichment analysis. Given that the two most highly connected proteins are linked to host cell invasion, it is not surprising that the enrichment analysis (see Figure 2) strongly indicates that this set of proteins is associated with merozoite movement into the host cell. One of the iHS proteins in this subnetwork, AMA1 (PF3D7_1133400), has already been identified as potential vaccine candidate, as blocking it with antibodies inhibits cell invasion. A second, P48/45 (PF3D7_1346700), also induces an antibody response in the host. The three proteins of unknown function (PF3D7_0705100, PF3D7_1327400, and PF3D7_1313500) are not surface-associated, but may play a crucial role in the virulence of the parasite. Disruption of PF3D7_0705100 has been shown to attenuate blood-stage growth and the other two may play similarly important roles..

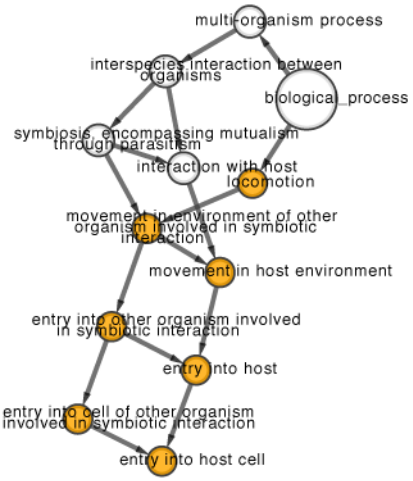


Figure 2. Enrichment analysis for proteins in the 288 node subnetwork. Analysis was done with BiNGO, using default settings. Gray color indicates a probability of 2 x 10^{-7} that the proteins represent a random set

3 Conclusions

The approach laid out here will help researchers understand how to circumvent drug resistance in the malaria parasite and suggest novel drug targets..

4 References

[1] Mu J, Myers RA, Jiang H, Liu S, Ricklefs S, Waisberg M, Chotivanich K, Wilairatana P, Krudsood S, White NJ, Udomsangpetch R, Cui L, Ho M, Ou F, Li H, Song J, Li G, Wang X, Seila S, Sokunthea S, Socheat D, Sturdevant DE, Porcella SF, Fairhurst RM, Wellems TE, Awadalla P, Su XZ. "Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs" Nat Genet, 42, 3, 268-271, Mar 2010.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Real-time photoacoustic tomography using linear array transducer and its phantom evaluation

Chul Gyu Song[1], Dong Won Lee[2], Bang Young Kim[1], Dong Ho Shin[1]
[1]Division of Electronic Engineering, Chonbuk National University, Korea
[2]Dept. of BIN Fusion Technology, Chonbuk National University, Korea

Abstract

Photoacoustic Tomography (PAT) is a promising medical imaging modality by reason of its particularity. It combines optical imaging contrast with the spatial resolution of ultrasound imaging, and can distinguish changes in biological features in an image. For these reasons, many studies are in progress to apply this technique for diagnosis. But, real-time PAT systems are necessary to confirm biological reactions induced by external stimulation immediately. Thus, we have developed a real-time PAT system using a linear array transducer and a custom-developed data acquisition board (DAQ). To evaluate the feasibility and performance of our proposed system, two types of phantom tests were also performed. As a result of those experiments, the proposed system shows satisfactory performance and its usefulness has been confirmed.

Keywords-Photoacoustic Tomography, Linear array transducer, Real-time imaging

I. INTRODUCTION

Photoacoustic Imaging(PAI) is a hybrid imaging modality which combines features of optical and acoustical imaging. Pure optical imaging like optical coherence tomography (OCT) has high image contrast but low imaging depth. PAI has high image contrast similar to optical imaging, but overcomes some weaknesses of optical imaging by using acoustical measurements[1-4]. In addition, PAI can measure biological changes because components of biological tissue have different optical absorption, which is a main factor in the photoacoustic effect. To confirm a biological reaction induced by external stimulation immediately using PAI methods *in vivo*, real-time Photoacoustic Tomography (PAT) systems have recently been studied by many research groups, In this study, we present a real-time PAT system for primary study on PAI using a linear array probe and a custom-developed 64ch data acquisition board (DAQ), and evaluate the feasibility of our proposed system using two types of phantom test.

II. REAL TIME PAT SYSTEM

A custom-developed trigger controller is dedicated to laser-emitting, and the trigger for acquisition timing generation and a linear array probe (L14-5/28, Ultrasonix) with 5 MHz center frequency and 128 element transducers were used to measure the photoacoustic signal. To amplify and filter the detected signal, a custom-developed pre-amplifier was placed ahead of the DAQ, and had 40 dB gain, 5 MHz passband and 64 channels. Figure 2 shows a schematic diagram of the developed real-time PAT system. To evaluate the feasibility and performance of our proposed system, two types of phantom test were performed.

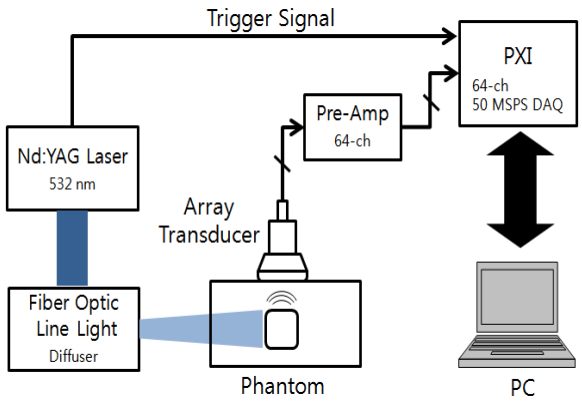


Fig. 1 Schematic diagram of proposed PAT system

The first phantom examines the computation of frame rate. Rectangular-shaped gelatin (Protein : 84 ~ 90 %, Water : 8 ~ 12 %) is prepared to imitate biological tissue, and a transparent polymer tube with 1 mm inner diameter is embedded at 8 mm below the surface of the gelatin with a tilt of approximately 5.3 degrees, as in figure 5. Then, the changes in the image were recorded while injecting magenta dye through the tube gradually for 10 seconds. The source was a Nd:YAG laser (Meditech, Eraser-k) with 532 nm wavelength, and the repetition rate was 10 Hz.

III. RESULTS AND DISCUSSION

The results of the first phantom test for real-time image acquisition are represented in Figure 3. In this picture, PAT images of magenta dye flowing through the transparent polymer tube are distinct as time goes on, and because of the laser repetition rate, the frame rate of the PAT system was 2 frames/sec. This result indicates that our system has

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

performance sufficient to reconstruct a 570 * 300 pixel image. Figure 7 shows, however, quite dissimilar thickness between the image and the actual phantom. This result is because the Fiber Optic Line Light is not suitable for the diffuser to cover the whole scan region, and shows only the upper boundary of the polymer tube.

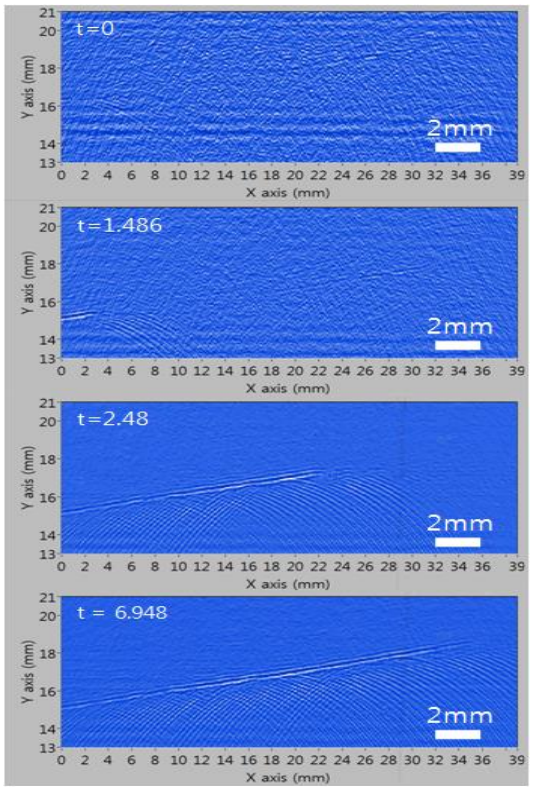


Fig. 3 Phantom for real-time PAT imaging

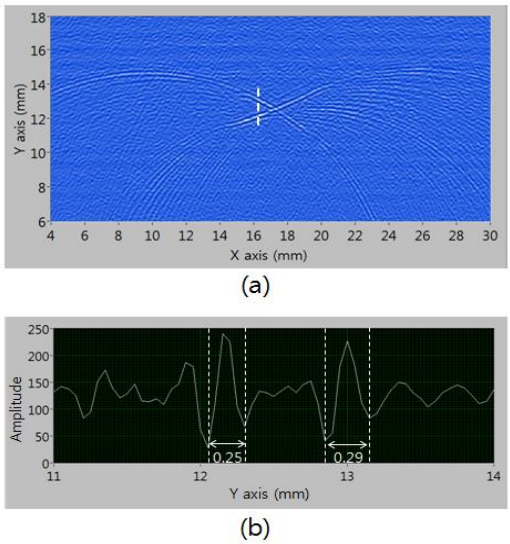


Fig. 4 (a) PAT image of hair phantom (b) Zoom of white dotted line

The results of the next experiment are shown in figure 4(a). The outline of the crossed hairs and their intersection are evidently classified, but the reconstructed image in figure 4(b), which is measured for the area under the white dotted line in figure 4(a), shows a measurement 10 times thicker than the real 0.03 mm hair thickness. When an acoustic speed of 1500 m/sec is assumed and since the transducer has a 5 MHz enter frequency, the predicted resolution is approximately 0.15 mm, which is also not enough to indicate actual thickness. Thus, the results of the experiment are predicted to show the maximum resolution of the developed PAT system, which was found to be approximately 0.3 mm.

IV. CONCLUSION

In this study, we developed a real-time PAT system for functional imaging *in vivo* and confirmed the feasibility of the proposed system through two types of phantom test. The results of phantom test show that our PAT system has enough performance to obtain a 570 * 300 pixel image at 2 frames/sec with approximately 0.3-mm resolution. Essentially, the frame rate of the real-time PAT system depends entirely on the laser radiation speed, because of the synchronization between the radiation and the detection. The maximum radiation rate of the laser source used was 10 Hz, and in the same conditions as the phantom tests, the system showed a rate of 9 frames/sec with a 10 Hz repetition rate, but poor image SNR, because a fast radiation rate reduces the source power. Therefore, further study for the improvement of the image reconstruction algorithm is being conducted in order to improve the process time in the present system and eliminate the side effects caused by the limited angular view of the linear array probe to enhance the contrast in reconstructed image.

V. ACKNOWLEDGMENT

This work was supported by the Human Resources Development of the Korea Insitute of Energy Technology Evaluation and Planning(KETEP) grant funded by the Korea government Ministry of Knowledge Economy(No. 20124030200080) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MEST 2011-0030075)

REFERENCES

[1] A. Oraevsky and A. Karabutov, "Optoacoustic tomography," in Biomedical Photonics Handbook, T. Vo-Dinh, CRC, Boca Raton, Ed., Chap. 34, pp. 1-34, 2003.

[2] L. V. Wang and H. Wu, Ed. Biomedical Optics : Principles and Imaging, John Wiley, 2007.

[3] Lihong V. Wang, Ed. Photoacoustic Imaging and Spectroscopy, 1st ed., Boca Raton : CRC Press, 2009

[4] Changhui Li, Andres Aguirre, John Gamelin, Anastasios Maurudis, Quing Zhu, and Lihong V. Wang, "Real-time photoacoustic tomography of cortical hemodynamics in small animals," J. Biomedical Opt., Vol. 15, No. 1, p. 010509, Jan./Feb. 2010

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

An Interactive Visualization Tool to Interpret Transcriptomics Data

Salil Pendse[1], Patrick D. McMullen[1]

Address: [1]The Hamner Institutes for Health Sciences, 6 Davis Drive, RTP, NC-27709, USA

Email: Salil Pendse - sPendse@thehamner.org (contact author); Patrick D. McMullen - PMcMullen@thehamner.org

***Abstract*—High throughput gene expression studies generate large numbers of genes of interest. Functional gene ontologies such as Reactome [1] and Gene Ontology [2] are used to infer mechanistic processes underlying gene expression changes. However traditionally, this analysis is difficult to communicate and is represented by means of tables. To overcome this problem, we have developed an interactive visualization tool that treats ontologies as hierarchical networks of cellular functions. This allows for consolidation of the typically overwhelming spreadsheets of gene expression data into parsimonious descriptions of inferred functional changes in the cell.**

***Index Terms*—Functional Genomics, Interactive, Transcriptomics, Network Visualization**

I. Browser Based Interface

A major weakness of using functional ontologies is that they are represented using tables and static two-dimensional graphics that are fundamentally limited in their ability to convey high dimensional data. We are designing a solution to this problem that takes advantage of the interactive potential of modern computational tools. This allows the consolidation of large amounts of data in a format that enhances data exploration and hypothesis generation.

Functional ontologies are nested descriptions of biology. Because of their hierarchical nature, there is useful information in the relationships between the terms that are associated with the data. We have preserved these parent-child relationships as edges in our functional annotation tool. Figure 2 is a screenshot of the tool running in Google Chrome browser. Ontology terms are shown as nodes if they are significantly enriched in the data (blue) or if they are required to connect significant nodes into a coherent network (white).

The tool incorporates multiple publicly available ontologies (Reactome and Gene Ontology) with the ability to add more as required. By selecting different conditions using sliders, one can instantly see which functional categories are enriched. Here we have used data from a study by Dere *et al.* [3] exposing C57BL/6 mice and Hepa1c1c7 mouse hepatoma cells to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and studying gene expression patterns of the response. Using the sliders one can see enrichment of categories occurring at selected combinations of experimental conditions. Further detail regarding a specific category can be obtained by mousing over the node of interest in the network. Clicking a category also gives us the list of genes from the current condition enriched in the selected category. The tool also allows us to combine up to three conditions to identify common pathways (Figure 1). The layout of the network itself is completely interactive, allowing us to drag and reposition any node as we see fit. Additionally, the tool allows us to export the network in various formats. The data from the network can be exported as flat tables or in a format readable by the popular visualization tool Cytoscape [4].

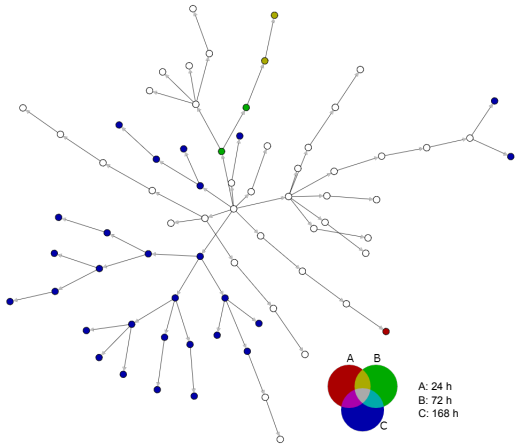


Figure 1. Combining multiple condition to view pathway differences. Here pathways enriched after mice were treated with TCDD for 24, 72 and 168 h are combined to see how function changes. It is immediately evident that at 168 h, TCDD has started eliciting immune response and the mice are no longer able to effectively clear the toxin from their liver[3]

II. Desktop Application

Creating these visualizations requires considerable time and programming knowledge. This prevents users with little programming experience from using the tool to quickly glean information from the data. To overcome this we have developed a spreadsheet-like desktop application that can be used to generate the networks and visualizations described above.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

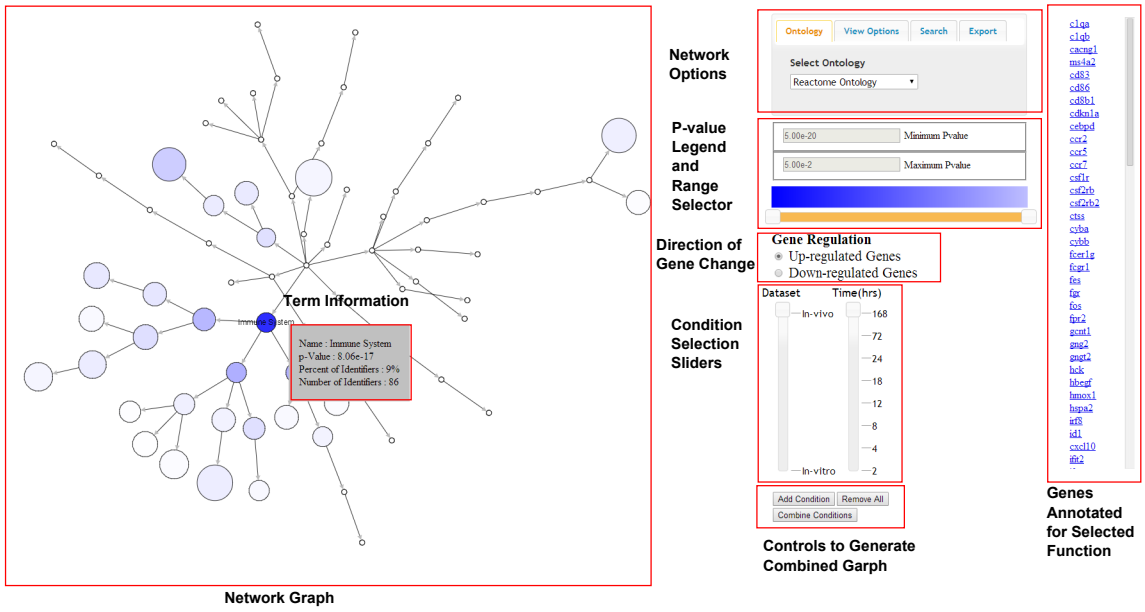


Figure 2. Screenshot of the interactive applet. The applet was created using Cytoscape.js JavaScript library

Figure 3 is a screen shot of the desktop application. The spreadsheet interface allows the user to upload data and enter all the information relevant to the study to be visualized. Columns in the spreadsheet refer to experimental conditions such as dose, time, cell line, etc.

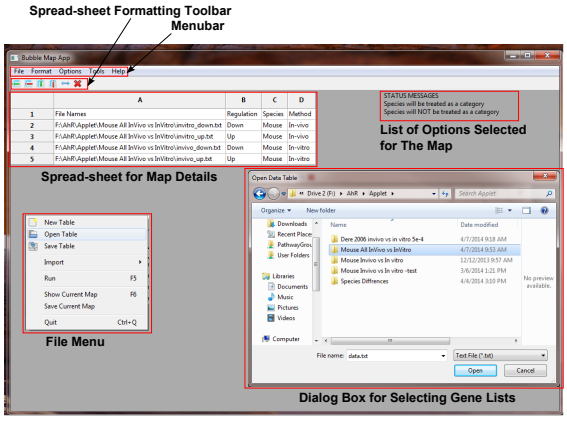


Figure 3. Desktop application for generating the interactive network. The application was created in Python using the wxWidgets library

Data derived from rats, humans or mice can be visualized in the same network. This allows us to compare not just across experimental conditions, but also across species. The ontologies used are stored as local copies when the application is installed. This allows for quick calculation of enrichment. However as the these ontologies are continuously being updated, a feature to update local copies has been included. This allows us to update the ontologies to their most current state, without having to manually download and extract information from the online database.

Together these tools provide a platform for intuitive interpretation of gene expression arrays and other high-throughput experiments.

Acknowledgments

This project was supported by Alternatives Research and Development Foundation(ARDF) project grant.

References

[1] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio, "The Reactome pathway knowledgebase.," *Nucleic acids research*, vol. 42, pp. D472–7, Jan. 2014.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, G. M. Rubin, and G. Sherlock, "Gene Ontology : tool for the unification of biology," vol. 25, no. 1, pp. 25–29, 2011.

[3] E. Dere, D. R. Boverhof, L. D. Burgoon, and T. R. Zacharewski, "In vivo-in vitro toxicogenomic comparison of TCDD-elicited gene expression in Hepa1c1c7 mouse hepatoma cells and C57BL/6 hepatic tissue.," *BMC genomics*, vol. 7, p. 80, Jan. 2006.

[4] M. E. Smoot, K. Ono, J. Ruscheinski, P. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization.," *Bioinformatics (Oxford, England)*, vol. 27, pp. 431–2, Feb. 2011.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

A Study of the Integration of Biomedical Devices with Information Technology Systems with an emphasis on Information Quality

Jim McGinnis, MS, ABD
STEM College, Computer and Information Sciences
University of Arkansas Fort Smith, Fort Smith AR, USA

The cost of Information Technology in the US per year is estimated at 1.7 trillion of dollars. While this seems to be a large investment that should produce large returns in terms of quality care, patient safety, and improved healthcare, there seems to be a disconnect between Information Technology(IT) and the medical industry/systems [1], [2]. A disconnect or distrust of the data between bio medical devices and IT systems can become evident when data from bio medical systems (e.g., Electrocardiogram or EKG) are integrated or converged across a network into an IT system such as an Electronic Medical Record System when proper policies and procedures are not in place. Integration is defined by Merriam –Webster as the combining and coordinating of separate parts or elements into a unified whole [3]. While an integration of an EKG is mentioned and used in this proposal, case studies on other integrations will take place as well. The assumption that all the data gathered via an interface or device will be of a quality nature to the receiving system is not necessarily correct. Problems with electronic medical records and other IT heath systems can lead to problems including loss of data [4]. This study proposes to look at the possible disconnect between IT and medical systems to try to ascertain why the investment does not produce the returns for healthcare. Some issues that will be researched include data quality, physician and medical staff acceptance and usage, and the problems that arise from implementation of new IT systems and the effects on healthcare patients and whether an integration project is a "success" [1].

The research will utilize a case study approach at various hospitals. The research would investigate completed or near completed integration projects concerning the integration of biomedical equipment and Information Systems (IS). The case studies will be conducted using interviews, surveys and collecting existing data. The case studies will compare data that was generated pre-implementation of the projects and post-implementation. The analysis will be based on comparing and contrasting the data collected pre and post implementation. The use of the case study will enable the research to utilize qualitative data as well as quantitative data. The settings (environment or culture) of the data collection can have a direct reflection on the end-user's perspective [5]. By using both qualitative and quantitative data, a case study can examine both the processes and the outcome of disconnect between IT and medical areas [6]. The research is important from the stand point of integration/quality, in how a project is conducted and how it will benefit users in the future and hopefully with the development of a functioning framework for integration and data quality (there is no standardized, simple approach to successful integration [7]. Few studies are available for IT or medical staffs to study for improving projects [8] [9]. The research that has been completed to this point suggests that studies are being performed on interoperability of device and IT with most looking at patient

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

outcomes but not on the data transfer quality [10] [9] [8]. This proposed research should enhance this area by building and incorporating an integration framework for use by IT and the medical field.

Citations

[1] R. J. Holden and B.-T. Karsh, "The Technology Acceptance Model: It's Past and it's Future in Health Care," 15 July 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2814963/?tool=pmcentrez.

[2] R. J. Holden and B.-T. Karsh, "The Technology Acceptance Model: Its Past and Its Future in Health Care," Elsevier, 2010.

[3] Merriam-Webster, "Merriam-Webster," 2012. [Online]. Available: http://www.merriam-webster.com/dictionary/convergence.

[4] K. Hobson, "The top Ten Health Technologhy Threats to Patient Safety," 8 December 2010. [Online]. Available: http://blogs.wsj.com/health/2010/12/08/the-top-ten-health-technology-threats-to-patient-safety/.

[5] E. H. Shortliffe, "Strategic Action in Health Information Technology: Why the Obvious has taken So Long," *Health Affairs,* pp. 1222-1233, 2005.

[6] Z. Zainal, "Case Study as a resarch method," *Jurnal Kemanusiaan,* 2007.

[7] E. Vedvik, A. Faxvaag and A. H. Tjora, "Beyond the EPR: Complemenatary roles of the hospital wide electronic health record and clinical departmental systems," *Biomed Central,* 12 June 2009.

[8] R. Amarasingham, L. Plantinga, M. Diener-West, D. J. Gaskin and N. R. Powe, "Clinical Information Technologies and Inpatient Outcomes," *Arc Internal Medicine,* pp. 108-114, 2009.

[9] B. Chaudhry, W. Jerome, W. Shinyi, M. Maglione, W. Mojica, E. Roth, S. Morton and P. G. Shekelle, "Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care," *Annals of Internal Medicine,* pp. 742-752, 2006.

[10] Directors, "Interoperability Definition and Background," HIMSS, 2005.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

SESSION

LATE BREAKING PAPERS AND POSITION PAPERS: BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Chair(s)

Prof. Hamid R. Arabnia

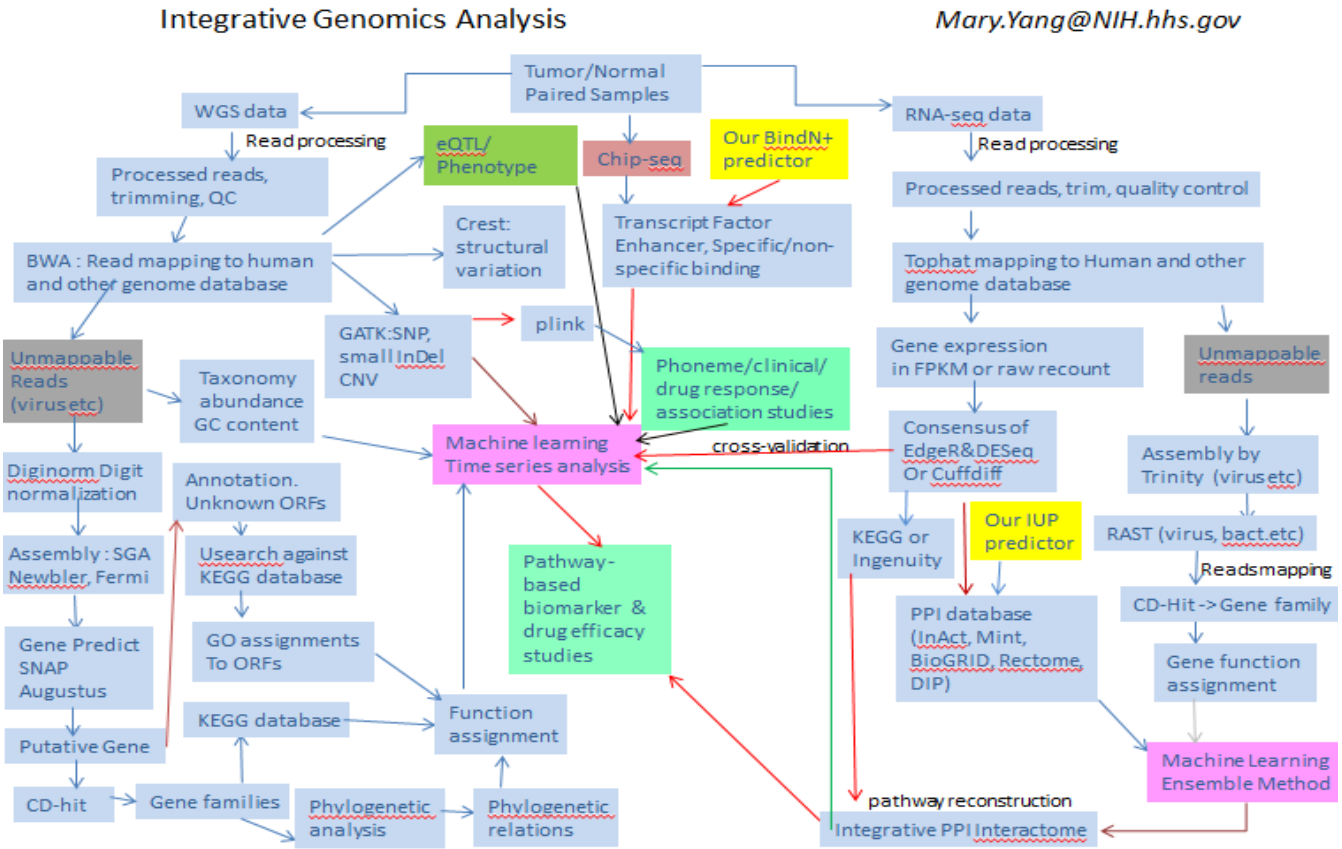
Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Invited Talk:
Integrative systems biology approaches to identify disrupted pathways in disease development

Mary Yang, MSECE, Ph.D.

Advancement of next-generation sequencing technologies has facilitated the identification of large number of genomic mutations, including SNPs (single- nucleotide polymorphisms), CNVs (copy number variations) and SV (structure variations). Large-scale genomic, transcriptomic and proteomic data have generated unprecedented opportunities to disease studies in the context of knowing the entire catalog of genetic mutations and related pathways. However genome-wide association studies (GWAS) still remain challenging, many genomic mutations identified by high-throughput approaches either contribute to small disease risk effect, or lesser population due complex molecular interactions and inter-patient heterogeneity. Motivated by our newly developed systems biology approaches as shown in the figure, we combine DEG (differential expression of genes) from RNA-seq and PPI (protein-protein interaction) data with the multi-variant eQTL (expression quantitative trait loci) mapping approaches from WGS (whole genome shortgun) data with phenotypic information to identify the disrupted networks relating to the disease development. In addition, incorporating lncRNA (long non-coding RNA) expression data with disease-disrupted networks has offered new insight into the regulatory structure of the disease-associated networks. Our integrative approaches make an important step toward better understanding disease mechanisms and finding better therapeutic drug targets.



Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Bio-: Mary Yang received the MS, MSECE and Ph.D. degrees from Purdue University, West Lafayette. She joined the National Human Genome Research Institute in 2005, where she has been working on various projects related to genomics and systems biology. She has been Founding Editor-in-Chief of International Journal of Computational Biology and Drug Design, a NIH PubMed fully indexed journal and is on editorial broads of Journal of Supercomputing and International Journal of Pattern Recognition and Artificial Intelligence. She is currently Associate Professor and Director of the Joint Bioinformatics Ph.D. Program of University of Arkansas Little Rock College of Engineering & Information Technology - University of Arkansas for Medical Sciences and HHS/FDA/NCTR in Arkansas. Dr. Yang has published over 50 journal papers in both computer science and biomedical sciences. She is the recipient of the Bilsland Dissertation Fellowship, the Purdue Research Foundation Fellowship, the ISIBM and IEEE Bioinformatics and Bioengineering Outstanding Achievement Award, and the NIH Fellows Award for Research Excellence.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

SMIR: a method to predict the residues involved in the core of a protein

R. Acuña[1], Z. Lacroix[1], J. Chomilier[2,3], and N. Papandreou[4]
[1]Scientific Data Management Laboratory, Arizona State University, Tempe, AZ, USA
[2]IMPMC, Sorbonne Universités, Université Pierre et Marie Curie, CNRS UMR 7590, MNHN, IRD UMR 206, Paris, France
[3]RPBS, Université Paris Diderot, Paris, France
[4]Department of Biotechnology, Agricultural University of Athens, Athens, Greece

Abstract - *Protein folding is the critical spontaneous phase when the protein gains its structural conformation. If errors occur in the process, the protein structure may fail to fold properly. We present a new method SMIR that identifies the residues involved in the protein core. A Monte Carlo algorithm is used to simulate the early steps of folding to determine the number of non-covalently bound neighbors. Residues surrounded by many others may play a role in the compactness of the protein and thus are called Most Interacting Residues (MIR). The original MIR method was updated and extended with a new smoothing method using hydrophobic-based residue analysis. SMIR is available as a web server. SMIR is free and open to all users as functionality of the Structural Prediction for pRotein fOlding UTility System (SPROUTS) at http://sprouts.rpbs.univ-paris-diderot.fr/mir.html. The new server also offers a user-friendly interface and access to previous results.*
Contact: Zoé Lacroix

Keywords: MIR, SMIR, simulation, protein, folding, lattice.

1 Introduction

Amino acids involved in inter residue contacts may play a role in the compactness of the protein and thus are called Most Interacting Residues (MIR). The MIR method was first introduced to simulate the origin of protein folding [1]. Starting from a random conformation, the folding process can be dynamically simulated in a discrete space (a lattice). Successive residues that collapse and form a local compact structure (linked to another one by an ex-tended polypeptide chain) form a fragment. The MIR method focuses exclusively on the early steps of the folding process. In its very first implementation, it aimed to delineate the fragments formed at this stage. For this reason, the method was calibrated with time limits to maximize the number of fragments before the folding process reaches a single compact domain. It assigned a score between 2 and 8 to each residue, corresponding to the mean number of non-covalent neighbors in the lattice. A high score indicates that the residue is buried, thus belongs to a fragment. A low score indicates that it is a low interacting residue, belonging to a piece of the chain which links two consecutive fragments. A correspondence between fragments and regular secondary-structure elements (SSE) was demonstrated on a set of 42 proteins, representative of various folds [1]. However, it has been shown that a pertinent analysis of globular protein structures with respect to folding properties consists in describing them as an ensemble of contiguous closed loops [2] or Tightened End Fragments (TEFs) [3]. Such description reveals that the ends of TEFs are fold elements crucial for the formation of stable structures and for navigating the very process of protein folding. Meanwhile, the MIR algorithm evolved and newer versions (including the actual presented one) aim at locating individual residues with very high mean number of neighbors (typically ≥ 6), which are called the MIRs. In the other limit, individual residues with low mean number of neighbors (typically 2) are the Least Interacting Residues (LIRs). Therefore, the residues identified as MIRs have the tendency to be buried at the early stages of the folding process. The comparison of MIR positions with the positions of the limits of closed loops, in proteins of known 3D structures, showed a statistically significant agreement. MIRs also significantly correlate with topohydrophobic positions, i.e., positions in multiple alignments of sequences of common fold occupied only by hydrophobic amino acids, and correlated to the folding nucleus [4], thereby giving a route to simulations of the protein folding process [5]. Thus, MIR is a potential method for an ab initio estimation of the residues which are important for folding and consequently, significantly sensitive to mutations.

It is important to keep in mind the difference between protein core and nucleus. Core is a static concept, and it results from the fact that a globular protein is a micelle, with an internal phase of hydrophobic character, and an external phase of hydrophilic character, statistically. The core of a protein can be derived by a simple accessible surface area calculation or with more sophisticated methods [6]. In contrast, nucleus is a dynamic concept, it relies on a model of folding, namely the nucleation condensation model [7]. In a few words, a small set of dispersed amino acids come into

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

contact during the folding because of the thermal vibrations of the molecule. They are hydrophobic, and once they form such a nucleus, the rest of the structure can be formed. Among proteins sharing the same fold, part of the nucleus is conserved. Be-sides, it is now documented that nonnative contacts are necessary for the folding, and they disappear once the stability is sufficient. Figure 1 illustrates the difference between core and nucleus in the case of a fibronectin.

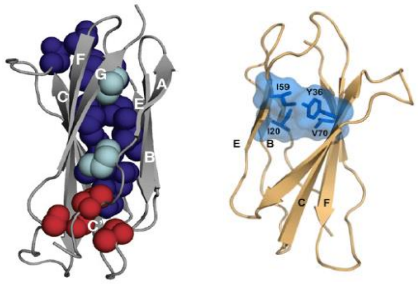


Figure 1: Difference between core (left) and nucleus (right) of the type III fibronectin [8, 9].

The knowledge of the residues constituting the folding nucleus is important for instance in the annotation of misfolding-related pathologies, but their experimental determination is not 100% secure. The role of prediction, at this moment, is a valuable complementary approach. The literature commonly admits that the number of residues involved in the folding nucleus is typically less than 10% of the sequence length, roughly one third of the hydrophobic residues. Initial MIR calculation slightly over predicts the nucleus. One guide line to improve prediction can be to produce a smoothing of the curve of NCN as a function of the sequence. This is one of the major improvements proposed with the SMIR method.

The SMIR method presented in this paper aims at improving the accuracy of MIR in the prediction of residues involved in the folding nucleus. Indeed it has been shown that MIR overestimates the folding nucleus of numerous proteins. The SMIR method is implemented and available as a server that supports the submission and the analysis of protein structures with MIR2.0 and SMIR. The server offers a dynamic interface with the display of results in a 2D graph.

2 Methods

The MIR method is an extension of previous simulations performed on cubic lattices, devoted to the complete folding of globular domains [10]. The MIR algorithm is a topological calculation, resulting from a series of energy-driven simulations of a protein backbone, where the mean number of non-covalent contacts is deduced for each residue. The analysis is performed at the early steps of folding and provides the number of Non-Covalent Neighbors (NCNs) for each residue in the sequence.

The simulation of the early steps of the folding is designed in the following manner. First, an extended initial conformation is produced for an alpha-carbon-only simplified representation of the polypeptide chain. Each alpha carbon is placed at random (while con-strained as a chain) on the nodes of a lattice. An extension of a cubic lattice, namely (2, 1, 0), originally proposed by Skolnick and Kolinski [11] is used (see Figure 2). Compared to the simple cubic lattice, it allows a wider range of backbone angles, from 64° to 143° between three contiguous alpha carbons. The number of first neighbors is also higher, 24, instead of 6. Side chains are discarded in the present simulation. Folding is produced by randomly selecting one amino acid, and submitting it to one of two available moves: end move for the N or C terminal positions, or corner move otherwise. Crankshaft move is no longer permitted with the (2, 1, 0) lattice. The new position can be occupied if it was previously empty, and the energy of the new conformation is computed by means of a statistical potential of mean force taken from the literature [12]. The Metropolis criterion is applied to accept or reject the new conformation.

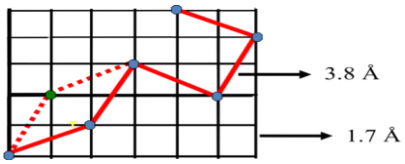


Figure 2: Details of the (2, 1, 0) lattice, with respect to the underlying cubic lattice. The dotted line indicates a possible move to a free node.

The process is stopped when roughly 10^6 to 10^7 Monte Carlo steps are reached, depending on the length of the query sequence. The full process is repeated 100 times, starting from 100 different initial conformations. The number of non-covalent neighbors (NCN) is recorded during each complete simulation. Two non-covalently bound residues are considered to interact if the distance between their respective alpha carbons does not exceed the upper limit of 5.9Å. The mean NCN is calculated at the end of the process and for all the initial conformations. The distribution of NCN along the sequence presents maxima and minima. We paid most of our attention to the maxima because we were aiming at the prediction of the core contacting residues, expected to be crucial for the formation of secondary structures [13] and whose prediction allows to determine the fold [14]. Therefore a residue i is accepted as a MIR if NCN(i) is equal or higher than 6. It results that more than 90% of the MIRs are hydrophobic (one of the six amino acids FILMVWY).

It has been demonstrated that for each protein, residues identified as MIRs constitute a non-trivial subset of the hydrophobic residues. Among families of folds (several

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

domains per family, similar structure, potentially different functions, and very divergent sequences) MIR occupy equivalent positions in the multiple alignments. There-fore, among families, a small number of hydrophobic positions are conserved as hydrophobic. They are compulsory for the folding to occur; they are deeply buried. For these reasons it seems reasonable to question whether they constitute the folding nucleus of the various folds. The answer is positive, as proposed by the presently available studies. They concern a very small number of families, because experimental evidence of the folding nucleus is not obvious and can show strong biases. Demonstration has been extensively proposed on two complete families, the immunoglobulins (56 structures of divergent sequences) and flavodoxins (43 structures).

One limitation of the MIR algorithm was the number of MIRs identified (by the threshold), typically around 15% of the amino acids, while the rate of amino acids expected to belong to the folding nucleus lies roughly in the range 5 to 10% This limitation also relates to the overall sharp variation in the graph. The SMIR extension addresses these issues, and it uses a Pascal triangle method to give smooth results. We also adjust the maxima that are identified in the smoothed graph to nearby (within 3 residues) hydrophobic positions, based on the accepted precision of the algorithm [15]. This is coherent with the expected accuracy for protein residue contact prediction of the contact prediction session of the Critical Assessment of protein Structure Prediction (CASP) experiments [16]. Hence, we continue to identify minima with a threshold but validate the extrema against the amino acids.

3 Results

3.1 Section and subsection headings

We model a protein as a chain of evenly spaced Cα atoms placed on a lattice [1]. We define a lattice unit (lu) to be 1.7 Å. Hence, Cα atoms are connected by vectors of the form (2,1,0), these vectors are $5^{1/2}$ lu in length which corresponds to 3.8 Å - the mean distance between adjacent Cα atoms. This results in 24 immediate neighbor positions for each point in the lattice. This represents the intersection of a 4x4x4 segmented cube with a sphere of radius 3.8 Å ($5^{1/2}$ lu) as shown in Figure 3.

The model does not take into account the presence of side chains, therefore the required separation is modeled with the 3.8 Å mini-mum distance requirement. Based on chain geometry, we limit the angle between some Cαs at position i and i+2 by requiring the distance between them to be from 4.1 to 7.2 Å (or from $6^{1/2}$ to $18^{1/2}$ lu). This corresponds to angles from 66 ° to 143 °, which is closer to the real angles in alpha and beta conformations. This is illustrated in Figure 4 where a residue i is fixed at [0, 0, 0] and all 24 possible positions for residue i+1 are represented as black vectors. There is a choice of 23 possible vectors for residue i+2. For the sake of clarity, only one position [0, 1, 2] (the green and red vectors) is shown. Red vectors are those that violate the distance (angle) restriction.

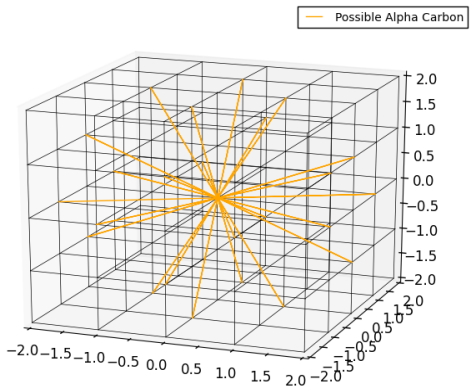


Figure 3: Vectors resulting from intersection of lattice with sphere at origin.

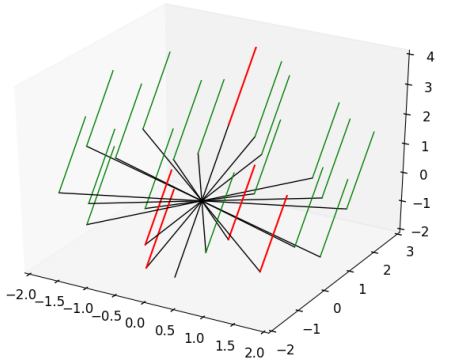


Figure 4: An example of angle restriction: vectors parallel or producing sharp corners thus violating the angle constraint are shown in red.

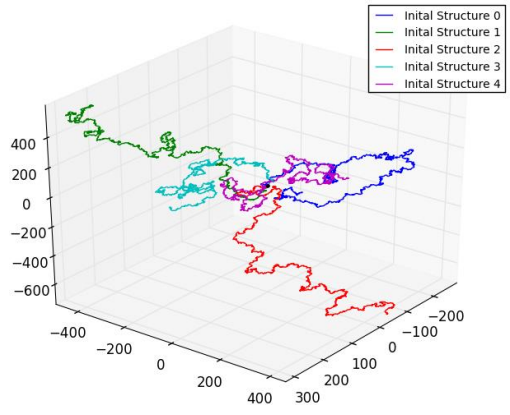


Figure 5: First five initial models.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

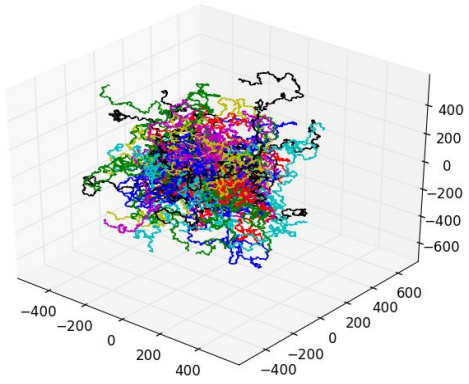


Figure 6: All initial models.

To initiate the simulations, 100 different starting models within this lattice are used. Figures 5 and 6 display respectively a sample of five and all models as a comprehensive plot. These models were computed randomly offline for chains of 1100 residues. For these models our only requirement is that they have some level of non-compactness [5]. Starting from the first residue located at position [0, 0, 0], the first n positions in the seed model will be used for an input model with n residues.

3.2 SMIR

The MIR method was first developed in 2004 [1, 5] and MIR 1.0 was first made available online as part of the RPBS server in 2005 [17]. The present SMIR server exploits MIR2.2 implemented with Fortran for server side simulation and a SMIR Javascript front end for interactive analysis. It has been found that the computation time for SMIR, once MIR results are available, is negligible on Intel Core 2 Duo based computers. The new SMIR smoothing method is implemented in Javascript with D3 [18] and has been primarily tested in Google Chrome 35. A browser based implementation allows users to retrieve this new analysis for any existing protein without the need to resubmit the entry to our submission server.

3.3 Submitting a protein

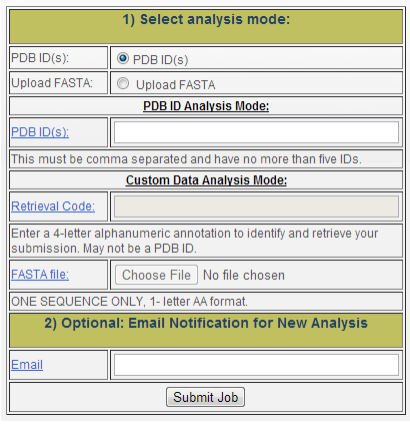


Figure 7: S/MIR Interface.

The interface in Figure 7 supports the submission of a PDB ID, a list of PDB IDs, or a FASTA file. In the latter case, the user will also enter a 4-letter alphanumeric code to identify the submission and later retrieve the results. The submission of an email address is optional. Should one be submitted, it will be only used for the purpose of informing the user of the availability of the results in the database with a reminder of the code. After submission, the server returns a SMIR status window (see Figure 8). Here the window displays the status for five proteins of PDB codes: 1AMM, 1DX5, 1I5I, 1QUC, 1ZAC. The top of the status windows lists the PDB ID(s) that have already been analyzed by MIR. Each different PDB is listed with a bullet point (e.g., 1AMM, 1DX5, 1I5I). If a protein has more than one chain, each available chain will be listed on that PDB's line and enclosed with parentheses (e.g., 1DX5(A), 1DX5(I), etc). The middle part of the window lists codes which are not valid retrieval PDB codes (e.g., 1QUC). The last part consists of the proteins that will be submitted to the server (e.g., 1ZAC). In this case, the PDB ID(s) will be added to the server queue for processing.

Figure 8: S/MIR status.

Each protein submitted to the server is displayed in the status window with the retrieval link to access the data once the execution is completed. If an email address was entered on the previous screen, a notification with a link will be sent upon completion. The proteins listed in the top of the SMIR status window are immediately viewable with a 2D graph (see Figure 9). If a PDB ID is not in the list of available proteins, it will be automatically submitted for analysis. Once the user's protein is ready to be analyzed, the server downloads the information associated with that PDB ID from the Protein Data Bank and runs MIR. After execution, the user may use the retrieval link or return to MIR query mode and enter that PDB ID to access the SMIR results. Additionally, the information that was generated for the new PDB ID is now available to other re-searchers for further use.

The graphical representation illustrated in Figure 9 is composed of three areas: legend for the MIR interface (top left), 2D display graph (top right) and data download (bottom). On recent browsers such as Chrome 24 or newer, the data can be downloaded with a CVS file. They can alternately be copied and pasted from a text box.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

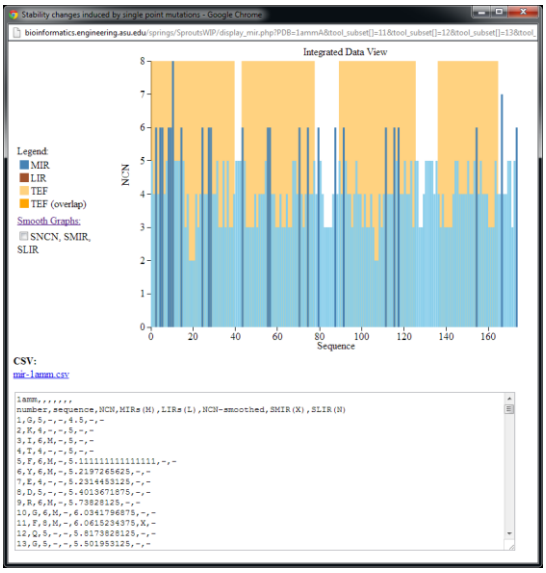


Figure 9: MIR Results.

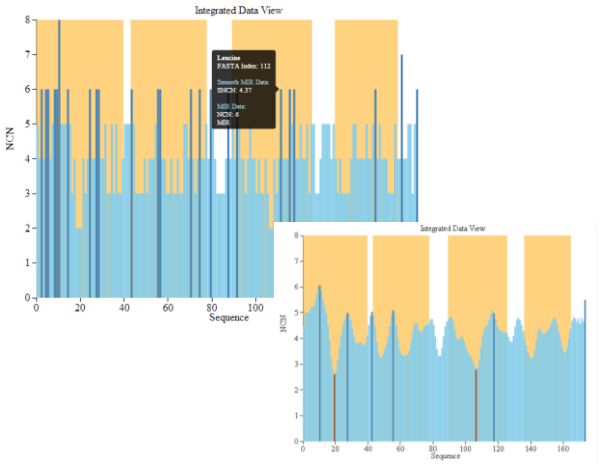


Figure 10: SMIR activated (dynamic window).

The 2D graph shown in Figure 10 displays MIR results in blue. Dark blue vertical bars indicate which residues are MIRs while dark red bars indicate LIRs (note that no LIR was shown in Figure 9). All bars plot NCN count at a position on the vertical axis. When browsing on the graph with the mouse, a black popup information box displays the amino acid name, its exact position in the protein (with respect to the FASTA file the protein is associated with), the number of NCN and the MIR status. The orange regions in the background indicate TEFs [1]. TEFs overlap on slightly darker orange areas. The SMIR method is activated with a checkbox. When SMIR is selected the 2D graph will show dynamically how MIR predictions (see top left of Figure 10) are replaced by SMIR predictions (see bottom right of Figure 10). When in smooth mode (i.e., when SMIR is selected), the dark blue and dark red bars indicate SMIRS and SLIRS (smoothed LIR, which are minima in the NCN curves) respectively.

3.4 Use Case and Discussion

We chose as an example, a case where the folding nucleus has been extensively studied. It is the TNf3 (PDB code 1ten), reported by the group of Jane Clarke [8]. This test case is interesting to tease the limits of our algorithm because a very small set of amino acids, four over roughly one hundred, is necessary to produce the so called Greek key topology of the native state. These four residues are highly conserved among the immunoglobulin superfamily, slightly less in the fibronectine type III superfamily. Φ-value analysis [19] have been experimentally determined, giving raise to the four amino acids forming the nucleus: L20, Y36, I59 and V70 [20]. MIR calculation illustrated in Figure 11 produces a high number of positions, 14 altogether.

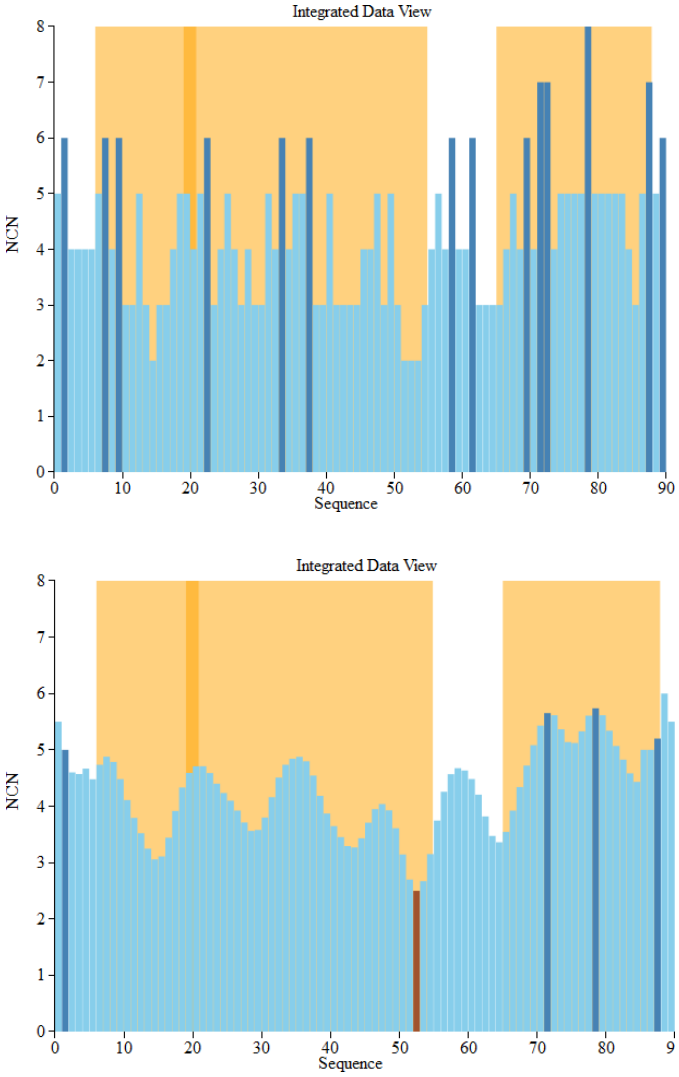


Figure 11: MIR (top) and SMIR (bottom) results for 1ten

The smoothing procedure proposed by SMIR gives a shorter list of four residues: L2, L72, M79, and F88. If one admit a window of ±2 AA, L72 is close to the list of residues involved experimentally in the folding. Although not determined to be SMIRs, L20, Y36, and I59 form maximums

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

in our smoothed results (refinement of SMIR thresholds may help in specific cases). This is an encouraging result for such a crude "toy model" using the only information of the sequence as an input. Actually, we do not believe the precision on positions can reasonably be more than ±2 AA.

4 Conclusions

Based on previous work [1, 5] we have presented the fundamental MIR algorithm and a method for increasing the readability and accuracy of residue interaction data. Our contribution over the previous MIR implementations is twofold: we have presented SMIR, an algorithm involving Pascal Triangle smoothing and hydrophobic residue analysis to calculate smoothed data. We have also implemented this algorithm in a new dynamic 2D graphical interface. Users may now view the smoothed MIR data for all proteins already existing in the SPROUTS database without needing to resubmit the protein for processing. These contributions refine the MIR technique so as to make MIR results more intuitive and useful to the scientific community.

One practical aspect of the prediction of MIR that can be important for wet biologists can be in the cases where they are faced with the production of inclusion bodies during the process of expression and purification. One of the ways used to circumvent this difficulty is to practice random mutations. The use of this server can be a suggestion not to mutate some positions suspected to be important for the structure, and consequently for the function, precisely the MIR. MIR and SMIR methods are also integrated in the SPROUTS workflow where they can be compared with stability analysis [21].

SMIR is hosted at the Université Paris Diderot on the server of the Ressource Parisienne en Bioinformatique Structurale (RPBS). RPBS provides scientists with a large range of resources devoted to the analysis of protein structure [17].

5 Acknowledgements

We acknowledge Pierre Tufféry for his help on using the RPBS resources, Dirk Stratmann for exciting discussions on benchmarks and method comparison and integration, and Elodie Duprat for shar-ing her results on the beta/gamma-crystallin superfamily. Mathieu Lonquety and Christophe Legendre contributed to the SPROUTS database where SMIR results are stored, and Fayez Hadji tested a preliminary version of the server. They are all thanked for their help. We also wish to acknowledge our collaborators at ASU: Rida Bazzi who is working with us on issues related to scientific workflow updates, Antonia Papandreou-Suppappola and Anna Malin who have worked on an alternative MIR method, and Banu Ozkan for evaluating SPROUTS functionalities and discussing future im-provement.

Funding: This work was partially supported by the National Science Foundation (grants IIS 0431174, IIS 0551444, IIS 0612273, IIS 0738906, IIS 0832551, IIS 0944126, and CNS 0849980) and by an invitation of the Université Pierre et Marie Curie.

Conflict of interest statement: Any opinion, finding, and conclusion or recommendation expressed in this material are those of the au-thors and do not necessarily reflect the views of the National Science Foundation.

6 References

[1] Chomilier, J., Lamarine, M., Mornon, J.P., Torres, J.H., Eliopoulos, E. and Papandre-ou, N. (2004) Analysis of fragments induced by simulated lattice protein fold-ing. Comptes Rendus Biologies, 327, 431-443.

[2] Berezovsky, I.N., Grosberg, A.Y. and Trifonov, E.N. (2000) Closed loops of nearly standard size: common basic element of protein structure. Febs Letters, 466, 283-286.

[3] Lamarine, M., Mornon, J.P., Berezovsky, I.N. and Chomilier, J. (2001) Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? Cellular and Molecular Life Sciences, 58, 492-498.

[4] Poupon, A. and Mornon, J.P. (1998) Populations of hydrophobic amino acids within protein globular domains: Identification of conserved "topohydrophobic" positions. Proteins-Structure Function and Genetics, 33, 329-342.

[5] Papandreou, N., Berezovsky, I.N., Lopes, A., Eliopoulos, E. and Chomilier, J. (2004) Universal positions in globular proteins - From observation to simulation. European Journal of Biochemistry, 271, 4762-4768.

[6] Bottini S., A Bernini, De Chiara, M, D Garlaschelli, O Spiga, M Dioguardi, E Van-nuccini, A Tramontano, N Niccolai 2013. ProCoCoA: a quantitative approach for analyzing protein core composition. Comput Biol Chem 43 :29-34.

[7] Itzhaki L.S., Otzen D.E., Fersht A.R. (1995) The structure of the transition state for folding of chmotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation condensation mechanism for protein folding, J. Mol. Biol. 25: 260-288.

[8] Lappalainen, I., Hurley, M.G. and Clarke, J. (2008) Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. Journal of Molecular Biology, 375, 547-559.

[9] Billings K., Best R., Rutherford T., Clake J. Crosstalk between the protein surface and hydrophobic core in a swapped fibronection type III domain, JMB 375 (2008) 560-571.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[10] Papandreou, N., Kanehisa, M. and Chomilier, J. (1998) Folding of the human protein FKBP. Lattice Monte-Carlo simulations. Comptes Rendus De L'Académie Des Sciences Série Iii-Sciences De La Vie-Life Sciences, 321, 835-843.

[11] Skolnick, J. and Kolinski, A. (1991) Dynamic Monte Carlo Simulations of a New Lattice Model of Globular Protein Folding, Structure and Dynamics. Journal of Molecular Biology, 221, 499-531.

[12] Miyazawa, S. and Jernigan, R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. Journal of Molecular Biology, 256, 623-644.

[13] Kister A., I. Gelfand (2009). Finding of residues crucial for supersecondary structure formation. PNAS 106: 18996-19000.

[14] Jones, D., Buchan, D., Cozzetto D., Ponti, M. (2012). PSICOV: precise structural contact prediction using spase inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28:184-190.

[15] Chomilier,J., Lonquety, M., Papandreou, N. and Berezovsky, I. (2006) Towards the prediction of residues involved in the folding nucleus of proteins. In Proc. DIMACS Workshop on Sequence, Structure and System Approaches to Predict Protein Function, May 3-5, 2006, Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) Center, Rutgers University. http://dimacs.rutgers.edu/Workshops/ProteinFunction/slides/chomilier.pdf

[16] Eickholt, J. and J. Cheng (2013). "A study and benchmark of DNcon: a method for protein residue contact prediction using deep networks." BMC Bioinformatics 14(Suppl): 512.

[17] Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., Renault, N., Wong, Y., Cantalloube, H., Chomilier, J. et al. (2005) RPBS: a web resource for structural bioinformatics. Nucleic Acids Research, 33, W44-W49.

[18] Bostock, M., Ogievetsky, V. and Heer, J. (2011) D-3: Data-Driven Documents. IEEE Transactions on Visualization and Computer Graphics, 17, 2301-2309.

[19] Fersht A. and Sato S. (2004) Φ-value analysis and the nature of protein folding transition states, Proceedings Natl. Acad. Sci. USA, 101: 7976-7981.

[20] Hamill S., Steward A., Clarke J. (2000) The folding of an immunoglobulin like Greek key protein is defined by a common core nucleus and regions constrained by to-pology, J. Mol. Biol., 297:165-178.

[21] Acuña, R., Lacroix, Z. and Chomilier, J. (2014) SPROUTS 2.0: a workflow to predict protein stability upon point mutation, submitted to ECCB 2014.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

WMS4HPC, a Workflow Management System bridge for High Performance Computing in life science data management

Etienne Z. Gnimpieba, , Carol M. Lushbough
Computer Science Department, University of South Dakota, 414 E. Clark St. Vermillion, SD 57069, USA,
{Etienne.gnimpieba; Nick.Weinandt; Carol.Lushbough}@usd.edu

Corresponding author: Etienne.gnimpieba@usd.edu, +1 605 223 0383.

~o~

Abstract

Summary: using High Performance Computing (HPC) resources in a workflow management systems (WMS) is a powerful way to address eScience challenges such as big data management. We present here WMS4HPC, a WMS connecting bridge with 4 degrees of freedom, for the integration and application of HPC cyberinfrastructure and local resources. The BioExtract Server, a Web-based WMS, has been enhanced to enable researchers to easily add their own analytic tools to the system through the iPlant collaborative infrastructure (IPC) allowing these tools to be executed on a HPC platform or a local system, shared with collaborators, and included in analytic workflows.

Availability and implementation: WMS4HPC was implemented in Java using RESTful API to connect Bioextract Server and iPlant cyberinfrastructure. BioExtract Server at bioextract.org is a free and open WMS with no login requirement. This implementation has been used for several peer review published workflow designs. Workflows such as RNA_Seq data analysis are available online on Bioextract.org workflow tab.

Keywords: Systems Biology**,** Bioinformatics, RESTful API, BioExtract Server, Data Integration, Systems Integration.

INTRODUCTION

Background: emerging developments in Big Data, Systems Biology, and Integrative Biology introduce an increasing number of challenges in life science research. The primary objectives of Workflow Management Systems (WMS) such as Bioextract Server are to simplify researchers' ability to access, apply, and share analytic tools, workflows and data [1]. Executing an analytic workflow on big data can be very difficult if the researcher's infrastructure cannot handle big data. High Performance Computing (HPC) cyberinfrastructures aim to assist researchers in handling big data and can improve a tool's execution performance (e.g. memory, run time)[2]. Additionally, it is not always optimal to execute small tools on an HPC cluster due to queue waiting times that degrade the Total Cost of Ownership (TCO). A goal in bioinformatics is to provide robust, efficient infrastructures combining the strengths of large HPC facilities and these local systems.

We focus here on the design of the WMS4HPC bridge that will handle powerful HPC resources and WMS features to facilitate the deployment and the application of analytic tools in big data management (extract, load, treat, store, analyze, share) such as RNA-Seq data analysis. By implementing it on Bioextract Server to leverage iPlant Collaborative we connected the researcher to a large community and tools repository for bioinformatics data analysis in diverse domains. And then contribute to the eScience challenges.

WMS4HPC, A WMS BRIDGE FOR HPC

Principle: WMS4HPC should provide 4 degrees of freedom, allowing the researcher to combine WMS features with local and community resources including tools, data, and HPC.

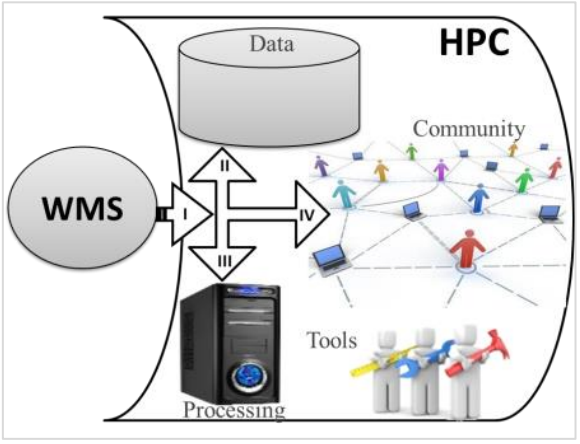


Fig.1. WMS4HPC mechanism in leveraging the HPC and the WMS with 4 degrees of freedom having extensions for frontier profile gateway (I), data space freedom (II), processing resources freedom (III), and community freedom (IV).

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

WMS4HPC consist of 4 degree connection (degrees of freedom) bridge which allows researchers to break principal walls (limits) between the problems, the workflow design process, the WMS features, and the HPC powerful resources using PaaS (Platform as a Service) philosophy. These degrees of freedom are related to ***frontier, data space, processing resources, and community*** (**Fig.1.).**

Degree I: free workspace frontier, user integration (federation, synchronization, workspace collaboration)

Frontier management is very complex due to security issues, conflicts in resources, redundancy, and personal or custom considerations (population immigration and emigration policy issues). In order to avoid these issues, we propose to a user who works on a local workspace to open a small secure gate (controlled through the WMS) by federating his HPC workspace from the WMS connection. This approach allows the user to have access to the HPC workspace and any HPC foreign user needs authorization (visa) to access the owner's local WMS workspace. Once the frontier is crossed, you have the ability (but are not required) to use all HPC powerful resources such as big data, processing, memory, security, community or networking.

Degree II: free data space, or data as a service (DaaS)

Workflow management system such as Bioextract.org has limitations in data management, especially the ability to handle big data. The distributive data repository system features offered by several WMS would benefit by extending their access to big data repositories and management systems provided by HPC. *Free data space* degree of freedom proposes then to extend user data management ability to use fast distributed data transfer protocols offered by HPC (e.g. iRODS in ICP cyberinfrastructure).

Degree III: free processing resources, or internet of things

Systems resource performance improvement (memory, cpu) is important to handle big data or execute heavy or complex tools. *Degree III* follows the IoT (Internet of Thing) philosophy and proposes to extend processing resources by employing a common resource sharing algorithm [3]. The main consideration here is the sequential dependency of tasks in the workflow. If some tasks can be executed in parallel, a significant performance improvement is possible.

Degree IV: free community, opening experimental results to others researchers through WMS features for scalability, and get involved in the community based research.

The WMS (e.g. Bioextract.org) allows researchers to share their workflows, results, and datasets with selected people or in the entire WMS community. The *Degree IV* allows the researcher to share resources (tools, data, workflow, problem, etc.) or results with the HPC community. Using semantic web tools such as ontologies and web services, we implemented a secured link to connect the WMS local workspace with the HPC large community (e.g. over 100 000 researchers in iPlant Collaborative infrastructure).

Performance and reliability

WMS4HPC is designed to integrate the main high quality criteria for good practices protocol in systems and data management and integration, even in life science and in computer science. It includes qualities such as: *easy use* (web-based integrated interface), *speed* (optimal resources), *reproducibility* (important for repeatability in published work), *scalability (*the WMS allows you to interconnect more than one OMICs level (genomics, transcriptomics, etc.), and the HPC allows you to cover large number of researchers), s*hareability* (improve your research networking by sharing with groups or the public), *long term reusability* (for publication, anyone can check and reuse it), *referential*(available online, your result, dataset, tools is identified and uniquely accessible).

APPLICATION

Bioextract Server

The BioExtract Server (bioextract.org) is a Web-based, distributed system designed to aid researchers in the analysis of life science data by providing a platform to facilitate the creation of bioinformatic workflows[1]. Scientific workflows are created within the system initially by recording the task sequence performed by the user. These tasks may include querying multiple data sources, saving query results as data extracts,

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

and executing distributed analytic tools. The series of recorded tasks can then optionally be saved as a reproducible workflow and is available for subsequent re-execution with the same or customized inputs and/or parameters. In the beginning, the BioExtract Server focused primarily on the management and analysis of genomics data. Subsequently, it has been expanded to include many life science domains. In order to handle the vast quantities of biological data generated by high-throughput experimental technologies, the BioExtract Server has leveraged iPlant Collaborative (IPC, www.iplantcollaborative.org)[2] functionality through their AGAVE REST API to help address big data storage and analysis performance issues in the bioinformatics field (e.g. RNA-Seq data analysis) [4]. Leveraging the IPC cyberinfrastructure has several key advantages: 1) it provides the ability to easily manage, analysis, and share big data, 2) it provides large scale computation support, 3) it is open for the community to contribute new analytic applications, and 4) it allows users to easily integrate their analytic tools with other popular application in the development of automated workflows[5].

The BioExtract Server has been in production for about five years and currently has approximately 500 active registered users, although researchers do not need to be registered to take advantage of much of the system's functionality. The primary enhancement to the system since 2011 is the integraton of the iPlant resouces through their AGAVE API. Additional enhancements include: the increase number of tools installed (from ≈ 100 to ≈300); additional data repositories (Biomart, Kegg, Brenda, Sabio-Rk); and domains covered for scalability analysis (from genomics sequence data to phenomics, proteomics, gene expression profiling, Systems Biology data)[6][7].

We applied WMS4HPC by leveraging the iPlant Collaborative cyber-infrastructure (HPC) on Bioextract Server (WMS). These two systems are freely accessible with a large bench of resources available for life science researchers. That application was supported by the AGAVE RESTful API and the Semantic Web annotation systems provided by iPlant collaborative [3][8]. At each degree of freedom in the WMS4HPC bridge, a java module and/or UI has been developed to hide the complexity of script code.

User interface implementation of WMS4HPC

Build in Java, the bridge is displayed to the user as friendly windows such as following figures.

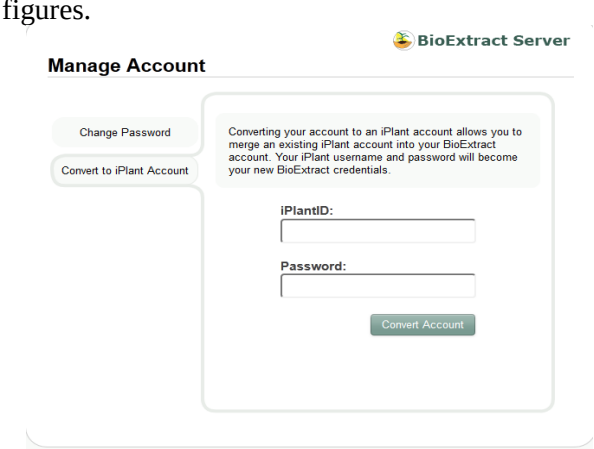


Fig.1. workspace frontier management for iPlant and Bioextract Server account interface with WMS4HPC

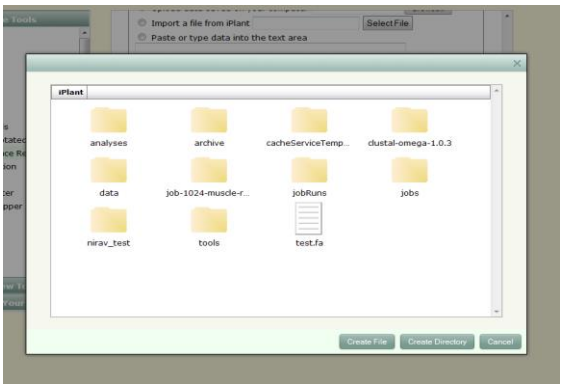


Fig.2. Example of iRODS (HPC) big dataset management interface through Bioextract with WMS4HPC.

Figure 2 show the implementation of the WMS4HPC data management system interface on bioextract server. This interface aims to avoid complexity in user script writing and make the ability to use HPC big data for workflow design. This implementation also facilitates data exploration and scientific discovery by integrating powerful, community-recommended software tools into a system that is robust enough to handle data while utilizing high performance computing resources like XSEDE (formerly known as TeraGrid) and others as needed to perform these tasks much more quickly.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

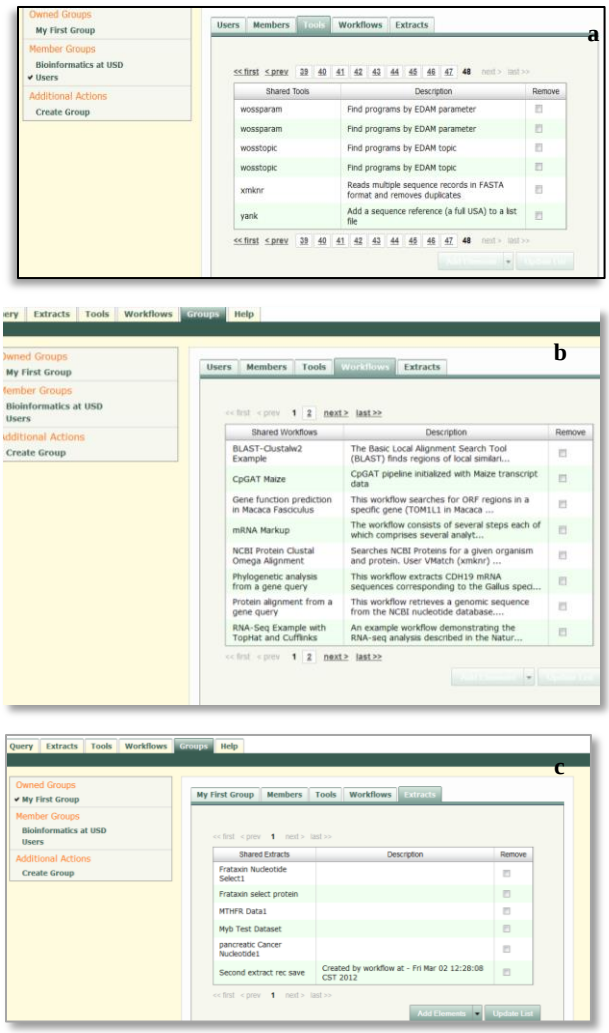


Fig.3. share bioextract tools(a), dataset (b) and workflow (c) within users

Figure 3 shows how user can easily share tools, datasets, and workflows to a selected group in Bioextract Server.

Deploy an analytic tool on iPlant

Develop or get an application which can be executed from the linux command line.

For example, if a java application is created which takes input through via the args array in the main method, it can be executed in the command line by typing:

java -jar ApplicationName.jar input1 input2

After the application is created, a file structure must be created to store the application and other required files.

Create the JSON File Describing Application.

In order for a user to deploy an application, they must submit a JSON file containing a description of the application they wish to deploy. The description must be in a format iPlant can process (below).

```
"inputs":[],
"parameters":[
   {
      "id":"tao",
      "details":{
         "label":"Type TAO ID",
         "description":"Type TAO Id",
         "visible":true
      },

      "semantics":{
         "ontology":[
            "xs:string"
         ]
      },
      "value": {
         "default": "",
         "required": true,
         "type":"string",
         "visible": true
      }
   }
],
```

The PhenoscapeQuery JSON file does not have any input files, but does have a parameter: The id of the parameter is important as it will be referenced by the wrapper script. The PhenoscapeQuery JSON file also contains the following line:

```
"modules":["purge", "load TACC", "load jdk32"],
```

The "load jdk32" is necessary as the PhenoscapeQuery application is a java application. "load jdk32" tells iPlant to have the java jdk ready before executing the PhenoscapeQuery application. Once the JSON file describing the application has been created, the wrapper and test scripts can be created.

Creating wrapper script.

The HPC cannot directly execute an application. iPlant instead relies on a wrapper script. The wrapper script uses the parameters and inputs described in the JSON file to execute the application as if it were directly executing it from the command line. The contents of the PhenoscapeQuery wrapper script are displayed below.

```
TAO="${tao}"
EXE=bin/PhenoScapeGeneQuery.jar
chmod +x $EXE
set -x
java -jar $EXE $Parameter
set +x
```

The TAO parameter is the id of the parameter

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

from the PhenoscapeQuery JSON file. EXE is the location of the application in the PhenoScapeQuery application folder.

```
EXE=bin/grninfer-linux.out

INPUT="${__infile}"

INPUT_F=$(basename ${INPUT})

iget -f  "${INPUT}"
chmod +x $EXE
set -x
$EXE $INPUT_F
set +x
```

The script parameter INPUT represents the input path described in the GRNInfer JSON file. The command

iget -f "{$INPUT}"

is one which iPlant uses to get the input file passed to the application and move that file to the location where the application is being executed. The test script can be left blank (but must still be present in the application folder). Once the wrapper and test scripts have been created and placed in the locally created file structure, the file structure can be uploaded to iPlant.

Uploading file structure to iPlant.

The file structure must be uploaded to iPlant before the iPlant is given JSON file describing the application a user wishes to deploy. iDrop can be used to move your locally created file structure to iPlant using the upload folder functionality.

Sharing an application with BioExtract:

cURL is a command line tool which can be used to move data to a url. If cURL is already installed on your machine, skip this step. Many linux operating systems have curl installed, but if your machine does not, it can be installed.

To Execute cURL Command in Command Line

The following command allows user1 to share their created tool with user2:

*curl -u "user1:pass1" -d "username=user2&permission=READ_EXECUTE" https://**$APIHOST**/apps-v1/apps/NameOfApp-AppNum/share*

Where **$APIHOST** is foundation.iplantcollaborative.org

Degree I: workspace on iPlant connect to Bioextract

As a guest, researchers can browse, search for data, access the Bioextract Server's public tools, and execute the public workflows. By registering, users have many more options available such as: saving query results, adding tools to their account, and creating, modifying, and sharing workflows with other users.

WMS4HPC implementation allow users to register with their iPlant Collaborative credential giving them access to their iPlant resources within the Bioextract Server, or synchronize with the existing Bioextract account. This is done through a secure protocol and nice UI integrated in the Bioextract login module.

Degree II: extension of your data repository on iPlant big data management system iRODS

Once the iPlant workspace is setup, a default big data store space is ready for usage. It contains two main parts. A file system for both application (or tool) deployment and dataset storage. Through Bioextract server, researchers can s*tore and use big datasets on iRODS using a data explorer window.* This interface aims to avoid complexity in user script writing and makes it easy to use HPC big data for workflow.

Degree III: use resources from anywhere

The implementation of our bridge provides the user the ability to deploy and use tools from iPlant or from their own local cluster. Then the researcher can choose to use local tools for small job workflow and iPlant tools deployment for big jobs in workflow design. That ability helps fix the problem regarding TCO in HPC usage. If the tool deployment on your local cluster is easy, some specificity for iPlant deployment is needed.

Deploying analytic tools on iPlant for workflow design on Bioextract.org

Integrating an application with the Bioextract Server involves storing the executable in a pre-described file structure, copying this file structure to iPlant, writing a formal application description in JSON, writing a bash script (which acts as a wrapper), and POSTing the JSON application description to the iPlant AGAVE API Apps endpoint. The JSON description includes information about parameters, a short text description of the application, and information

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

about which kinds of computing nodes are required for use (e.g. TACC cluster, iPlant local cluster, Atmosphere VM). Once iPlant's basic validation has taken place, the application will appear in the list of applications under the iPlant node on the Tools page of the Bioextract Server. At this point the application can be executed, shared, and integrated into an automated workflow.

Degree IV: share resources (data, tools), results, workflow, through the community (user, group or public)

WMS4HPC allows users to get involved in the community through Bioextract Server sharing features or iPlant's sharing ability with their large community. cURL is used to share tools and resources with the iPlant community. During the sharing process, the annotation of these tools assigns each tool to a specific domain using ontologies. The SSWWAP web semantic module integrated in iPlant plays a key role in the assignment process [9].

Examples use cases

Since the implementation of the WMS4HPC Bridge, over 70 tools have been deployed by users from different domains. Several workflows have been designed in the Bioextract Server using IPC HPC resources for peer review publications, including workflows related to population genetics, GWAS and phenomics (submitted), and RNA-Seq data analysis (that improve running time by 800% compare to the normal execution)[10]. A simple example of a workflow that incorporates an iPlant analytic tool can be found at bioextract.org on the Workflow tab. The workflow is "**NCBI Protein Clustal Omega Alignment**" with steps that include: 1) Query NCBI Protein for Arabidopsis Argonautes, 2) remove the duplicates using VMatch, and 3) perform a multiple sequence alignment using clustal omega (installed at iPlant).

CONCLUSION

Systems integration in life science research has become a complex challenge as data sets have grown. The ability to handle big data by HPC remains inaccessible for the common biologist. A WMS is the easiest way to hide the coding-level complexity of bioinformatics. WMS4HPC shows how a good integration of HPC and WMS could be possible and how to minimize the gap between life science researchers and these useful technologies. This work represents one more step in the improvement of the WMS philosophy shown on Bioextract.org. Our future work consists of integrating semantic (meaning) in the WMS workflow design process (free meaning degree of freedom). Then any resource can be annotated and intuitive workflows can be built using reasoning methods. To reach that step, every resource needs a performance evaluation (tools performer, data reliability, process quality). These reliability modules are under design to provide certified tools for output results value.

Funding: This work was supported by the National Science Foundation [IOS-1126481] Integrating the BioExtract Server with the iPlant Collaborative; and The National Institute of Health SD BRIN [P20GM103443].

Reference

[1] C. M. Lushbough, D. M. Jennewein, and V. P. Brendel, "The BioExtract Server: a web-based bioinformatic workflow platform.," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W528–32, Jul. 2011.

[2] S. A. Goff, M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, A. Muir, N. Merchant, S. Lowry, S. Mock, M. Helmke, A. Kubach, M. Narro, N. Hopkins, D. Micklos, U. Hilgert, M. Gonzales, C. Jordan, E. Skidmore, R. Dooley, J. Cazes, R. McLay, Z. Lu, S. Pasternak, L. Koesterke, W. H. Piel, R. Grene, C. Noutsos, K. Gendler, X. Feng, C. Tang, M. Lent, S.-J. Kim, K. Kvilekval, B. S. Manjunath, V. Tannen, A. Stamatakis, M. Sanderson, S. M. Welch, K. A. Cranston, P. Soltis, D. Soltis, B. O'Meara, C. Ane, T. Brutnell, D. J. Kleibenstein, J. W. White, J. Leebens-Mack, M. J. Donoghue, E. P. Spalding, T. J. Vision, C. R. Myers, D. Lowenthal, B. J. Enquist, B. Boyle, A. Akoglu, G. Andrews, S. Ram, D. Ware, L. Stein, and D. Stanzione, "The iPlant Collaborative: Cyberinfrastructure for Plant Biology.," *Front. Plant Sci.*, vol. 2, p. 34, Jan. 2011.

[3] R. S. H. Istepanian, S. Hu, N. Y. Philip, and A. Sungoor, "The potential of Internet of m-health Things 'm-IoT' for non-invasive glucose level sensing.," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2011, pp. 5264–6, Jan. 2011.

[4] R. Dooley, M. Vaughn, D. Stanzione, S. Terry, and E. Skidmore, "Software-as-a-Service: The iPlant Foundation API," in *5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers*, 2012.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[5] C. M. Lushbough, E. Z. Gnimpieba, and R. Dooley, "Life Science Data Analysis Workflow Development using the BioExtract Server leveraging the iPlant Collaborative Cyberinfrastructure," *Concurr. Comput. Pract. Exp.*, 2014.

[6] J. Fisher and T. A. Henzinger, "Executable cell biology.," *Nat. Biotechnol.*, vol. 25, no. 11, pp. 1239–49, Nov. 2007.

[7] E. Z. Gnimpieba, D. Eveillard, J.-L. Guéant, and A. Chango, "Using logic programming for modeling the one-carbon metabolism network to study the impact of folate deficiency on methylation processes.," *Mol. Biosyst.*, vol. 7, no. 8, pp. 2508–21, Aug. 2011.

[8] C. M. Lushbough, E. Gnimpieba, and R. Dooley, "BioExtract Server, a Web-based workflow enabling system, leveraging iPlant collaborative resources," in *2013 IEEE International Conference on Cluster Computing (CLUSTER)*, 2013, pp. 1–3.

[9] D. D. G. Gessler, G. S. Schiltz, G. D. May, S. Avraham, C. D. Town, D. Grant, and R. T. Nelson, "SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services.," *BMC Bioinformatics*, vol. 10, p. 309, Jan. 2009.

[10] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.," *Nat. Protoc.*, vol. 7, no. 3, pp. 562–78, Mar. 2012.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Wavelet Packet Based Diagnosis of Sleep Apnea using ECG Data

Syeda Quratulain Alir[1], **Varun Jeoti**[1], **and Samir Brahim Belhaouari**[2]
[1]Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, Malaysia
[2]Department of Mathmatics, Alfaisal University, Kingdom of Saudi Arabia

Abstract— *Sleep apnea is a nocturnal ailment related to breathing. The traditional diagnosis of sleep apnea using Polysomnography is quite inconvenient and expensive. This paper presents a technique to diagnose sleep apnea using only a single feature of ECG, i.e, RRI. Based on the physiological aspects of sleep apnea, a frequency band of interest (F_bOI) has been chosen from 0-0.125Hz. Discrete Wavelet Packet Transform (DWPT) has been deployed on RRI and a suitable decision variable is derived thereby helping segregate the normal and apnea cases. Testing and validation has been performed using MIT BIH online database [1]. The developed technique accurately diagnosed all the sleep apnea cases in MIT BIH sleep apnea database.*

Keywords: Sleep Apnea, ECG, RRI, DWPT

1. Introduction

Sleep Apnea is a nocturnal ailment. An episode of apnea is considered to occur when there is absence of breathing for ten seconds or more. In severe cases, single episode of apnea can last for more than a minute. According to American Academy of Sleep Medicine [2], a person is considered to have sleep apnea if he has more then 10 apnea events per hour. Although, it is roughly as common as asthma [3], it remains unrecognized from primary care physicians [4]. Clinical measures to diagnose sleep apnea involve nocturnal polysomnography of patient. The diagnosis requires at least one night of polysomnogaphic recording of patient's usual sleep hours. Clinical polysomnography involves Electrocardiogram (ECG), Electroencephalogram (EEG), Electromyogram (EMG), Electrooculography (EOG) and several other electrophysiologic measures [5]. A comprehensive study of the nocturnal activity is performed based on the signals achieved. Declaration of the person being normal or having sleep apnea is made after this study [5]. Polysomnography is the best method to accurately pin-point the sleep disorders in people. Although polysomnography owns a very high accuracy but it is very costly as it involves a number of expensive electro-physiological tests. Moreover, the associated fatigue of spending a whole night in sleep laboratory with a number of electrical probes connected to patient's body also demoralize the patients to investigate the sleep disorders they are facing.

Considering the demerits of the traditional diagnosis, research has begun on signal processing based diagnosis of sleep apnea [6]–[19]. Few physiological signals, someway or the other, get affected by disruption in breathing caused by sleep apnea. Several signal processing tools, like Fourier transform, wavelet transform etc., are in-use for the extraction of the discriminatory features in the signals mentioned above. Several researchers have deployed signals like EEG, SpO2 and ECG, both individually and in combination of each other for developing a signal processing based solution to sleep apnea diagnosis [6]–[19]. Although, some algorithms have shown 100% accuracy for diagnosing sleep apnea [12]–[15] but, issues like complexity, memory inefficiency, need of human intervention etc. have to be further explored. There is scope of improvement in the signal processing diagnostic algorithms for simplicity, time efficiency and robustness in diagnosis.

In this paper, a less complex and faster method for diagnosis of sleep apnea using ECG data from MIT-BIH online database [1] has been presented. Discrete wavelet packet transform (DWPT) is deployed on a specific narrow band of frequency in the spectrum of the RR interval of ECG, where the changes are expected to occur based on physiological aspects associated with sleep apnea. The frequency transformation and suitable test statistics generated from it are followed by a comprehensive statistical study to obtain the right threshold for segregation of normal and apnea cases. The performance is evaluated and validated in terms of probability of detection and probability of error, accuracy, specificity and sensitivity.

2. Materials and Methods

2.1 Study Group

As mentioned in the introduction, the ECG database provided by MIT-BIH online database [1] has been deployed for this research. In 2000, this database was created for sleep apnea challenge. The datasets are single channel ECGs that were extracted from polysomnographic recordings. The sampling rate is 100 Hz and average duration of recording is 8 hours. The online database has provided two data-sets, one for training and testing (data-set A) and the other for validation (data-set A_v). We have sub-divided data-set A in A_l and A_t for the training phase, whereby a threshold is achieved, and for testing of the developed technique, respectively. In order to observe the behavior of our developed technique in a diverse and richer database, it would be validated using A_v, which contains 35 unique cases of normal and apnea

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

patients, all shuffled up. Similarly, as in testing, the decisions are compared with the human annotations provided by MIT BIH online database.

2.2 Motivational Literature

It has been noticed that in case of sleep apnea RRI shows some low frequency components which are absent otherwise. M J Drinnan et. al [9] highlighted these low frequency components using Fourier Transform and have developed a technique that relies on the ratio of the spectral power between 0.01 and 0.05 cycles/beat and 0.005 and 0.01 cycles/beat to detect sleep apnea. Probably because of the less efficiency of the segregation scheme the overall technique could only yield around 90% accuracy. This technique motivated us to develop a uni-threshold segregation scheme that would be obtained after a detailed statistical study of the data under consideration.

2.3 The System Model

Fig. 1: The system model

As depicted in Fig. 1, the system model is based on three core phases, i.e., RRI-Pre-processing, designed technique and the developed decision block. In Fig. 1, $x(i)$ represents RRI data stream, $x_k(m)$ are the processed RRI data blocks, d_k refers to decision variable extracted by the technique phase and A/N represents the decision block output refereing a RRI data block as abnormal or normal.

Initially, in RRI pre-processing, the RRI data is prepared according to the need of the next phase. First of all, $x(i)$ is divided into N blocks, each block length M. The signal in k^{th} block is then represented by,

$$x_k(m) = x(i - kM), \tag{1}$$

where $0 \leqslant m \leqslant M-1$ and $0 \leqslant i \leqslant N-1$ and $0 \leqslant k \leqslant N/M - 1$. Later, $x_k(m)$ is corrected by removing outliers and false shoot up in $x_k(m)$. RRI often have missed or impractical magnitudes due to several reasons including equipment sensitivity and programming limitations. In order to eliminate the issue of false shoot up/shrink magnitudes, two thresholds (th_l=0.4 and th_u =2) have been defined for lower and upper cut-off limits. If a sample of $x_k(m)$, crosses the defined threshold, its value is adjusted by taking the mean of two previous and two later samples. The false values in $x_k(m)$ are replaced with the corrected values in $x_k(m)$, which is the output of pre-processing block.

Technique block is the backbone of the methodology as it provides the decision variable as its output while taking in $x_k(i)$ as its input. The schematic diagram of this block along with the sub-blocks is depicted in Fig. 2. The frequency domain analysis of RRI is considered as a tool to discriminate between normal and abnormal cases. Logical determination of frequency band of interest (F_bOI) is of core importance, as rest of the analysis totally depend on it. Band of interest (F_bOI) has been defined from $0 - 0.125$Hz. This selection has a physiological explanation as sleep apnea events consist of respiratory arrests lasting over 10 sec, including the awakening response after apnea. The minimum limit respiratory cessation is 10 sec which corresponds to frequency of 0.1Hz, while the longest apnea time usually observed lasts approximately 2 min (0.008 Hz) [20]. Therefore, SA-positive subjects are expected to have higher power in $0.008 - 0.1$Hz. In this research, we are performing short time analysis of ECG signal, whereby the whole night of ECG signal is divided in blocks of 5 minute intervals. Due to programming limitations the upper limit of 0.125Hz has been used. Given the presence of new frequencies in F_bOI during apnea event, the technique relies on microscopically studying F_bOI.

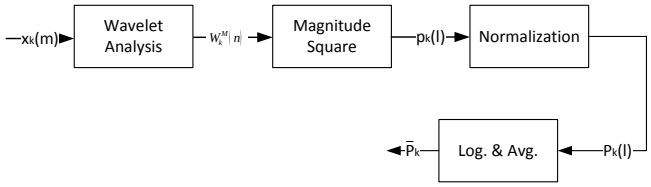


Fig. 2: The components of technique

The target of the wavelet transformation is to microscopically study the F_bOI. Wavelet multi-resolution analysis would be performed on $x_k(m)$ in order to achieve F_bOI and decomposition of F_bOI for frequency content analysis. Till the achievement of F_bOI, detailed coefficients are discarded to ensure minimal memory usage. However, both approximate and detailed components are used while decomposing and analyzing F_bOI. In order to fit into wavelet multi-resolution framework, $x_k(m)$ is considered to be the the approximate coefficients at scale $m = 0$ ($S_{0,n}$), defined by

$$S_{0,n} = x_k(m), \tag{2}$$

where n is the control parameter for translation in the wavelet transformation. In order to achieve maximum simplicity, Haar function has been deployed as base wavelet, which later proved to be sufficient enough to segregate normal and abnormal cases successfully. The general formulas for approximate($S_{m,n}$) and detailed coefficients ($T_{m,n}$) using Haar function are:

$$S_{m,n} = \frac{1}{\sqrt{2}}[S_{m,2n} + S_{m,2n+1}], \tag{3}$$

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

and

$$T_{m,n} = \frac{1}{\sqrt{2}}[S_{m,2n} - S_{m,2n+1}]. \quad (4)$$

The wavelet decomposition is performed using Eq. 3 and 4. From $m = 1$ to $m = 3$ detailed coefficients are discarded later till $m = 6$ detailed coefficients are further decomposed. The wavelet transform vector would be,

$$W_i^6 = (D_{i-7}, D_{i-6}, ..., D_i), \quad (5)$$

where i =12. W^6 depicts that the transformation vector has been achieved after 6 iterations. The achieved wavelet transform vector basically provides an insight to the defined F_bOI. Hence, our focus is to analyze and mark the differences in this vector for normal and abnormal cases. Now in order to obtain a decision variable from the achieved wavelet transform vector, few steps of mathematical manipulations have to be performed. First of all, for avoiding the value cancelation between positive and negative value while averaging, magnitude squared DWPT coefficients are calculated, as shown in Eq. 6

$$p_k(l) = |W_i^6(l)|^2 \quad (6)$$

where l is the length of vector. The magnitude squared DWPT coefficients are given by $p_k(l)$, where l represents the length of the vector.

Then, in order to avoid block to block variations in peak values, normalization has been done. Normalized $p_k(l)$ is given by $P_k(l)$,

$$P_k(l) = \frac{p_k(l)}{max(p_k(l))} \quad (7)$$

However, it is seen that the Apnea attributes in $P_k(l)$ are a small fraction of unity. In the proposed method, the relative emphasis of these small fraction coefficients is increased by adopting a logarithmic approach. We convert $P_k(l)$ to their respective decibel value $10log_{10}P_k(l)$. By doing so the peak value of unity becomes $0dB$ while all other smaller fractions are in a range from $-20dB$ to $-80dB$ whose range is far narrower than their linear counterpart in $P_k(l)$. Now an average of this set is taken - an average that is no longer skewed to unity if it was taken without logarithm.

Let this average be denoted by $\overline{P}_k$, and defined as:

$$\overline{P}_k = \frac{\sum_{n=1}^{l} 10\log_{10}p(l)}{l} \quad (8)$$

where k denotes the kth block. We assume that $\overline{P}_k$ carries the apnea attributes and can act as the suitable decision variable (DV).

3. Discussion on Decision Variable

If the underlying statistics of this DV is Gaussian under two cases of normal and apnea, it can be usefully used for detecting apnea. In order to validate this assumption, we now undertake a histogram analysis of $\overline{P}_k$ in two cases separately - one where $\overline{P}_k$ is for normal cases and another when it is for apnea cases. $\overline{P}_k$ is the output of the technique block which is the actual decision variable. In the next section, $\overline{P}_k$ would be used to represent the decision variable, instead of the assumed DV, d_k.

The role of decision block is to provide robust and reliable detection based on the DV ($\overline{P}_k$) provided by the developed technique. probability density function (PDF) of the $\overline{P}_k$, in both the apnea and normal cases. If the PDF were Gaussian, the P_{fa} and P_{md} could be easily determined. In this section, we will verify whether the $\overline{P}_k$ is indeed Gaussian or not. If not, whether it can still be used in the same manner. Towards this end, we take the ECG data from MIT online sleep apnea database. For the purpose here, we have deployed 50% data of A denoted by A_l.

In order to investigate the statistical trends followed by $\overline{P}_k$, a histogram is constructed using the value of $\overline{P}_k$ from data in A_l. The histogram achieved is shown in Fig. 3.

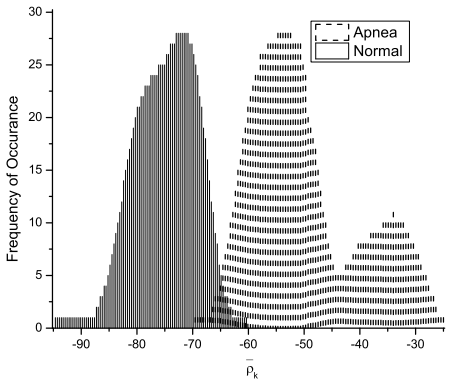


Fig. 3: Histogram of $\overline{P}_k$ values in apnea and normal cases for A_l

Fig. 3 shows an approximate Gaussian curve for normal case whereas a mixture Gaussian curve for apnea. The two regions are quite distinct from each other. The region of confusion is only over the area of overlap that is small. So if a threshold is chosen midway between the overlap region. The distinction can be made fairly accurately. Infact the tail of the curve in overlap region is quite Gaussian and can be used for estimating the error probability of miss detection and false alarm. In what follows we assume a mixture model for decision variable and obtain a best fit model for the same.

In order to evaluate threshold, *Th*, Gaussian curve fitting has been done on the histograms, as shown in Fig. 4. MAT-

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

LAB 2009a, curve fitting tool (cftool) has been deployed to curve fit the histogram. Individual characteristic equations have been obtained, for both cases.

$$f_n(\alpha) = \beta_1 \times e^{(-((\alpha-\gamma_1)/\zeta_1)^2)} + \beta_2 \times e^{(-((\alpha-\gamma_2)/\zeta_2)^2)}, \quad (9)$$

Eq. 9 represents the curve-fit equation for the normal case, where α, γ, ζ, and β denotes $\overline{P}_k$, mean, variance and relative strength of each curve in the gaussian mixture.

The available range of the coefficients for curve fitting is obtained whereby the 95% confidence bound values are utilized. For the normal case, summary of the available and chosen values of variables is mentioned in table 1.

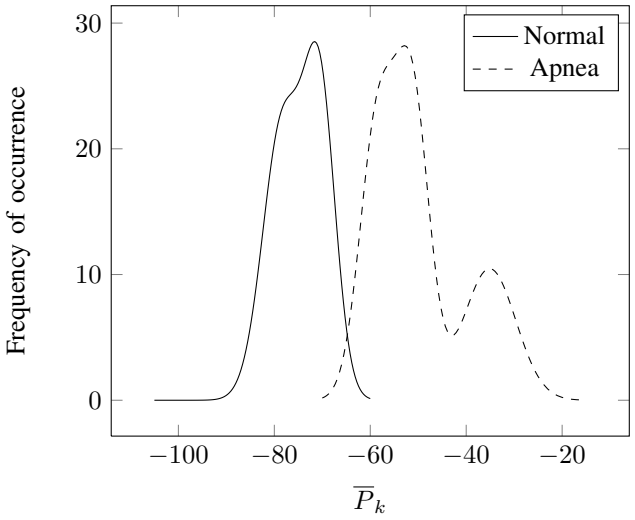


Fig. 4: Curve fitting of the data shown in Fig. 3

Table 1: Variable values for Abnormal case for 50% learning data

Variable	Chosen Value	Range
$\beta 1$	23.94	(22.68, 25.2)
$\gamma 1$	-70.53	(-70.73, -70.32)
$\zeta 1$	4.613	(4.451, 4.776)
$\beta 2$	21.89	(21.22; 22:57)
$\gamma 2$	-78:3	(-78:62,77.98)
$\zeta 2$	5.831	(5.557; 6.106)

$$\begin{aligned} f_a(\alpha) \quad = \quad & a_1 \times e^{(-((\alpha-b_1)/c_1)^2)} + a_2 \times e^{(-((\alpha-b_2)/c_2)^2)} \\ & + a_3 \times e^{(-((\alpha-b_3)/c_3)^2)} \end{aligned} \quad (10)$$

Eq. 10 represents the curve-fit equation for for apnea. For the apnea case, summary of the available and chosen values of variables, of Eq. 10, is mentioned in table 2, where α, b, c, and a denotes $\overline{P}_k$, mean, variance and relative strength of each curve in the gaussian mixture.

Table 2: Variable values for Abnormal case for 50% learning data

Variable	Chosen Value	Range
a1	23.47	(20.97; 25.96)
b1	-51.22	(-51.74,-50.7)
c1	4.932	(4.533, 5.33)
a2	21.91	(19.74, 24.08)
b2	-58.44	(-59.08,-57.81)
c2	5.27	(4.832, 5.708)
a3	10.47	(10.13, 10.81)
b3	-35.11	(-35.34,-34.89)
c3	7.715	(7.326, 8.103)

From Fig. 4, it is clear that although both the curves are not pure Gaussian but instead superposition of Gaussian functions. However, their tails in the confusion area are quite close to the gaussian tails. In the confusion area, an interesting phenomenon is observable whereby we notice that for normal curve sometimes,

$$|\alpha - \gamma_1| < |\alpha - \gamma_2| \quad (11)$$

Because of this relationship, $e^{-(\alpha-\gamma_2)/\zeta_2}$ in Eq. 9 would be increased so high as compared to $e^{-(\alpha-\gamma_1)/\zeta_1}$, that the other Gaussian equation becomes negligible and hence, in confusion area, Eq. 9 reduces to,

$$f_n(\alpha) = \beta_2 \times e^{(-((\alpha-\gamma_2)/\zeta_2)^2)} \quad (12)$$

While depending upon the values of γ_1 and γ_2 sometimes,

$$|\alpha - \gamma_2| < |\alpha - \gamma_1| \quad (13)$$

Hence, at some α, either of the curve would be negligible reducing the Eq. 9 to only one gaussian curve equation. Lets assume i denotes the curve with potentially higher value, hence,

$$f_n(\alpha) = \beta_i \times e^{(-((\alpha-\gamma_i)/\zeta_i)^2)} \quad (14)$$

On the similar grounds, the reduced equation in the confusion area for apnea curve can be written as:

$$f_a(\alpha) = a_i \times e^{(-((\alpha-b_{i1})/c_i)^2)} \quad (15)$$

From Eq. 14 and 15 it is obvious that in confusion area there exists only two gaussian curves. Accordingly, the maximum likelihood criteria can be used to maximize the detection of apnea. The optimum values of P_{fa} and P_{md} are thus uniquely obtained using the threshold, Th, as discussed in next section.

3.0.1 Threshold Extraction

In order to achieve the optimum threshold, a new function $H(\alpha)$ has been defined,

$$H(\alpha) = f_1(\alpha) - f_2(\alpha). \quad (16)$$

The optimum threshold that maximizes the detection of apnea is the intersection point of $f_1(\alpha)$ and $f_2(\alpha)$ i.e., when $\alpha = Th$, $H(Th) = 0$. *Th* has been obtained using the

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

well known Newton Raphson's Method with initial value $\alpha = -70$. By this method $th = -64.82$. Accuracy, specificity and sensitivity have been calculated for A_l.

4. Performance Evaluation

The performance of the developed technique is evaluated using probability of false alarm (P_{fa}), probability of miss detection (P_{md}), probability of detection (P_d), probability of error (P_e), accuracy, sensitivity and specificity. These measures were evaluated using confusion matrix parameters, as mentioned in table 3.

Table 3: Confusion Matrix

	Predicted Positive (P)	Predicted Negative (N)
Positive	TP	FN
Negative	FP	TN
Positive=SA	Negative=Normal	F=False, T=True

The probability of false alarm can be evaluated using the parameters of confusion matrix as follows.

$$P_{fa} = \lim_{N\to\infty} \frac{FP}{TN}, \tag{17}$$

where N denotes number of experiments. This equation represents the ratio of false alarm with respect to the correctly determined negative cases.

The probability of miss-detection can be evaluated using the parameters of confusion matrix as follows.

$$P_{md} = \lim_{N\to\infty} \frac{FN}{TP}, \tag{18}$$

where *N* denotes number of experiments. This equation represents the ratio of miss-detection with respect to the correctly determined positive cases.

The error introduced in our case is either a false alarm or a miss-detection, hence the net probability of error would be the sum the probabilities of these two entities. Mathematically,

$$P_e = P_{fa} + P_{md}. \tag{19}$$

The probablity of error free detection or simply the probability of detection can be determined as:

$$P_d = 1 - P_e. \tag{20}$$

Specificity indicates the ability of a classifier to detect negative cases, i.e. normal cases. It is calculated using Eq. 21.

$$Specificity = \frac{TN}{(TN + FP)} \times 100\% \tag{21}$$

Sensitivity represents the ability of a classifier to detect the positive cases, i.e. *SA* cases. It is calculated using Eq. 22.

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100\% \tag{22}$$

Accuracy represents the overall performance of a classifier. It indicates the percentage of correctly classified positive and negative cases from the total number of cases. It is calculated using Eq. 23.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100\% \tag{23}$$

5. Results & Discussion

5.1 Summary of Achieved Results

This section highlights the values of performance evaluation measures evaluated from the tables presented in section 5.2 with the help of the definitions discussed in section 4.

Table 4: Cases based performance evaluation using A_t

Evaluation Measure	Value
Specificity	5/5
Sensitivity	12/12
Accuracy	18/18

Table 5: Cases based performance evaluation using A_v

Evaluation Measure	Value
Specificity	10/10
Sensitivity	25/25
Accuracy	35/35

Table 6: Block based performance evaluation using A_v

Evaluation Measure	Value
Specificity	99.08%
Sensitivity	98.59%
Accuracy	98.89%
P_{fa}	0.006
P_{md}	0.017
P_e	0.023
P_d	0.977

Table 7: Block based performance evaluation using A_t

Evaluation Measure	Value
Specificity	99.34%
Sensitivity	98.24%
Accuracy	98.93%
P_{fa}	0.01
P_{md}	0.014
P_e	0.024
P_d	0.976

It is observed that the value of P_{fa} is decreased while calculated for A_v. It is because of the fact that A_v is a larger database and so owns higher number of normal blocks than that of A_t. Hence TN cases increases, subsequently decreasing the P_{fa}.

5.2 Discussion

In order to test the developed method for block based detection, ECG data stream is divided in 5 minute data blocks. The status of each minute, as per annotated by human experts in the database, is checked and if there is a single apnea minute in a block the whole block is marked as apnea. An annotation array, Arr_a, is created for storing the status of all the blocks.

Initially, the developed scheme is tested using data set A_t. There are a total of 18 cases in A_t, out of these 18 cases, 10 are abnormal, 5 are normal and 3 are border line. ECG signal from each case is taken one by one that later passes through our model, as described in section 2.3 and final decision for each block is determined as either normal or abnormal (for making our diagnosis more sensitive, boarder line cases has also been merged up into abnormal cases after their due classification). A decision array Arr_d stores the status for

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

each block as decided by the developed scheme. At the end a comparison is made between Arr_a and Arr_d in order to acquire the accuracy status of developed scheme. The performance has been evaluated using probability measures along with accuracy, specificity and sensitivity, as discussed in section 4.

Later, in order to validate the consistency of results obtained through this technique, it was tested for a larger database Similar procedure was performed for the dataset A_v that contains a total of 35 cases, out of which 20 are apnea, 10 are normal and 5 are borderline. Performance was evaluated as a comparison with that of the testing data.

As summarized in 5.1, the proposed scheme segregate apnea patients from normal people with 100% accuracy. The scheme has first been tested for 18 different data sets followed by blind validation of 35 more cases.

Chazel et. al. [15], have deployed multiple techniques while using both the RRI and EDR signals, in order to achieve 100% accurate classification of sleep apnea patients. In contrast our method uses only the RRI feature of ECG and used single DWPT based technique to achieve to same accuracy. In 2008, Ahsan et. al., [12] have discovered a method for 100% classification accuracy of sleep apnea patients using Wavelet transform. But this technique also deploys two features of ECG signal, i.e., RRI and EDR. Study of two features needs more computation as compared to our technique where the analysis of only one feature is required. Using Spectrogram, McNames et. al. [13], have achieved 100% accuracy but they have also used two features i.e., RRI and S-amplitude of ECG wave. Jarvis et. al. [14], have presented a spectrogram based solution but the threshold of the technique is not standard, the whole data has to be analyzed, a new threshold is produced and a decision can only be made after the analysis of whole data. On the other hand we have trained our classifier based on the statistics of learning data and a standard robust threshold is achieved. Our proposed scheme makes decisions based on five minutes data blocks and after the analysis of 12 blocks (1 hour) a tentative decision can be made that may further be strengthen by studying more blocks.

6. Conclusion

The main targets of this study were development of an ECG based sleep apnea detection algorithm that is fast and simpler as compared to other methods in literature. Most of the available schemes mentioned in literature are complicated either because of more number of discriminatory features or deployment of more then one technique.

The scheme proposed in this research focuses on the least number of features to be incorporated. Only a single feature, RRI has been used and a single decision variable $\overline{P}_k$ is selected. This unitary feature and discriminatory variable selection has made the algorithm quite simple and memory efficient.

Moreover, during the wavelet analysis, only F_bOI is focused and till the achievement of F_bOI, detailed components have been discarded, hence conserving the memory. This technique not only helped in conserving the memory but it also increased the accuracy as precisely the activity during apnea event is focused and redundant detail is discarded. Hence, all the analysis iteration process is performed on a small frequency band instead of utilizing all of the involved frequencies. Moreover, the proposed method takes (on average) 0.70 minutes to diagnose a case as normal or abnormal (using MATLAB 20008b on Intel Core i3).

In a nutshell, a comprehensive statistical study has been performed using the learning database. The best fit applied to the probability density function validates that a simple binary maximum likelihood detection can be used for the purpose. Accordingly, a unique threshold is obtained that minimizes the decision error. Based on the threshold, sleep apnea is diagnosed in each block of every patient's record in the testing database. Overall classification of the patients is performed by using the AASM 1999 criteria on the diagnosed blocks -AASM 1999 criteria says that a person is said to suffer from sleep apnea if he has 5 or more episodes of sleep apnea per hour in whole night recording. Validation has been performed by using a richer data-base containing records of 35 people.The decisions of our scheme are compared with the human annotations provided at the MIT-BIH online database. Validation process revealed that this scheme has the 2 capability to accurately segregate the apnea and normal cases. Hence, the technique serves to minimize the chance of un-diagnosed sleep apnea with 100% accuracy while being very simple and fast.

References

[1] Physionet organization and computers in cardiology mit-bih online database (sleep apnea challenge 2000), MIT-Physionet, www.physionet.org, 2000.

[2] A. A. of Sleep Medicine Task Force, Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research., Sleep 22 (1999) 667–689.

[3] M. H. Kryger, A Women's guide to sleep disorders, McGrawHill, 2004.

[4] V. Hoffstein, S. Linde, No More Snoring, John Wiley & Sons, 1999.

[5] R. D. Cartwright, Treating Sleep Disorders: Principles and practice of behavioral sleep medicine, Wiley, 2003, chapter3: Sleep Apnea: A challenge for Behavioral Medicine.

[6] D. Liu, Z. Pang, S. R. Lloyd, A neural network method for detection of obstructive sleep apnea and narcolepsy based on pupil size and eeg, IEEE Transactions on Neural Networks 19,2 (FEBRUARY 2008) 308–318.

[7] L. Almazaydeh, M. Faezipour, K. Elleithy, A neural network system for detection of obstructive sleep apnea through spo2 signal features, International Journal of Advanced Computer Science and Applications 3(5) (2012) 7–11.

[8] M. R. Azim, S. A. Haque, M. S. Amin, T. Latif, Analysis of eeg and emg signals for detection of sleep disordered breathing events, in: 6th International Conference on Electrical and Computer Engineering ICECE, Dhaka, Bangladesh, 2010.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[9] M. J. Drinnan, J. Allen, P. Langley, A. Murray, Detection of sleep apnoea from frequency analysis of heart rate variability, Computers in Cardiology 27 (2000) 259–262.
[10] C. Maier, M. Bauch, H. Dickhaus, Recognition and quantification of sleep apnea by analysis of heart rate variability parameters, Computers in Cardiology 27 (2000) 741–744.
[11] A. Hossen, Identification of obstructive sleep apnea from normalsubjects: Fft approaches wavelets, International Journel of Biometrices and Bioinformatics 4 (2012) 22–33.
[12] A. H. Khandoker, C. K. Karmakar, M. Palaniswami, Computers in Biology and medicine 39 (2009) 88–96.
[13] J. N. McNames, A. Fraser, Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram, Computers in Cardiology 27 (2000) 749–752.
[14] M. R. Jarvis, P. P. Mitra, Apnea patients characterizedby 0.02 hz peak in the multitaper spectrogram of electrocardiogram signals, Computers in Cardiology 27 (2000) 769–772.
[15] P. D. Chazel, C. Heneghan, E. Sheridan, Automatic classification of sleep apnea epoches using the electrocardiogram, Computers in Cardiology 27 (2000) 745–748.
[16] P. K. Stein, Domitrovich, Detecting osahs from patterns seen on heart-rate tachograms, Proceedings - Computers in Cardiology (2000) 271–274.
[17] J. E. Mietus, C. K. Peng, P. C. Ivanov, A. L. Goldberger, Detection of obstructive sleep apnea from cardiac interbeat interval time series, Computers in Cardiology 27 (2000) 753–756.
[18] M. Ballora, B. Pennycook, P. C. Ivanoc, A. Goldberger, L. Glass, Detection of obstructive sleep apnea through auditory display of heart rate variability, Computers in Cardiology 27 (2000) 739–740.
[19] C. W. Zywietz, V. V. Einem, B. Widoger, G. Joseph, Sleep apnea detection in single channel ecgs by analyzing heart rate dynamics, in: 2001 Proceedings of the 23rd Annual EMBS International Conference, October 25-28, Istanbul, Turkey, 2001.
[20] N. Oliver, F. Flores-Mangas, Healthgear: Automatic sleep apnea detection and monitoring with a mobile phone, Journal of Communications 2 (2).

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

BIOPHOTONIC ALGORITHM OF DNA BIOPHOTONIC LANGUAGE OF DNA

Dr. Boucherit Taieb, Yagoubi Abdelkader
BOUCHERIT Laboratory, Oran, Algeria.
07, road kaddoursalah houari Delmonte Oran, Algeria
Sponsored by **Dr.Abdelmalek Boudiaf, Minister of Health**, Algeria

Abstract -*We are aware that DNA has a very large capacity to store information's on the constitution of the individual as a higher capacity hard drive, billions of combinations formed by four amino acids are the basic archiving this storage but many questions remain unanswered at this time to science, namely the differentiation of cells, how the information comes to DNA, why bone cells is another heart, etc.. What is the phenomenon that differentiation Science is only concerned with the architectural structure of DNA sequencing*
That will shed new light on DNA and answers to real images.

1 Introduction

We have demonstrated in "WorldComp'13", that the DNA emits an electromagnetic field, *"paper ID: BIC 2373."*, *"Visualization of organs Electromagnetic field and DNA"*. The whole of Biophotons emitted by the DNA, form a coherent field carrying information as a Biophotonic radiation. This information which as holographic images form, this information is in the form of holographic images as I explained in my last publication that is a reminder that in contrast to the digital basic algorithm (0,1), however the Biophotonic algorithm of the brain and DNA use the base (colors, forms). The DNA emits an electromagnetic field, therefore coherent radiation that enables the communication between the DNA of all cells of the organ. This communication occurs in the form of holographic images contain all the information from the infinitely smallest to knowing subatomic.What will follow will bring the evidence by real images as well as the explanation of the photonic communication system of the DNA and the human brain architecture, and image quality.

2 Material &Methods

2.1 Material

The material is simple, it consists off has composite materials also all equipment off has laboratory off physics and chemistry, has computer & digital camera.

- Sensors.
- ChemicalMaterials.
- MaterialsPhysics.
- Materials composite.

2.2 Methods

- V.I.S System makes it possible to manufacture the organ in the compositematerials through their emitted energy.
- 1st step of manufacturing organ proceeds to taking photos from different angles of the composite and processing them by computer in the next step.

3 Theory & explanation

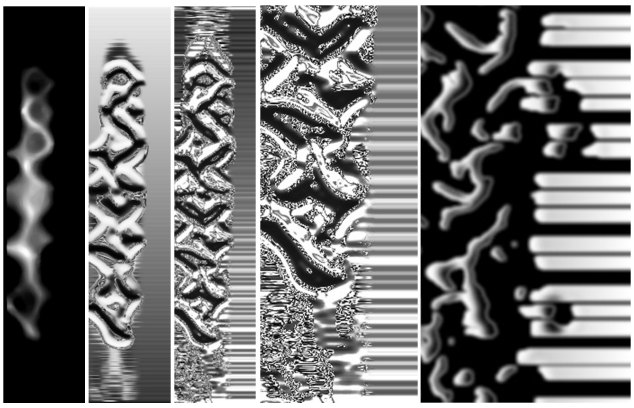
3.1 Theory

Any matter made of an assembly of molecules, which are also made of an assembly of atoms formed by electrons with negative charge which turn around a core consisting of protons positively charged & neutrons neutral charge.
DNA is made of an assembly of atoms; therefore a system electrically charged creating an electric field around it.
The structure of the DNA is made of two strands rolled up on them even forming a double helix, each one made of a formed on deoxyribose skeleton and phosphorus on which come to articulate bases nitrogenized to follow in a specific sequence, which is guanine with cytosine and the adenine with the thymine. This has long been known, and I would not talking over on this subject, therefore what is really interesting is the provision in length of DNA as well as the oval shape of ribosomes which answer us about the transmission antenna configuration & emission/reception of electromagnetic wave.
DNA is an electromagnetic antenna which receives and transmits information, an antenna as its elongated structure allow captures the electrical pulses, view from above its circular design and a little oval giving it the magnetic character.
DNA is an oscillator; DNA emits an electromagnetic field is a laser photons; it transmits a transmission electromagnetic wave and receives the emissions from other DNA.
Information of the emissions and the receptions is made of holographic images form, the image algorithm is not binary "0" and "1" but photonic; only as (colors & forms), therefore the holographic images which are read in real-time treated and dispatched to the other DNA informing about the actual state of the cells, therefore DNA As well as being a transmitting and receiving antenna, interprets the received information in its holograms form and archive them; is a super quantum computer.

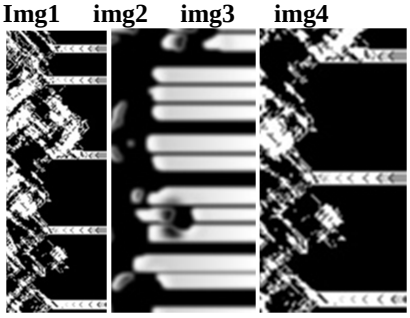
Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

3.2 Process & technical explanation:

The V.I.S System enables us to visualize the DNA in real holographic image and thus allows us to obtain it as photonic algorithm; the image is real and not digital. It also permits us to see all image characteristics by enlarging; we remark and you can check by yourself that, we enlarged the picture more and more, the result is appear more details, the phenomenon of pixelization doesn't exist.The images will become apparent demonstrate that in incontestable way.



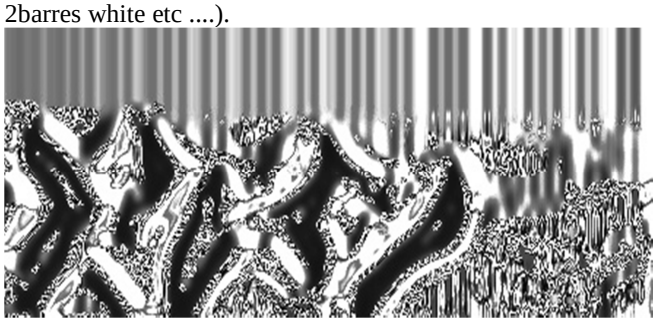
Img1 img2 img3 img4 img5



Img6 img7 img8

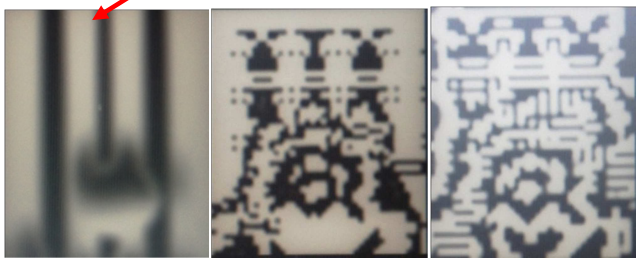
Img1 : real image of DNA double helix
Img2 : electromagnetic field of the DNA
Img3 : radiation of the DNA
Img4 : inferior part of the DNA with radiation
Img5 : enlarging lasing form & radiation
Img6 : reception phenomenon by DNA
Img7 : enlargement lasing
Img8 : enlargementt laser at the reception

I will now study the radiation and its nature, we can already make an important remark is that it is periodic (3barres white 2barres white etc ....).



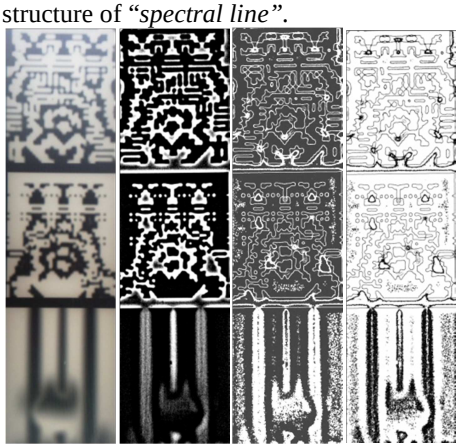
img9

Img10
Let us now study a single part



Img11 img12 img13

Img9: coherent nature of electromagnetic radiation
Img10: enlargement
Img11: expansion of a *"spectral line"*
Img12: internal structure of the *"spectral line"*
Img13: inverse image
We note that the V.I.S system lead us to show the internal structure of *"spectral line"*.

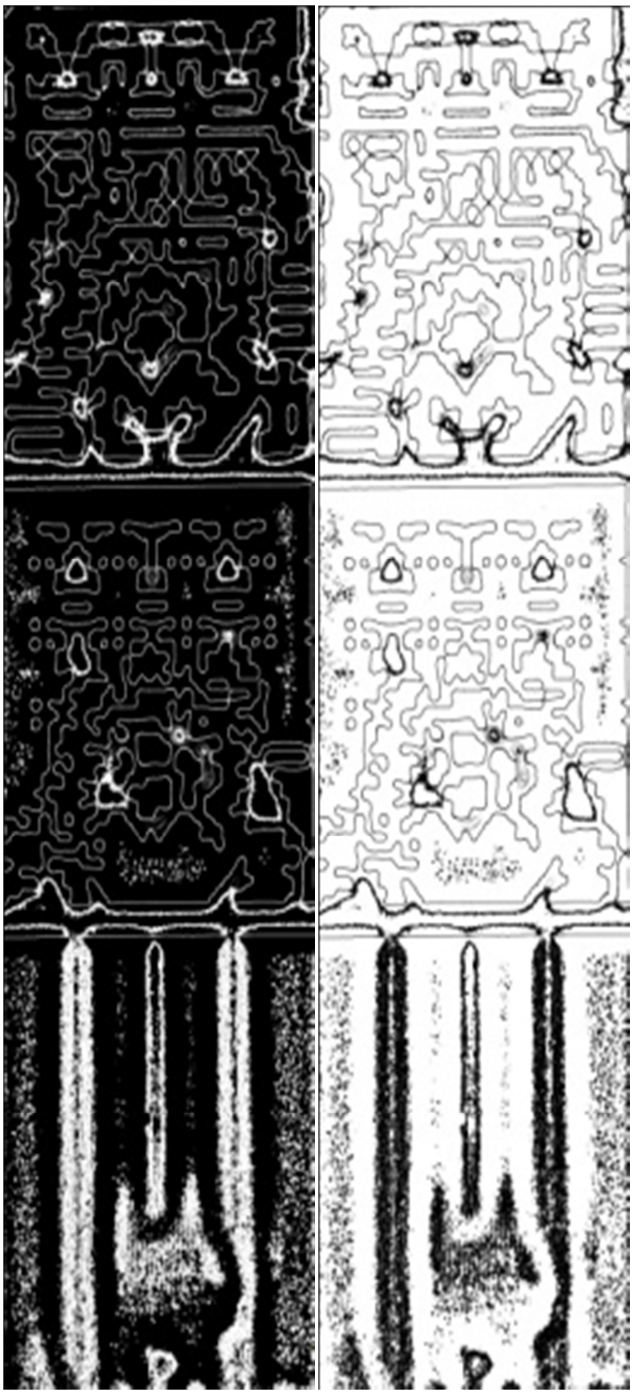


Img14 img15 img16 img17

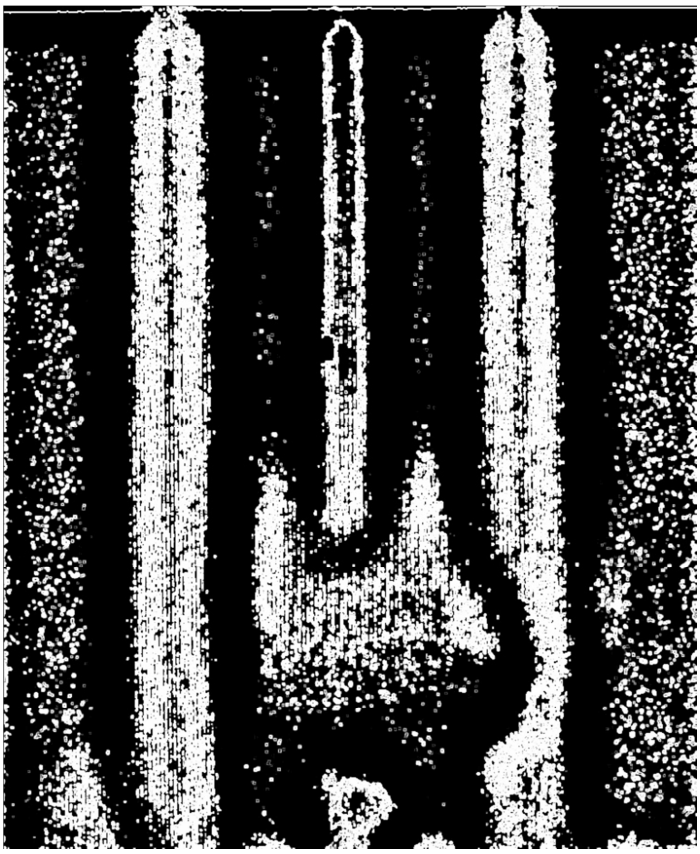
Img14: img11 img12 more img13 more on each other to visualize the line, its structure and its inverse.
Img14: first process of visualization by VIS system; smallest in structure
Img14: second process of visualization by VIS system
Img15: third process of visualization by VIS system
Img16: fourth process of visualization by VIS system
Img17: inverse image 16.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

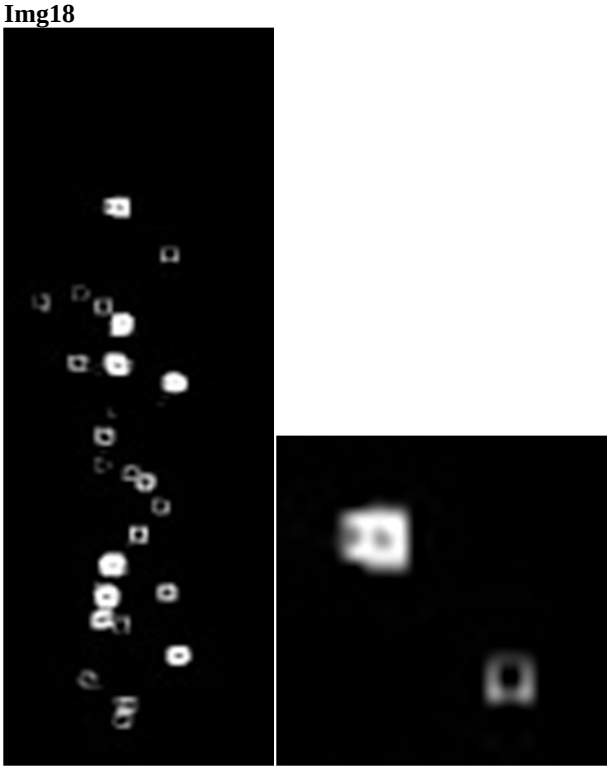
The process of visualizing the infinitely smallest is completed, The enlarging of image 16 and 17 shows us the astonishing & incredible images that no microscope in the world can show, with all details and especially it makes us understand clearly the functioning mode of DNA *"Biophotonic"* system.
At this stage, I respectfully ask you dear board of examiners & scientific committee to take the image 16 and a maximum enlargement and see the details of the images which you will see.



Img16 **img17**



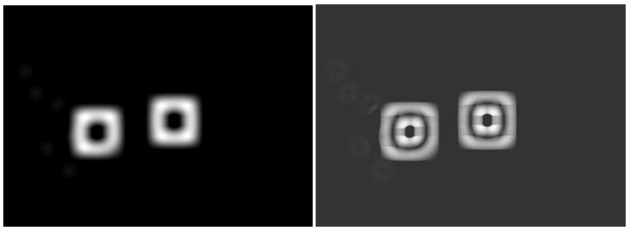
Img18



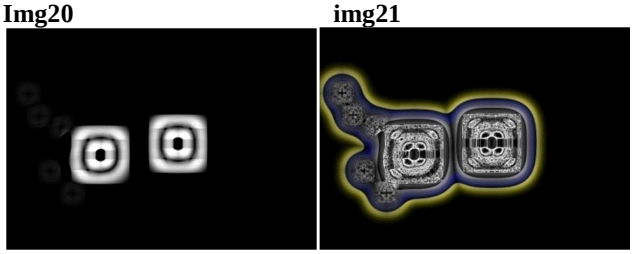
Img19 img20
Img19 :internal structure of the spectral line
Img20 :enlarging &visualization of three photons

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

I give the proofs that it is three photons, we will studying the imprints images of photons



Img20 **img21**

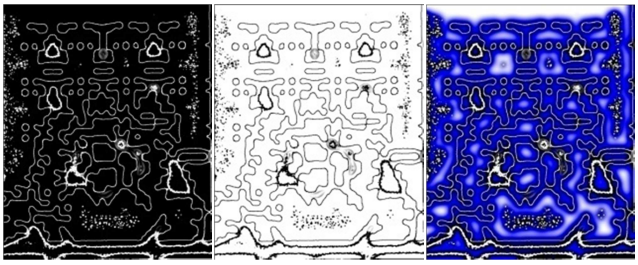


Img22 **img23**

The images "21-22-23" show us effectively that it is indeed, “the photons” and the transmission and reception of DNA is Biophotonics.

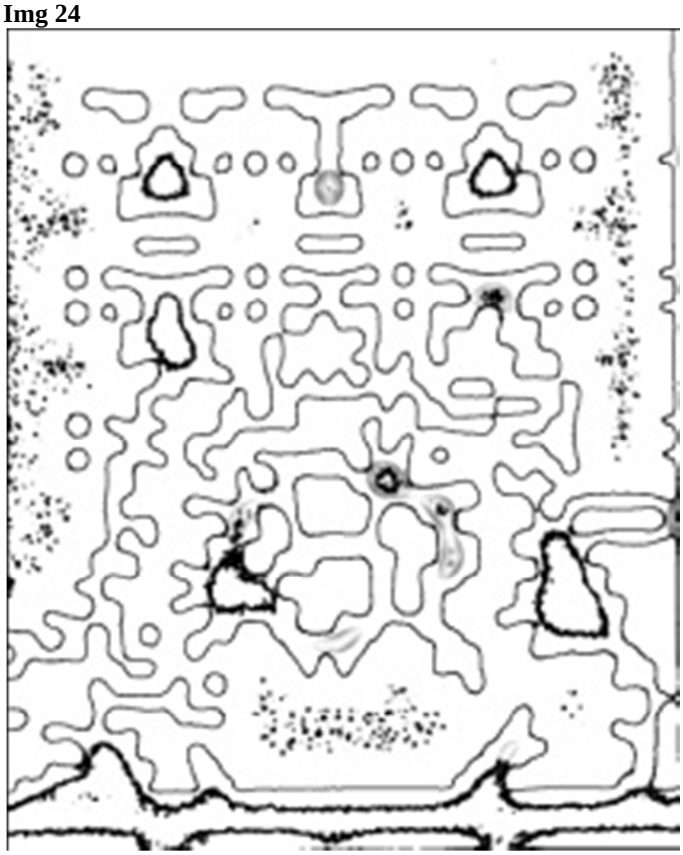
3.3 Schema and images of Biophotonics functioning

The DNA receives information about status of all body cells and broadcast the information’s about the state of appropriate cell by Biophotonics effect, each Biophoton save one or more information’s on the cell state, this phenomenon is done according to a precise diagram and a given circuit made of corridors, the photons make trajectories the hallways, during which they record information, which pass to get inside DNA where the information’s will received, treated and stored.
The VIS images show us perfectly clear this phenomenon, I point out that this theory and the images which demonstrate this theory are exposed for the first time in this publication.



Img24 **img25** **img26**

Img 24

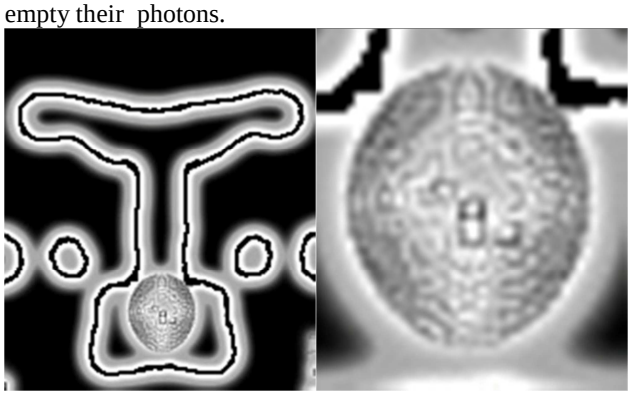


Img25

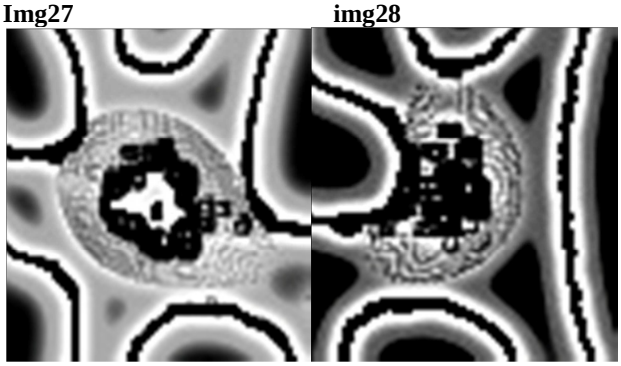
Img24:functioning informations diagram follow by photons
Img26 :inverse of img25

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

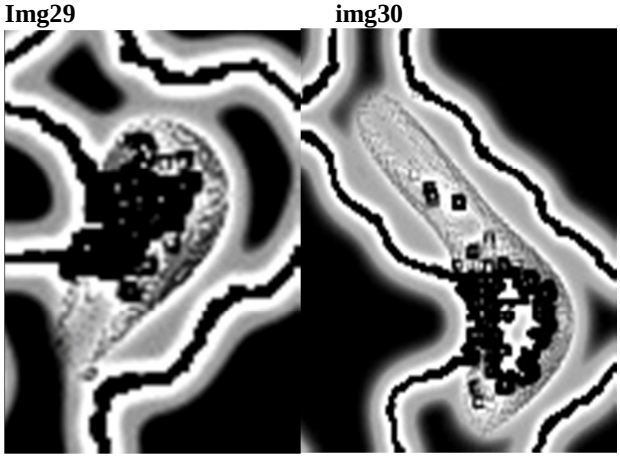
As I specified above the photons take stimulated corridors, and they attract with each other, forming a ball or Boson, each one is made of thousands of photons travel toward other DNA or to other body cells, this is the transmission of information by biophotonic effect.
Conversely, ie at reception, the coming of the spherical ball formed by thousands of photons take trace a stimulated corridors and stick themselves to their walls to empty itself, we notice that when the photonic Ball empty their photons, more it extends lengthens and lose his spherical form to become an oval form.
I show you below the images of photonic ball and images corridors and oval balls emptied photons balls.
I show below you the images of spherical photonic balls as well as images of corridors tracked, and the oval balls which empty their photons.



Img27 **img28**



Img29 **img30**



Img31 **img32**

Img27 : Photonic Ball& corridors
Img28 : Photonic Ball contain photonic beads
Img29 : Biophotonics Ball which empty their photons
Img30 : Biophotonics Ball& the track traced by the photon
Img32 : Photons are emptied, they traveling in a specific way, we note that biophotonics Ball became oval.

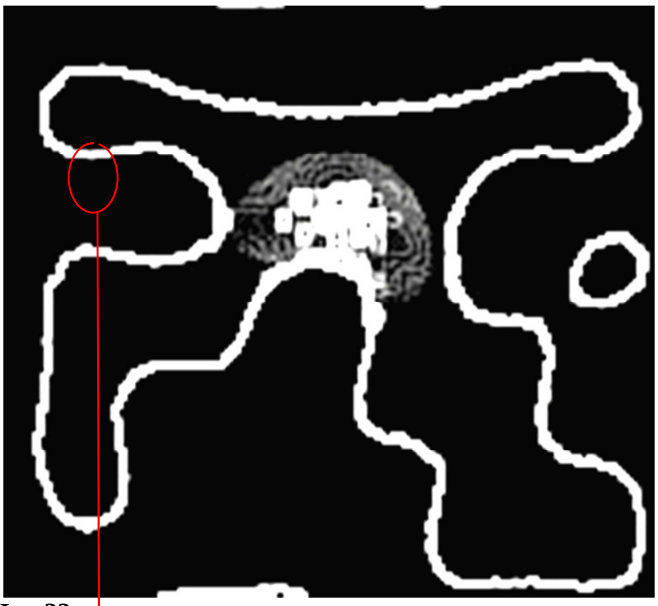
3.4 Proof of the Biophotonic effect

Images "05" "06" "07", and "08" show us conclusively that DNA broadcast and receives a transmission electromagnetic field with a specific frequency &period.
Images "17" "19" & "20" show us that is a photonic emission.

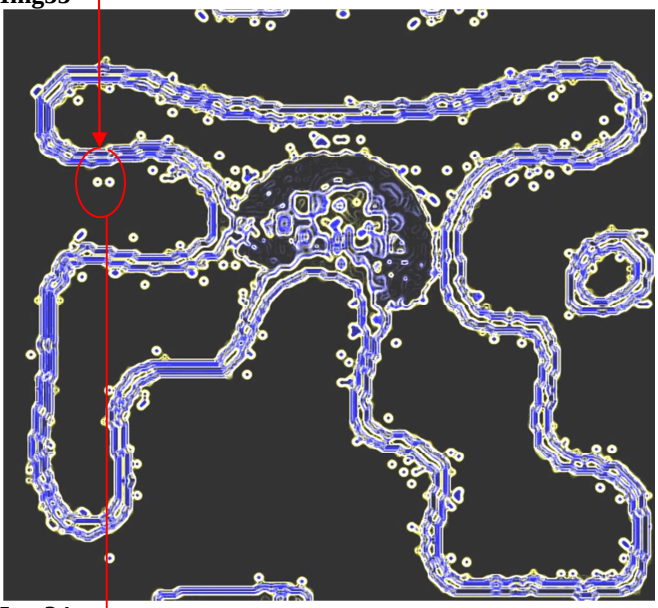
3.5 Biophotonic theory

The DNA with its structure and its helical form; become an antenna for transmission and reception of broadcasted electromagnetic waves, it looks strangely like the telephone relay antennas, or television broadcast station, that we see everywhere in our environment.
As any material consists of atoms which themselves are composed of protons and neutrons, forming the core which electrons rotating around, that enables us to say that matter therefore DNA is electrically charged, and as atoms are the center of perpetual movement, this implies that DNA is an oscillator.
The structure made of four amino-acids arranged in billion combinations is responsible for the storage and archiving DATA obtained from the received information's as it can be a library of information for transmission.
The received or transmitted Information are dispatched through Biophotons, each one owns in him even one or more information, the collection of this information is done by circuits or corridors which each Biophoton should follow to charge or discharge information.
DNA transmission with each other simultaneously done by "Biophotonic Balls".In fact, each Biophoton in emission phenomenon; cleave to other billions of Biophoton forming a "Biophotonic ball" moving towards others DNA. it come to Einstein theory which provides that light is made of grains traveling at the light speed as a "ball of light".
The VIS system visualizes the "ball of light" and how it is formed which are in reality the cluster of photons that will stick each to other forming the famous "ball of light", it shows to the smallest detail, their bursts showing the variation in their forms of spherical to elongated form vanish after when they are completely emptied of their photons.
Schematically like as the wagon train each Biophoton is a wagon which travel from station towards another station, each station are routed to an terminal, or they are removed from the locomotive for discharge their goods. They are subsequently loaded as related information each of others forming a "light ball" routed with a photonic speed towards another station where they are removed from each other to their discharged goods, this diagram is closest to simplify this phenomenon and rather just in reality.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America



Img33



Img34

Img35

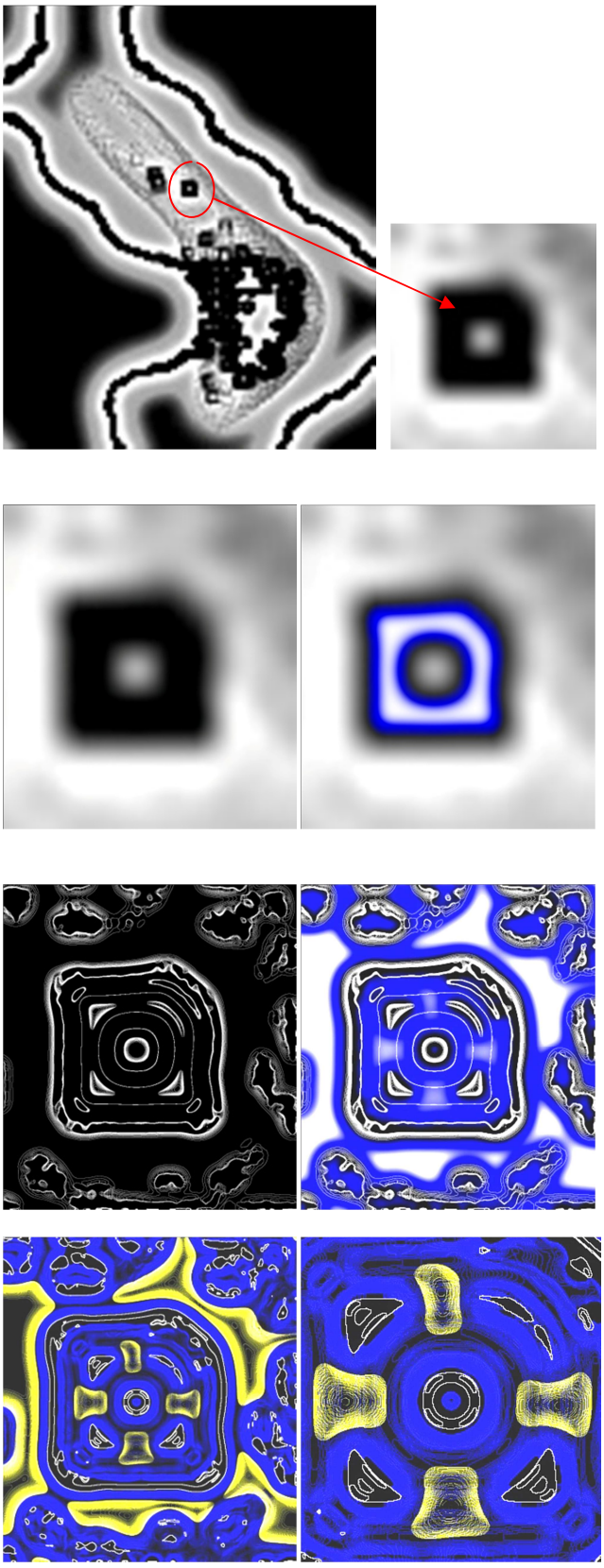
Img33 :Biophotnic Ball which is charged or discharged info
Img34 :corridors which the biophotons follow
Img35 :Biophotons

currently V.I.S system is the only one (system/tool) at the present time which makes it possible to give us images of the infinitely small with real precision, it enable us to see the invisible that exists in holographic images with photonic algorithm (colors, forms), the following images demonstrate that as below.

4 Conclusion :

The DNA with its helical structures which is transmission antenna & reception analyzes the received data and a result of their classifications in databases using combinations of four billion amino acids (Adenine-Thymine-Guanine-Cytosine).
DNA is a quantum computer, any material is formed of atoms which have a core made of protons having a positive charge and neutrons neutrally charged, and thus electrons which negatively charged rotate around this core.
We can say that the matter is electrically charged and comprises an electromagnetic field.
The DNA emits an electromagnetic field and also receives an electromagnetic field, the DNA mass is seat of very important vibration movement, thus it is an oscillator.
The cells communicate with each other via their DNA by electromagnetic field it is the source and the storage of these photons, the whole of these photons or more precisely Biophotons is a coherent field which carries information as images, these images correspond to a *photonic algorithm whose base is "colors and forms"* contrary to the numerical algorithm whose base is the “0 and 1”, therefore rather than receiving images as those which we knows, *DNA receive holograms images that each of them consists of an infinite number of images or " Data Bank image " which explains how come cell differentiation.*
I visualized electromagnetic radiation, and showed that this radiation is coherent and therefore an issue.
I let you the appreciation of the real images obtained by the VIS system which shown us the path followed by "Potonic-balls"; as well as the corridors which the biophotons must follow as they arrive at destination and starts to empty “photons ball,” the beauty of these images is unquestionable because they are real and no equipment also sophisticated in the world can realize.
Le VIS système est une nouvelle méthode permettant d'étudier par des images de l'exploration de l'infiniment petit avec une réelle exactitude défiant tout autre système d'exploitation, and allows us to compare images obtainedwithexistingtheories in particlephysicsdomain, medicine and manyotherdisciplines.
Other publications will be presented dealing in the elementary particles domaine.
i show you as below some pictures dealing the particle physics.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America



Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

BIOINFORMATIC STRUCTURE & BIOPHOTONIC ALGORITHM OF THE BRAIN

Dr. Boucherit Taieb, Yagoubi abdelkader IT technology engineer,
Lalam Abdelkhalek electronician developer, Boucherit mounir health student.
BOUCHERIT Laboratory, Oran, Algeria.
07, road kaddour salah houari Delmonte Oran, Algeria
Sponsored by **Dr. Abdelmalek Boudiaf, Minister of Health**, Algeria

Abstract –*The brain is an organ that Lodge in the cranial box. Its weight is abt 1.5 kg, it consists of abt 12 billion of nervous cells, 100 billion of neurons, and 120 trillion of connections, it makes arrangement of matters as the best performer, complex and impressive computer that exist actually. It can deal with millions of messages that is received each second, save them and sort out all kind of information. The brain controls the heart blood and digestive as well as respiratory systems and other systems at the same time enabling us to read and think.*

Information saved by human brain, can fill pages of several millions of books that is the equivalent of millions of university libraries. Human brain possesses lot of capacities to learn and think.

To understand how does this arrangement of neurons work, the inter connection of neurons is similar to the structure of chip that will enable us to build the structure of future mega-computers.

The algorithm with which functions the human brain is different from the digital algorithm of the computer, the first one uses forms and colors, while the second one uses 0 and 1. Our computers use chips with limited capacity of storing and treatment of information using electric energy, while the brain uses an electric and chemical energy to dispatch & send-off information.

1 Introduction

The human brain consists of about ten thousand millions of nervous cells, about hundred million of neuron and billions of connections. Each of us have the most sophisticated and the most performed computer that can exist. Logically this biological assembling is a source of inspiration that enables us to create future computers. In my humble opinion, what is really needed is to study it in the least details trying to understand how it functions and how is the whole structured; starting by:

- Nervous cells.
- Neurons.
- Connection and transmission modes of the images.
- Its ability to learn from its previous experience.
- Its ability to interpret the reactions in numerical tasks and model it technically.

as per my last publications at Worldcomp'11 « Vitreous Imaging System », Worldcomp'13 « Mass Micro Reconstruction ». I dutifully bring you an architectural technical model of the human brain. I let you the pleasure to judge the architecture of the brain and the quality of the obtained images from the data bank images or the memory images already listed in my previous publication.

2 Materials & Methods

Materials

The material is very simple, it consists of a composite materials also all equipment of a laboratory of physics and chemistry, a computer & digital camera.

- Sensors.
- Chemical Materials.
- Materials Physics.
- Composite Materials.

2.1 Methods

- The MMR2 make it possible to manufacture the organ in the composite materials through their emitted energy.
- The first step of manufacturing of the complete organ proceeds to taking photos from different angles of the composite and processing them by computer in the next step.

2.2 Theory & explanation

A neuron is a specialised cell in communication and information treatment; it is composed of cellular body;

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

ramification or dendrite that receive the information's from an axon which broadcasts the information; the axon is long enough with a diameter of 10 µm, with a synapsis that transmit the information from a cell to other cells, the signal which travels along axon to synaptic knob.
The function of the neuron is to circulate the information from neuron to others, i.e. it is between the organism and its internal environment and with itself.
These hundred billion of neurons constitute a complicated Network with hundred thousands of connections in the neuron. All is in the *Cranial Box* closing the encephalon.
The encephalon is composed of two cerebral hemispheres that are the brainstem and the cerebellum.
The encephalon has no direct contact with *Cranial Box,* it's protected by a set of thee sheets called the meninx floating a liquid called "*cephalorachidian liquid"*.
Both of the cerebral hemispheres are divided, each one into 04 zones: frontal, parietal, temporal and occipital zone.
Each zone is responsible of different precise functions.

Digital images coding:
How are the multimedia contents, particularly, the images coded in the computer? In the computer science each information «text, picture, sound…» is coded under a binary form which means 0 and 1. The smallest information unit is called *«bit» «binary digit»*, a set of 8 bit is called *«byte»*.
A *byte* enables to store a letter, a figure. This grouping of numbers by set of 8 enables the best legibility similar to what we appreciate on decimal base, to group the figures by three in order to distinguish the thousands, Eg: 1 256 245 is more legible than 123245.

How the information is coded in binary system?
For the figures operation is carried via a reconversion in base 2. A natural who he is a positive whole or nil. The number of figures that we want to use. With 1 byte it is possible to obtain 2 (= 2^1) value : 0 and 1

A natural number is a positive integer or zero. The number of bits to use depends on the range of numbers that you want to use.
With a bit, it is possible to get 2 (= 2^1) values: 0 and 1
With two bits it is possible to represent 4 (= 2^2) different values: 00, 01, 10 and 11
With a byte (8 bits), it is possible to represent 256 (= 2^8) values or the integers between 0 and 255
For a group of (n bits), it is possible to represent (= 2^n) values or integers from 0 (= 2^n)

So how can we count with 4 bits? With 24 bits?

The base (2) operate exactly as the base (10); except for exactly for its mesur unit. Ex : in base (10) « eleven » is written «11» either «$10^0 + 10^1$».

In base (2) «eleven» is written as «1011» either «$2^0 + 2^1 + 2^3$» ($1*2^0 + 0*2^1 + 1*2^2 + 1*2^3$) n base-2, « onze » s'écrit « 1011 » soit «$2^3 + 2^1 + 2^0$» ($1*2^3 + 0*2^2 + 1*2^1 + 1*2^0$) La valeur d'un octet est comprise entre 0 et 255.

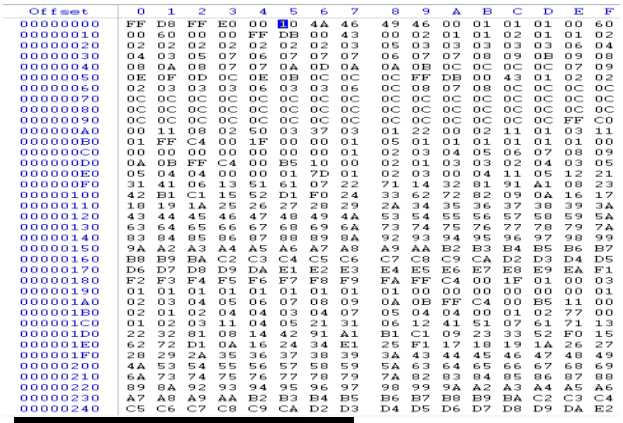
The picture coding :
Two categories of pictures coding

1. **Vectorial coding :** The picture is coded by a set of mathematic formulas.
2. **The Bitmap coding** : The image is encoded as point table

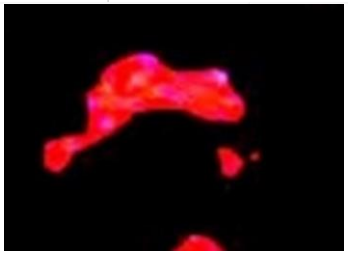
Vectorial image, bitmap image
Example: representation of a circle in vector or bitmap coding

2.3 Process & technical :

0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	1	0	0	0
0	0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0



Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00000000	FF	D8	FF	E0	00	10	4A	46	49	46	00	01	01	01	00	60
00000010	00	60	00	00	FF	DB	00	43	00	02	01	01	02	01	01	02
00000020	02	02	02	02	02	02	02	03	05	03	03	03	03	03	06	04
00000030	04	03	05	07	06	07	07	07	06	07	07	08	09	0B	09	08
00000040	08	0A	08	07	07	0A	0D	0A	0A	0B	0C	0C	0C	0C	07	09
00000050	0E	0F	0D	0C	0E	0B	0C	0C	0C	FF	DB	00	43	01	02	02
00000060	02	03	03	03	06	03	03	06	0C	08	07	08	0C	0C	0C	0C
00000070	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C
00000080	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C
00000090	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	0C	FF	C0
000000A0	00	11	08	02	50	03	37	03	01	22	00	02	11	01	03	11
000000B0	01	FF	C4	00	1F	00	00	01	05	01	01	01	01	01	01	00
000000C0	00	00	00	00	00	00	00	01	02	03	04	05	06	07	08	09
000000D0	0A	0B	FF	C4	00	B5	10	00	02	01	03	03	02	04	03	05
000000E0	05	04	04	00	00	01	7D	01	02	03	00	04	11	05	12	21
000000F0	31	41	06	13	51	61	07	22	71	14	32	81	91	A1	08	23
00000100	42	B1	C1	15	52	D1	F0	24	33	62	72	82	09	0A	16	17
00000110	18	19	1A	25	26	27	28	29	2A	34	35	36	37	38	39	3A
00000120	43	44	45	46	47	48	49	4A	53	54	55	56	57	58	59	5A
00000130	63	64	65	66	67	68	69	6A	73	74	75	76	77	78	79	7A
00000140	83	84	85	86	87	88	89	8A	92	93	94	95	96	97	98	99
00000150	9A	A2	A3	A4	A5	A6	A7	A8	A9	AA	B2	B3	B4	B5	B6	B7
00000160	B8	B9	BA	C2	C3	C4	C5	C6	C7	C8	C9	CA	D2	D3	D4	D5
00000170	D6	D7	D8	D9	DA	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	F1
00000180	F2	F3	F4	F5	F6	F7	F8	F9	FA	FF	C4	00	1F	01	00	03
00000190	01	01	01	01	01	01	01	01	01	00	00	00	00	00	00	01
000001A0	02	03	04	05	06	07	08	09	0A	0B	FF	C4	00	B5	11	00
000001B0	02	01	02	04	04	03	04	07	05	04	04	00	01	02	77	00
000001C0	01	02	03	11	04	05	21	31	06	12	41	51	07	61	71	13
000001D0	22	32	81	08	14	42	91	A1	B1	C1	09	23	33	52	F0	15
000001E0	62	72	D1	0A	16	24	34	E1	25	F1	17	18	19	1A	26	27
000001F0	28	29	2A	35	36	37	38	39	3A	43	44	45	46	47	48	49
00000200	4A	53	54	55	56	57	58	59	5A	63	64	65	66	67	68	69
00000210	6A	73	74	75	76	77	78	79	7A	82	83	84	85	86	87	88
00000220	89	8A	92	93	94	95	96	97	98	99	9A	A2	A3	A4	A5	A6
00000230	A7	A8	A9	AA	B2	B3	B4	B5	B6	B7	B8	B9	BA	C2	C3	C4
00000240	C5	C6	C7	C8	C9	CA	D2	D3	D4	D5	D6	D7	D8	D9	DA	E2



The brain use a basic system as below :

- Formes ⟶ which replace 0
- Colors ⟶ which replace 1

Forms generally all forms that exist are included in a sphere, the point through every possible and imaginary forms.

Colors : the human brain use the visible spectrum which is located between the ultra-violet and infrared.

This is the recognition algorithm brain, or neuronal algorithm, the images are captured by the eye, they are digitized by retina cells (cells in cones and cells in sticks) and are transmitted by neuronal algorithm to the brain which receives them decodes them and reconstitutes them in real holographic

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

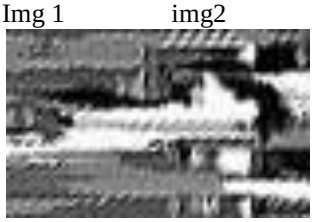
images; therfore the numerical algorithm used by computers gives us unique images.
The algorithm uses the neural brain gives us holographic images or pictures « DATA BANK IMAGES » i.e. each hologram contains an infinity of images each of which visualizes specific information

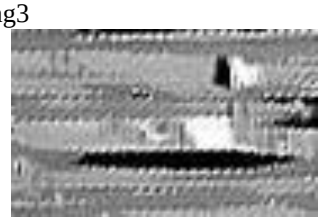
Example: Kidney: holographic images images of scenes


Img 1


img2


img3

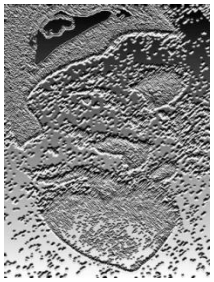

Img4


img5

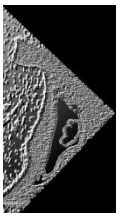
Img1: images obtained by MMR system
Img2: holographic image represented by the brain we notice that if we make an enlarging the holographic image2 we will discover precise images of many things known like a boat in img3 or of another boat different from the first in img4 or a submarine in img5,
To note a remarkable feature is that the more we enlarge the hologram more details emerge, there is no phenomenon in pixelation. A holographic image of the brain is an image that contains a within itself a very precise infinity of images relating information, and does not obey the phenomenon of pixelation during its expansion.
To be clear: the algorithm coding images of the brain gives holographic images, each holographic image enlarging brings up very specific recognizable image, it does not follow the typical raster, so we in a holographic picture a large number of images.
Holographic images : writings and symbols
But each global holographic image contains itself several holographic image contain other images enlargement
Exemple :

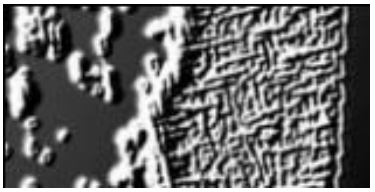

Img1

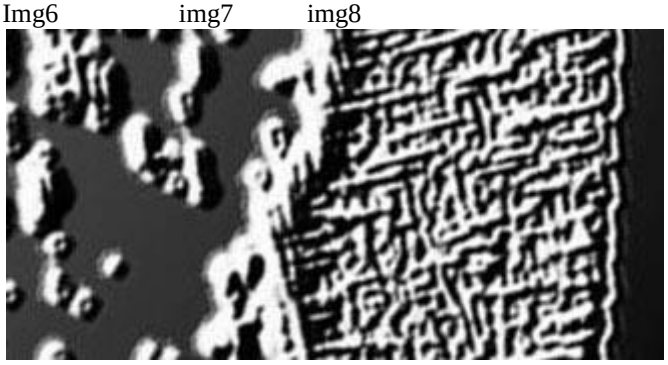

img2


img6


Img6


img7


img8


Img9

Each hologram gives us pictures on a specific dmaine.
The img6 pictures reveal to us writings and symbols another holographic image mapping reveal to us more sites etc... is well known to all disciplines and other disciplines unknown.
I want to say a special publication process memory images, here I give this example to explain and understand that perceived by the brain images is actually a holographic image containing an infinite holographic images of each it deals with a specific and particular discipline.
That said, I return to the bioinformatics brain structure MMR System3 (Micro Mass Reconstruction 3) allows us to represent biological stucture as a computer structure, starting from an image the system gives us an architecture in computer network.
the technique is to posed a specific sensor on a body that will capture all the energy of the body, this énergeie contains all the information in this organ microscopically and macroscopically, the second step in the reconstruction of the composite body, this reconstruction will contain all the information of the study body, just like a photocopy of the original.
It will be sufficient in a third step to take pictures of the organ and study in general appearance and microcopique the pictures taken are images 'data bank' that is to say that each image contains an infinity of images dealing each with a particular aspect

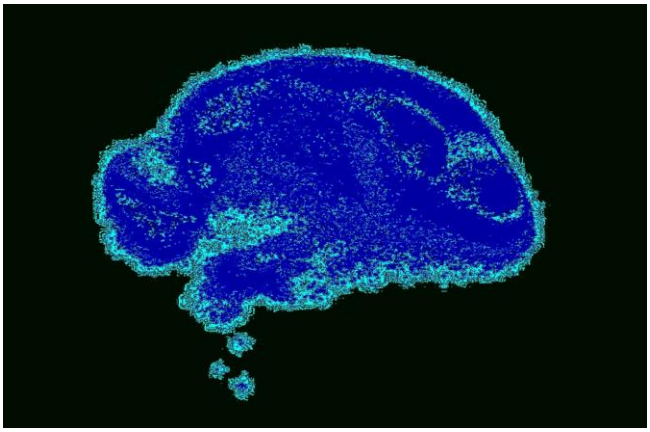
2.4 MMR3 images of the brain

The brain rebuilt out of composite material starting from the collecting of its energy

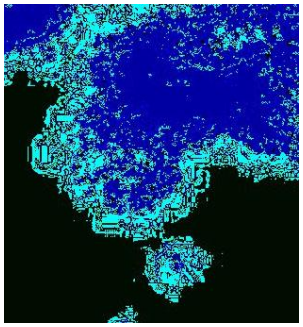
Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

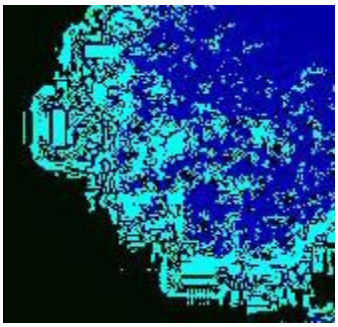

Img 10

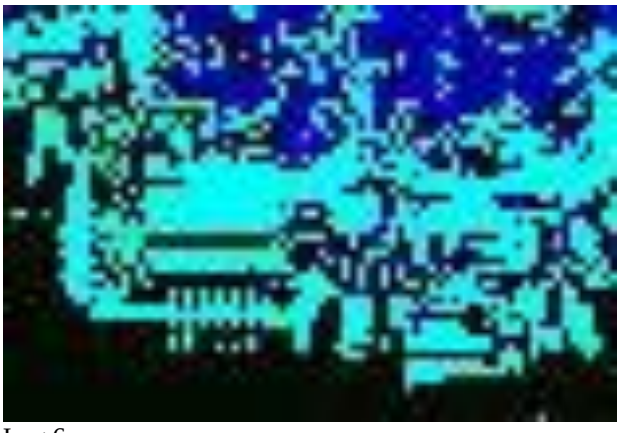

img 11

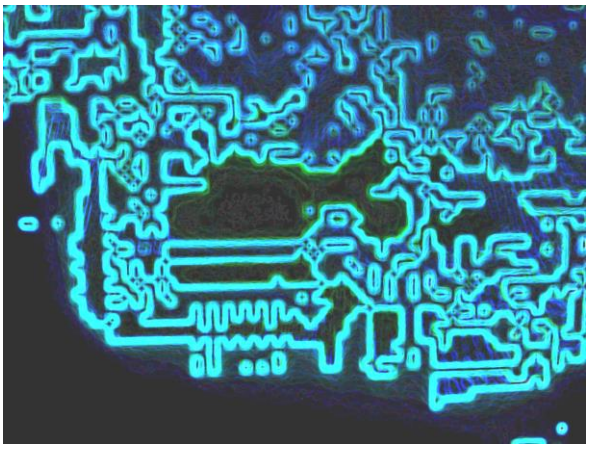

Img 12

Img 1 : image of the composite material brain starting from the collecting of its energy
Img 2 : negative of the image img 1
Imag 3: application of the technique MMR.
The enlarging of the image img 3 enables us to see ls structure bio-informatisue brain in its mondres details.


Img 4


img 5

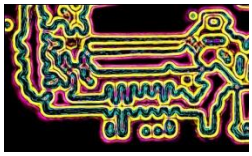

Img 6

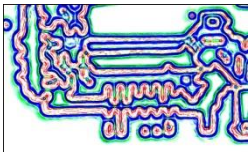

Img 7

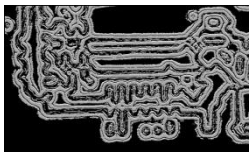
Img 4 : enlarging Bioinformatics structur
Img 5: enlarging Bioinformatics structur
Img 6: Bioinformatics circuits
Img 7: schematization of the Bioinformatics circuits

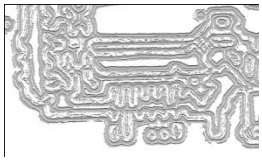
2.4.1 The bioinformatics brain :

MMR3 or masses micro reconstruction allows us to schematize the whole of the nervous cells, neurons and inter neuronal connections in data-processing circuits, this representation enables us to see and include the diagram of this unit and to represent it in circuits, it appears clearly that the provision of this unit follows a very precise diagram very near to the computer diagrams.
The architecture of the neuron is quite precise, it consists of a cellular body, ramifications and of a prolongation


Img 8

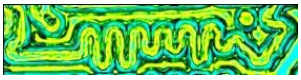

img 9


Img 10


img 11

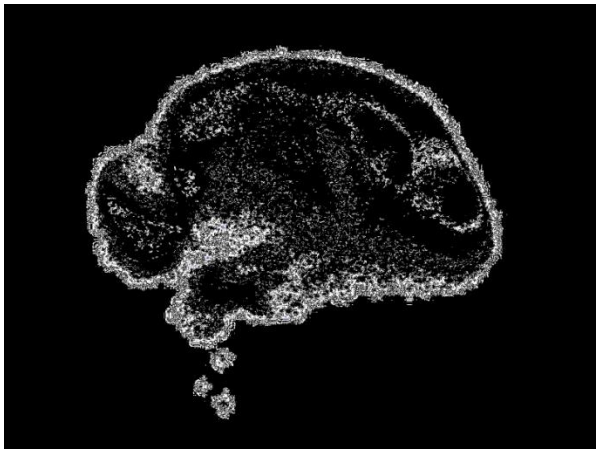
Img 8: data-processing schematization, circuits
Img 9: data-processing diagram circuit
Img 10, img 11: data-processing circuit
The more important the enlargement is and the more discovering us the details very characteristic of the images.


Img 12

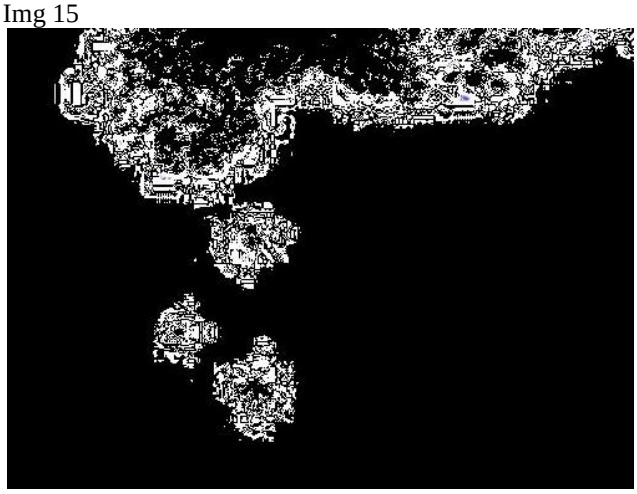

img 13


Img 14

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America



Img 15



Img 16

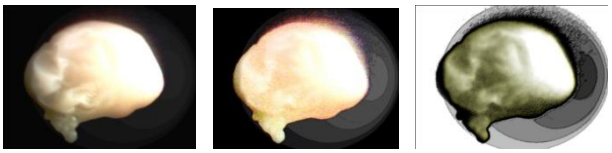
Img 17

Img 15: Bioinformatic structur of the Brain
Img 16: enlargement of bioinformatic circuits
Img 17 : enlargement of bioinformatic circuits

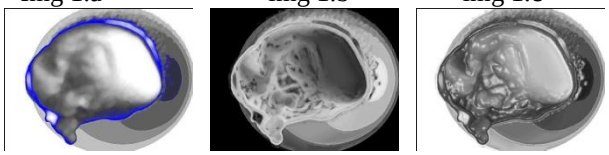
We notice that the arrangement is very accurate, the MMR3 gives us a very accurate picture of the appearance and layout in all circuits, which are very similar to computer circuits.

The human brain is composed of four zones.

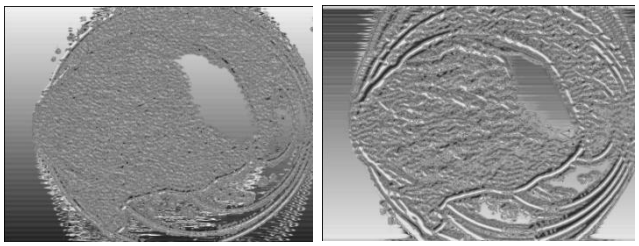
- wave : of superior frequency higher than 12Hz and power of some microvolt's
- The Alpha wave : their frequency is between 8.5 and 12 Hz
- The Theta wave : frequency 4.5 between and 8 Hz
- The Delta wave : frequency 4 Hz primarily collected during the Dreams period.



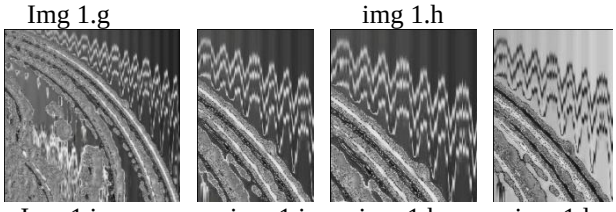
Img 1.a img 1.b img 1.c



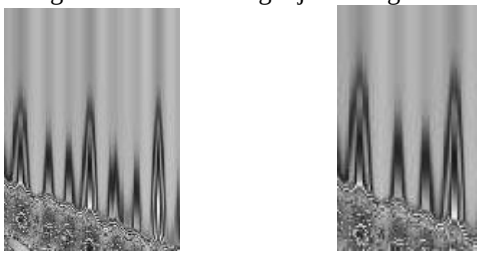
Img 1.d img 1.e img 1.f



Img 1.g img 1.h



Img 1.i img 1.j img 1.k img 1.l



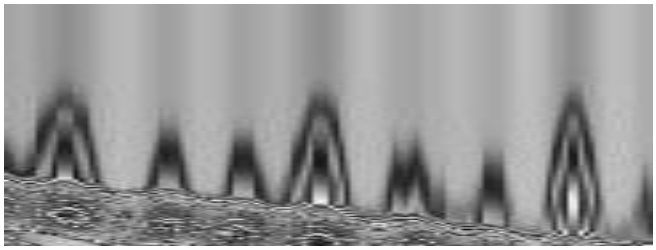
Img 1.m img 1.n

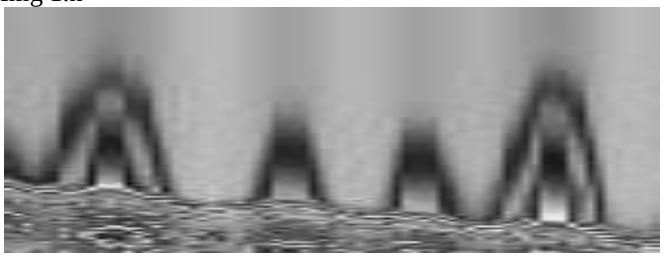
Img 1.a : brain in composit material.
Img 1.b : images illustrating the four fields electromagnetic of the brain.
img 1.c : four field visible field.
Img 1.d : delimitation of each field.
img 1.e : four field visible.
img 1.f : four field visible.
Img 1.g : other image reflecting the electromagnetic field.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

img 1.h : electromagnetic field image.
img 1.i : enlargement.
img 1.j : enlargement.
Img 1.k : wavelength of the field.
img 1.l : electromagnetic wave.
img 1.m : wavelength of the electromagnetic field.
img 1.n : period of electromagnetic wave.

With M.M.R system we can see the electromagnetic fields of the brain, we remark that there are four fields; the first is Frontal, the second is temporal, the third is parietal and cover the frontal and the temporal, and the fourth occipital it cover all the others.
The images visualize the electromagnetic wave broadcasting we distinguish the wave period with a great clearness.


Img 1.n


Img 1.o

we note the broadcast of an electromagnetic wave in the nous donne la dispositions image 1.n, with the periodic repetition of the wavelength observed in the image 1.o.
So the MMR system allows us to visualize the electromagnetic fields of the brain, it also allows us to visualize the emission of electromagnetic waves with a repetition period of the wave, we can easily identify the frequency and period, all it's visible in images that are real images, as opposed to their detection by an electroencephalograph which can only be measured.

3. Conclusion

The brain is a bio-computer quantum which works with a photonic algorithm whose base is : colors and forms. The colors are represented by three colors which give the whole visible spectrum ranging from infrared to ultraviolet ; the forms are represented by the point and the sphere; we know that any forms lies between the point and the sphere. The photographs which the brain receives are not fixed; but rather a holographic film whose each holographic photograph is made of an infinity of photographs representing all information of the object observed at the moment T or he was seen.
The brain is an assembly of complex circuits bioinformatic or CHIPS BIOLOGICAL working with a photonic algorithm.

Energy is also different, because the brain works with the energy produced by the degradation of sugar in ADP (adenosine di phosphate) in ATP (tri adenosine phosphates) and ENERGY.
The Storage capacity of information this biological quantum computer is unlimited, storage starts in the uterus to finish with death, i.e. with the destruction of the brain.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Review of Research on Biological literature text mining

Kong Meijing[1],Zhang Le[2,3],Wang Jun[4]
[1,4]College of Computer and Information Science ,Southwest University, Chongqing 400715, China
[2]Department of mathematical sciences, Michigan Technological University, Houghton, MI 49931, U.S.A.
[3]College of Computer and Information Science, Southwest University, Chongqing, 400715, China.
[1]a369@swu.edu.cn, [2]zhanglcq@swu.edu.cn, [4]kingjun@swu.edu.cn

Abstract - *Currently, the field of biomedical research is booming, a lot of biomedical knowledge in unstructured form in all forms of text file, and now it is the exponential trend increase, how to solve the contradictions between massive growth of information and knowledge of text slowly, in a credible way to find useful patterns in the text is a challenge. In recent years, biomedical text mining technology which is one branch of an efficient automatic access to new exploration-related knowledge, has great progress. This review describes the major methods and results of biomedical text mining research. namely information retrieval and literature search tool and the main aspects of biological text mining biomedical named entity recognition , text categorization, abbreviations and synonyms of the word recognition, relationship extraction, forming hypotheses and integrated framework for the above work, and finally the recent developments in the field is summarized and discussed.*

Keywords: Bioinformatics; text mining; computer applications.

1 Introduction

Text Knowledge Discovery (Knowledge Discovery in Texts, KDT) is focused on the discovery and excavation from the text inductive knowledge, such as useful models, trends, rules of knowledge and other computer procedure. KDT, namely text mining technology (Text Mining), is the products of the combining theory with technology which include artificial intelligence, machine learning, natural language processing, and data mining and related automatic text processing, such as information extraction, information retrieval, text classification[1].

With the development of computer network technology and the emergence of network version of the journal, medical bibliographic databases and other full-text database, biomedical text mining become one of the more active areas, and even some people in this area named KDiBL (Knowledge Discovery in Biomedical Literature). Using text mining techniques to deal with the massive scientific literature and biomedical text annotation data stored in databases, which found that research and innovation knowledge (such as genes, proteins, diseases, drugs and their relationship) is a hot research field of the current artificial intelligence and data mining. Text mining is a method for automatic computer processing of natural language text, it does not completely solve the pursuit and realization of natural language processing, and machine learning methods positioned to achieve limited objectives apply knowledge extraction and excavation.

The purpose of biomedical text mining is to help researchers in the massive literature more efficiently identify the information they need, find hidden relationships, and apply mathematical algorithms, statistical methods and data processing methods literature analysis and processing, so the information overload pressure from investigators passed on the computer.

Applications of the biomedical field text mining technology can improve biomedical information construction and management efficiency. Constructing biomedical databases, is the driving force to promote the advancement of biomedical text mining technology[2]. Through information extraction technology, you can build special proteins role relational database which is based on the purpose of disease diagnosis and drug design, Such as breast cancer. Alzheimer's disease-related protein effect relationship database. Proteins database described by the network, will greatly benefit disease diagnosis, drug design, progress related to the promotion of biomedical research[3].

The greater sense of biomedical text mining, is that biomedical text mining techniques can help biomedical researchers perform faster and more effective research in several areas, such as information retrieval technology can help users quickly and efficiently find useful information in a massive collection of document information; information extraction technology can extract specific factual information from the biomedical literature, that the entire biological knowledge to build the network, the relationship between organisms forecasts, development of new drugs will have great significance; text classification techniques can be screened coarse-grained to narrow the search scope for further information processing in preparation; assume that technology can be tapped to generate experimental hypotheses and experiments suggested in the literature in order to get biologists to verify new scientific discoveries. Therefore, this research has attracted wide attention from computational linguistics, bioinformatics, machine learning and other aspects of the researchers, the paper describes the biological literature focuses on information retrieval tools and biomedical named entity recognition, abbreviations and synonyms recognition, biomedical entity relation extraction, was found to form hypotheses and integrated framework for the relationship between the above work, the other associated technical evaluation and so on.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

2 Biological Literature Information Retrieval

As one of the main achievements demonstrate the biological literature and academic exchanges, the number of large, fast growth rate of far more than the other disciplines. NCBI (National Center for Biotechnology Information), 1988 as a U.S. National Information Resource Center of Molecular Biology established, it's main task is to establish a public database of computational biology research, and develop some software tools for analyzing genome, spread biological information. It offers 28 free database, PubMed is one of these databases which is retrieval system using a uniform interface (Entrez).As important source of literature in the field of life sciences PubMed, has been favored by the majority of medical researchers, with using the Boolean model and vector space model and a dictionary to automatic query expansion to make information more collective. Most users through Entrez database access and query the PubMed database (Entrez / PubMed), but there are some drawbacks in the search interface and functionality of the system [4]. Such as, It's difficult for new users to grasp limited search and complex retrieval methods of Mesh and it's quite boring for user to read these lengthy list of search results because of lacking of analytical tools to navigate the search results. To improve efficiency, many dedicated search engine came into being. So, we will briefly present some common biological literature search engines to retrieve text and non-text material .

2.1 Handling Text Material

With the rapid growth of biomedical literature, information retrieval researchers need to get the aid they need literature, following a relatively common example of HubMed and GoPubMed .

HubMed (http://www.hubmed.org/) is an optional search interface combined with external network services and provides a function to improve the literature search, browsing and retrieval efficient[5]. Users can create and visualize clusters of related articles, reference data in a variety of formats, to get updated daily publications and browse to the full-text links. HubMed is the EUtils of the NCBI Web Services, interpreting the PubMed data in another way. HubMed and PubMed use the same syntax to search and get the same result, but the expression is different between HubMed and PubMed. The result of HubMed can be sorted by date information or relevance. GoPubMed (https://www.gopubmed.org) is a system combined with semantic network of biomedical information retrieval ,to classify PubMed search results, which can help users quickly locate the most relevant literature, and base search result on visual statistical analysis from multiple perspectives[6]. GoPubMed works by using the extraction ontology terms (Go terminology, MeSH term and UniProt protein), and the search results are divided into four categories: What (ontology terms), Who (Author), Where (city and periodicals), When (publication date), which users can quickly browse required documents by classified Navigation. In addition, GoPubMed can analysis literature information.

2.2 Handing non-text material

The traditional retrieval method is based on paper level which meet the needs of most of the researchers [7].However, it had to spend a lot of time and effort to collect relevant literature and determine the correlation, and these items embedded in the article in-depth content, such as, embedded charts, tables, diagrams, maps, photographs, etc., it is hard to be retrieved. In fact, these non-text not only to assist researchers on article content flexible and in-depth analysis, but also to promote knowledge discovery, so it has important value. In addition, charts and data are often more convincing, giving the impression that clear at a glance.

For the image retrieval , first cellular localization image search (SLIF) system[8, 9] , SLIF extract and analyze specific types of images, the use of geometric moments, word processing and morphological image processing to extract all graphic images in the full text of journal articles BM and to identify these images depict fluorescence microscopy[10] , and then determine the subcellular capture Location digital features (calculating SLF6 features and converting the output of a single fraction). Although the images provide important evidence, without reference to the relevant text often cannot understand.Yu examined the three associated text: the captions of figure, the main abstract sentences and the associated sentences appearing in the full text[11]. Conclude that the summary sentences can be used to summarize the image content and other relevant text description usually does not include the indications described in experimental procedures and conclusions of the experiment.

BioText Search Engine provides free of charge biologists search engine which is based on web Service[12] . The current system indexes all Open Access articles available at PubMed Central. In order to improve the efficiency of the figures information retrieval, BioText Search Engine for articles extracted from local storage, use Luccene open source search engine indexing tool kit[1]. Currently BioText Search Engine is mainly based on keyword search, which is including keyword search the full text and abstracts, captions of figures and tables illustrate retrieval. Searcher through "view all figures and tables from this article" link associated with the article realize all charts or tables, but also through "view full article" link experiment with the full text of the link graph. BioText Search Engine's development objective is to provide users with a more comprehensive search portal, such as author, journal name, and support like genes, proteins, organisms, species and other local features retrieval. Another search engine FigSearch can be extracted by retrieving summary chart containing the gene names [13]. Compared with the BioText Search Engine, it's more targeted, but shows the limitations of its retrieval.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

While researchers have been working on how to improve the retrieval efficiency, but there are still many urgent problems for the non-text information retrieval. First, the accuracy rate is not high, and the current precision of the system cannot meet user's expectations. We can find three factors affecting its precision by analyzing .Above all, it's difficult to accurately extract the target. Then, whether the target object accurately indexing, indexing words that accurately describe whether the target object topic. Finally, there is no a standard to describe metadata. Second, paying for resources lead to restrictions on access to get full literature as result of copyright issues or other issues. Third, the quality of graphics or table is not high, for example, data are incomplete or unclear, leading to its value in use not high.

Thus, we can predict the trends that how to improve the efficiency of non-text information retrieval. Following is about this.

First, the type of retrieve objects will continue to be expanded, currently the main targets for article retrieval of charts, tables and photos. In addition, the academic literature also contains a variety of formulas, flow charts, scientific data and other resources, which play an important role in research work.

Second, Content-based retrieval methods will be more and more attention. Retrieve content based on depth refers to the semantic content according to the target object, and analyze contextual understanding, in order to achieve a deeper level of technical retrieval methods, such as color, texture and shape of the images, some distribution of tones, shoot theme, scenes in the photo. This retrieval method which directly extracts content from the target object can perform similarity retrieval and meet the user's needs to retrieve multi-level, interactive and intuitive query. Third, meta-data [14] will continue to be improved due to target objects belonging to a new resource type, and the type, format and dimensions are quite different. It's impossible to accomplish description of all the items through the establishment of a metadata Description Standard. So, it's difficult and complicated to build object metadata. Database developers and experts in various disciplines required more effort to construct the metadata standards, which will also become a major research focus in the future.

Fourth, to develop complex integrated retrieval will be a new direction. The current retrieval focus on the integration of text information .However, integration of non-text information retrieval, text information retrieval and cross with non-text information has not yet been solved. Meanwhile, it's a challenge that researches encapsulate different sources with the same subject text information, scientific data, pictures, charts, maps, photographs and other information into a composite object.

Overall, the challenges facing the biological field of information retrieval is how to combine biological background knowledge, correct understanding user queries and various biological entity names appearing in the biological literature. This will be a great present or future research focus.

3 Text Mining

Because of the multitude of biological literature, so a lot of life science issues, and bioinformatics problems, need to use text mining methods, for example, we can find the relationship between genes and diseases, protein interactions, and so on.

The main task of text mining are term recognition, information extraction, discover relationships, and applied to the field of bioinformatics. Currently, some active area of research are named entity recognition, text categorization, relationship extraction, synonyms, abbreviations extraction, hypothesis formation, text classification and more integrated framework.

3.1 Terminology Recognition

3.1.1. Named entity recognition (NER)

Biological field named entity recognition is to find the name of the biological entity from the biological literature, which is marked as correct category (such as genes, proteins, disease, drugs, etc.). It is an essential part of biology literature retrieval and information extraction.

One of the basic tasks of biomedical text mining named entity recognition (Biomedical Named Entity Recognition, Biomedical NER), its purpose is to identify the name of the specified type from the collection biomedical text, such as protein, gene, ribonucleic acid, DNA and so on. This is a critical step for relationship extraction and other potential information.

However, the new named entities continue to emerge and the names of entities which has not been recognized are increasing. Currently there is not a complete biomedical named entity dictionary to contain various types and following features also resulted in the identification difficulties.

First, the length of the entity name is very long, the phrase may consist of multiple-word. For example, "normal thymic epithelial cells".

Second, more than a common noun to show an entity. Many biomedical named entity with "and" or "or" connected parallel structure, they share a common noun, so it is difficult to correctly identify named entities，for example,"91 and 84 kDa proteins"。

Third, multiple forms of expression of the same name of the entity, specifically, a lot of biomedical named entity has written a number of different forms, for example TNFA, TNF an alpha, infalpha and TNFalPha all represent the same gene.

Fourth, different names share the same form of expression. Specifically, the same word or phrase that can represent different types of biomedical named entities, depending on the context can be inferred, for example, IL22 both expressed protein name, and said DNA name.

Fifth, the nesting phenomena of biomedical named entity still exist, e.g.<PROTEIN> <DNA> kappa 3 </ DNA> binding factor </ PROTEIN>, where "kappa 3" is the gene name , and "kappa 3 binding factor" is the

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

name of the protein, and therefore it's necessary to solve the problem of overlapping candidate named entity.

Sixth, a large number of biological named entities using abbreviations.

Seventh, it is difficult for all the names to be enumerated and be included in the dictionary.

On the whole, these features brought great difficulties to determine the boundaries of the entity's name, therefore, biomedical named entity recognition will be a challenging study. Currently, they are about three recognition method of biomedical named entity. Following three categories are dictionary-based approach, heuristic rules-based approach and based on machine learning methods.

First, dictionary-based approach, the natural language text are compared with dictionary entry which include a large number of pre-existing biomedical named entity name, based on the results match. The algorithm requires a larger refinement over the name of the library as well as matching strategy. Implementation of this method is more direct and easy to understand, but because of the large number of names in the biological field, frequently updated and expression are inconsistent, so have to maintain a biological dictionary library is very difficult. Second, Rule-based methods, is to identity the entity and other text which is divided into different categories by defining rules. In fact, the field of bioinformatics is complex and many varied forms of entity names, and rules-based method requires a lot of knowledge structure, but it is impossible to cover all the rules, resulting in lower recognition rate. In addition, the new entity emerging name in the dictionary cannot be updated in order to contain all the new entity name, resulting in a lack of portability. Third, Machine learning-based approach, is to transform the named entity recognition problem into classification problem. By using the classification tools, the training text for machine learning, and thus to distinguish various types of entities named in the final test text identifies named entities. This method is to extract the need to identify the name of the entity model, as the standard for classification method in a tagging corpus of training. The method reflects the greater advantage with flexibility, adaptability to a specific environment, as well as the ability to deal with small sample. But it needs a large number of corpus, for manual annotation to serve as a training set, usually due to uneven data lead to "over-learning."

3.1.2 Synonyms and acronyms

Biomedical literature growth while biomedical terminology is also growing and abbreviations occupied high proportion, for example, IFN, TPA and so on. There is no discipline for the formation of a lot of abbreviations, and acronyms also has a high degree of ambiguity. In general, the extended form of the abbreviations has more than evidence to determine its category .Anyway, to identify acronyms must largely dependent on context, and cannot rely on existing biological dictionary. Biomedical entities have multiple names and abbreviations, if there is an automated method of collecting synonyms and abbreviations to help researchers conduct literature research ,it's will be very useful. In addition, if all synonyms and abbreviations entities are mapped to a term which represent the same concept, the other text mining tasks can be completed more efficiently.

Many researchers use an online database to generate a list of synonyms gene names.Yu and Agichtein combines AbGene gene named entity recognition system, using statistical methods ,SVM classifier and automatic mode extraction[15], as well as hand-generated rules algorithm to extract synonymous in the full text of journal articles . According to statistics, their system's recall rate is about 80% and accuracy rate is about 9% then the total value of the F-measure is about 30%. Thus, the problem which automatically identify abbreviation and its full name in a single article has been basically solved, and the identification systems have achieved a higher accuracy and recall rate. Future research will combine identify abbreviations with other text mining tasks, and apply them to real biomedical text mining system.

As we all know, synonym extraction accuracy of gene and protein names is also generally lower, so there are more challenging. Although the list of synonyms automatically updated, to improve the performance of document retrieval and text mining system is valuable, but the accuracy of automatic extraction system is too low. However, the current work in progress, the official are standardizing gene names and symbols of the protein, so in the future the low accuracy problem caused by informal name may be reduced.

3.2 Information Extraction

Information extraction of the biological field is to extract the relationship which may be protein-protein interactions, or gene regulation relationship, according to user-defined interactions from biological literatures. This has important significance for creating a whole network of biological knowledge to predict the biological relationships, and development of new drugs.

There are two extraction method which are co-occurrence and natural language processing methods (Natural Language Processing, NLP) [16]. Co-occurrence method is based on a class of entities from the literature, which assumes that the closer the biological literature entities, the higher their relevance. For example, if two entity objects appear in a sentence and have a high frequency of co-occurrence of two entities by text mining dealing with a large number of abstracts, it imply the existence of two entities associated with a high probability target. Such method is relatively easy to implement, but the assumption is not reliable. The other is natural language processing method, is based on large-scale corpus analysis and complex language model to extract relationship by grammatical structure and semantic relations .Although it's more complex and difficult to achieve, but the association is more accurate and reliable.

On the basis of the relationship between two methods, some researchers have tried to extract the biological protein and build pathway from the literature and the network

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

(Network). And we all know that extracting genes, proteins or gene ontology (GO) has a direct significance. Chiang and Yu's Meke system using gene ontology (GO) as a function of the name of the dictionary coding, the genes and gene products with the name of the dictionary LocusLink combined sentence arrangement while taking advantage of the system to determine patterns associated with gene function , then the model combined with the naive Bayes classifier to extract sentences containing information about the function of the gene product[17]. 2004 BioCreative Task two also extract relevant gene ontology (GO) encoding gene from the free text.

Although, in recent years, the biological literature information extraction system from a simple rule-based extraction for advanced development of the use of computer language parser, however, there are still many problems to be improved, the current text mining systems cannot meet the actual needs of researchers and performance evaluation criteria system is not yet mature, so biologists and computer scientists need to cooperate further strengthened.

3.3 Hypothesis Generation

Relationship extraction is mainly extracted relationships between entities that can clearly be found in the text, while the hypothesis tries to reveal the more explicit relationship that does not exist in the text but inferred by other. The purpose is to reveal previously undiscovered deserves further study the relationship. Text mining is defined as "by automatically extracting text from different information resources, the discovery of new information previously unknown." With extracts from the published literature to determine the information compared to pay more attention to the unknown text mining access to information. For such an experimental biological sciences, the biological text mining aims to dig out from the literature experimental hypotheses and experiments suggested that in order to get biologists to verify new scientific discoveries.

In fact, all the research hypotheses are formed idea of using "non-relevant literature complementary structure" of the 1980s, Swanson proposed is simply ABC model: A affected B, and B on C, then A may influence C in the 1980s and early 1990s Swanson ABC model gives a lot of use and tap new hypothetical example, such as: fish oil treatment of Raynaud's disease [18]; magnesium deficiency and headaches relationship[19].

Hypothesis mining system is not yet a standard tool for biologists, but someday there is a need to continue research and improve the ability of the system handle a large number of different types of data. At present these data must be manually researched by scientists and need better ways to evaluate these system, which can record the improvements of the system and to make a clear choice.

3.4 Text Categorization

Reading massive biomedical literature bring great difficulties for the researchers in recent years, but the widespread use of text mining and natural language processing techniques has solved this problem, in which text classification is an important part. Text classification attempt to automatically test document or some parts whether the document contain the interesting features of a particular subject. Interesting information is not explicitly specified by the user, but to provide positive training set (set of documents have been found to contain features of interest) and negative training set. Text classification system should be able to automatically extract features that can distinguish between positive and negative, and which are applied to the candidate documents and then to make decisions.

Accurate text classification systems are particularly useful for database administrators, who may have to browse a lot of database to find a small amount of literature which contains some valuable information, but more biomedical information is created in form of text, the database manager need to convert the information to the encoded data and there is a strong need that text classification method is applied to biomedical text.

Yeh organized the text mining contest that is the part of KDD international competition in 2002[20]. And the task is to evaluate papers FlyBase data set, and according to the Drosophila gene products to determine whether it should manage that paper. The performance of the best entries are created by manually generating a set of rules which are based on tagging, semantic dictionary and semantic restrictions and the F is 78% by detecting the raining document. Another effective method is based on the text whether there are the gene products to classify biomedical papers, and after extracting feature, to use the Naive Bayesian classifiers, have a good performance.

The study that applying text classification to the actual work of biomedical managers and indexing has only just begun. one of the tasks of 2006 TREC Genomics Track is text classification problems[21] . This task is trying to imitate manual annotation, in order to find the document experimental evidence of genetic information contained in the mouse genome informatics (MGI) system browsing, and finally completed the text of the Standard Generalized Markup Language collection (SGML) format. To evaluate the performance of the task and to meet the needs of managers and other researchers for the future .Text classification technology can improve the efficiency of biological databases were screened, so improving the text classification in biomedical research must continue.

3.5 Integration Framework

To address a wide variety of user needs, many research groups are developing integrated text mining framework. The MedScan systems of Novichkova and others developed, is a synthesis of the dictionary and syntactic, semantic template as a set of entity relationship extraction of universal biomedical text mining system[22]; Glenisson developed TXTGate, perform the text profiles and clustering of genes by using multiple online databases

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[23]; Becker created the PubMatrix tool which combining multiple queries PubMed list the two-dimensional gene lists contained the names and functions of terms[24]; BioRAT system of Corney is another template-based system that combine template design tools with web spider that is used for locating and retrieving the full text of journal articles combination[25]; Textpresso system is an Ontology-Based Information Retrieval and Extraction System for Biological Literature[26].

However, there are also some problems which all of these systems are still in the research and development phase and the evaluation tends to simple. In addition, the system has not conducted a thorough evaluation of the user, and the ability to meet the needs of the biomedical research community remains to be seen, this is just to satisfy a stage biomedical researchers needs.

3.6 Evaluation Of Biological Text Mining Technology

In the field of bioinformatics, The technical evaluation of the text mining systems which were mentioned in the above has two indicators which is precision and recall rate[27], precision is identified by the creature in the named entity named entity of the ratio ,namely accuracy rate (P) = the number of correctly predicted named entity / entities to predict the total number of names; Recall rate refers to the experiment identified by all the correct ratio of biological named entity of the experimental data, that the recall rate (R) = total number of correctly predicted named entity named entity number / the total number of named entities of a text, we can see precision and recall rate which reflects two different aspects of recognition quality. So, we can use harmonic mean as a balance recall and precision rates, namely, F-measure = 2PR / [P + R].

4 Summary and Discussion

In summary, we can find the biomedical text mining has tremendous potential. The experts in various fields especially medical researchers are also gradually realizing that the job prospects. So, the next major challenge in text mining work is to develop efficient text mining tools, making it essential for biomedical researchers so that they can continue more productive under the pressure of the rapid growth of information. Research must focus on helping biomedical researchers to solve practical problems, and less on independent systems to meet user demand output. Of course ,it needs to build a bridge between biologists and computer experts in the field , because the field of researchers led by a computer background , but only biologists to propose effective evaluation methods and mining tasks . Thus, biologists and computer linguists should make further cooperation in the field of text mining show the diversity and innovation, so as to understand the actual needs, access to domain knowledge to develop an effective text mining tools.

For the future, although the biological literature mining related technologies carried out some research, but still need to continue to look at areas in-depth study, for example, efforts to improve the identification and biological information extraction performance of biological entities, in addition, in terms of literature and words must be easier to obtain in the full text of journal articles, which did not mention in the Summary and Mesh terms. Current text mining research has shifted from the title and abstract, but still get the full text of copyrighted restrictions. Therefore, the research community must cooperate with publishers to obtain a wide variety of content for text mining. Next, in addressing the specific text mining tasks, more research is needed to determine what features work and what types of features are useful. Feature space for text mining is a huge array of feature types, feature types, including (but not limited to) the words, concepts, keywords, format, author, references and links. Popular long- term bags (bag-of-words) method can be applied to a variety of texts from different sources, but ignores the location of the document and paragraph information. Finally ,Obviously, the main theme of the future development of the coordination and cooperation between disciplines must work together text mining researchers , publishers and biomedical researchers can produce by providing a consistent , measurable, verifiable results system to meet user needs researchers must take the lead in coordinating efforts to achieve biomedical text mining full scientific potential . Obviously, the main theme of the future development is the coordination and cooperation between disciplines must work together text mining researchers, publishers and biomedical researchers can produce a consistent, measurable, verifiable results system to meet user needs. Researchers must take the lead in coordinating efforts to achieve biomedical text mining full scientific potential.

5 Acknowledgements

I thank Mr. zhang because of his constructed advice. And I am grateful to the authors who is quoted be on this paper.

References

[1] Jensen, L.J., J. Saric, and P. Bork, Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet, 2006. **7**(2): p. 119-29.

[2] Zhu, F., et al., Biomedical text mining and its applications in cancer research. J Biomed Inform, 2013. **46**(2): p. 200-11.

[3] Blaschke, C. and A. Valencia, The Functional Genomics Network in the evolution of biological text mining over the past decade. New biotechnology, 2013. **30**(3): p. 278-285.

[4] Wang, J., I. Cetindil, and S. Ji, Interactive and fuzzy search：a dynamic way to explore MEDLINE. 2010.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

[5] Eaton, A.D., HubMed: a web-based biomedical literature search interface. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W745-7.

[6] Doms, A. and M. Schroeder, GoPubMed: exploring PubMed with the Gene Ontology. 2006.

[7] Shah, P.K., et al., Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics, 2003. **4**: p. 20.

[8] Lewis, J., et al., Text similarity: an alternative way to search MEDLINE. Bioinformatics, 2006. **22**(18): p. 2298-304.

[9] Cohen, W.W., R. Wang, and R.F. Murphy, Understanding captions in biomedical publications, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003, ACM: Washington, D.C. p. 499-504.

[10] Yu, H. and M. Lee, Accessing bioscience images from abstract sentences. Bioinformatics, 2006. **22**(14): p. e547-56.

[11] Yu, H. and M. Lee, BioEx: a novel user-interface that accesses images from abstract sentences, in Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. 2006, Association for Computational Linguistics: New York, New York. p. 189-192.

[12] Hearst, M.A., et al., BioText Search Engine: beyond abstract search. Bioinformatics, 2007. **23**(16): p. 2196-7.

[13] Liu, F., et al., FigSearch: a figure legend indexing and classification system. Bioinformatics, 2004. **20**(16): p. 2880-2.

[14] Hearst, M.A., et al. Exploring the efficacy of caption search for bioscience journal search interfaces. in Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. 2007. Association for Computational Linguistics.

[15] Yu, H. and E. Agichtein, Extracting synonymous gene and protein terms from biological literature. Bioinformatics, 2003. **19 Suppl 1**: p. i340-9.

[16] Yandell, M.D. and W.H. Majoros, Genomics and natural language processing. Nat Rev Genet, 2002. **3**(8): p. 601-10.

[17] Chiang, J.H. and H.C. Yu, MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics, 2003. **19**(11): p. 1417-22.

[18] Smalheiser, N.R. and D.R. Swanson, Assessing a Gap In the Biomedical Literature - Magnesium-Deficiency And Neurologic Disease. Neuroscience Research Communications, 1994. **15**(1): p. 1-9.

[19] Swanson, D.R., Fish Oil, Raynauds Syndrome, And Undiscovered Public Knowledge. Perspectives In Biology And Medicine, 1986. **30**(1): p. 7-18.

[20] Yeh, A.S., L. Hirschman, and A.A. Morgan, Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. Bioinformatics, 2003. **19**: p. i331-i339.

[21] Hersh, W.R., et al. TREC 2006 Genomics Track Overview. in TREC. 2006.

[22] Novichkova, S., S. Egorov, and N. Daraselia, MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics, 2003. **19**(13): p. 1699-706.

[23] Glenisson, P., et al., TXTGate: profiling gene groups with text-based information. Genome Biol, 2004. **5**(6): p. R43.

[24] Becker, K.G., et al., PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics, 2003. **4**: p. 61.

[25] Corney, D.P., et al., BioRAT: extracting biological information from full-length papers. Bioinformatics, 2004. **20**(17): p. 3206-13.

[26] Muller, H.M., E.E. Kenny, and P.W. Sternberg, Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol, 2004. **2**(11): p. e309.

[27] Hirschman, L., et al., Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics, 2005. **6 Suppl 1**: p. S1.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

SBMLChecker, a Semantic approach for SBML model reliability evaluation.

Mathialakan Thavappiragasam, Carol M. Lushbough, Etienne Z. Gnimpieba
Computer Science Department, University of South Dakota, 414 E. Clark St. Vermillion, SD 57069, USA,
{Mathialakan.Thavappi; Carol.Lushbough; Etienne.Gnimpieba}@usd.edu

ABSTRACT

In Systems Biology model design, reliability evaluation constitutes a requirements challenge. In order to apply the models on a given process or on work for in silico study, a systems biologist needs to be ensured of the models quality. The key problem remains the relation between the model and the biologist question. Several algorithms was designed to validate models but they only check correctness of syntax (e.g. Online SBML validator). These algorithms do not consider semantic annotation of a model defining biological context of the model. In our approach we have measured the model reliability using a combination of meaning (semantic) and syntax. This approach allows researcher to identify a model that really fits his needs and application domain. It also provides unique identification to each model element (compound, reaction, and compartment) in order to facilitate any Systems Biology operation such as merging, splitting, and simulation. It is implemented in Java and connected to the model database BIOMODELS using Restful API, our algorithm implementation called SBMLChecker is available online at http://jacksons.usd.edu/SBMLC/. The command line version has been deployed on BioExtract server, at bioextract.org that it to be integrated in automatic sharable scientific workflow.

Keywords: semantic, syntax, annotated URL id, SBML, biological model

1. INTRODUCTION

System Biology Markup Language (SBML) is the common format to represent a Biosystem mathematical model. Used by over 250 tools (SBML.org), it remains lacking in many aspects in order to provide the appropriate model in the right context. The reliability of a model depends considerably on the context related to the model design. The development of semantic annotation of biological elements allows systems biologists to connect design context (domain ontology) to a model.

Semantic in biological modeling

There are several organizations (EBI [2], NCBI [3]) maintaining databases (Biomodels [2], Protein, Gene, etc.) and/or ontologies (gene ontology [4]) in order to manage biological components (e.g., reaction, species, etc.) in a standard way. They try to categorize the already defined components and identify relationships among them. Each database assigns unique id to each element and keeps tracking relevant details (e.g. properties, description) with these ids. Furthermore, we can find several web applications (e.g., KEGG Mapper) that provide services to map the same components from different places [5]. Some of them provide web services, especially RESTful services, that could be used by software tool developers (web services for GO terms and annotation provided by EBML-EBI) [6]. A single component can be annotated by multiple databases and or ontologies. The SBML defines annotation-tag to annotate biological components, it has resources with the details of database and id for each annotation [7]. E.g. the reaction MTHFR, [5,10-methylene-tetrahydrofolate] + [NADPH] → [5-methyl-tetrahydrofolate] in BIOMD0000000018 has the annotations "urn:miriam:ec-code:1.5.1.20", "urn:miriam:kegg.reaction:R01224". This reaction has Enzyme id 1.5.1.20 and KEGG reaction id is R01224.

SBML reliability evaluation in existing tools (Online SBML Validator)

The model reliability checking should ensure their correctness on both syntax and semantic (meaning). The Online SBML validator introduced by SBML.org provides the services to test syntax and internal consistency of an SBML model. This system checks the following aspects of a model [1]:

- Consistency of measurement units associated with quantities (SBML L2V4 rules 105nn)
- Correctness and consistency of identifiers used for model entities (SBML L2V4 rules 103nn)
- Syntax of MathML mathematical expressions (SBML L2V4 rules 102nn)

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

- Validity of SBO identifiers (if any) used in the model (SBML L2V4 rules 107nn)
- Perform static analysis of whether the model is over determined
- Perform additional checks for recommended good modeling practices
- Perform all other general SBML consistency checks (SBML L2V4 rules 2nnnn; highly recommended)

However, this system does not consider the entire annotation information to evaluate the meaning of models. In order to analyze the semantics and syntax of models, we have designed a tool that extends the web services provided by the online SBML validator

2. METHOD

Principle

The reliability level of a model is calculated based on its validity of its syntax and semantics. Correctness of models on syntax is examined with the usage of web services provided by the online SBML validator. Semantic strength is measured by the annotated URL id of each model's component.

Design and algorithms for SBMLChecker

SBMLChecker does two way analysis, one for semantic strength and another for syntax correctness (Figure 5.), and generates reports R1, R2 respectively.

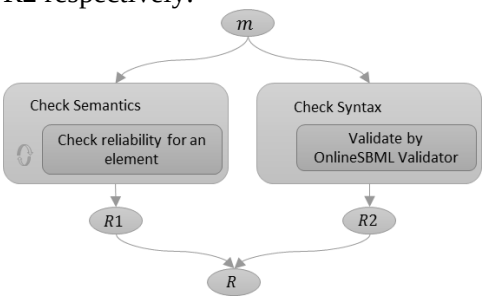


Figure 1. Global mechanism for model reliability evaluation

```
Data: model m
Result: Relaiablity report
declare report ;
for all elementsList L=(L1,...,Ln) do
    for all DBs/Ontologies for Li do
        for all elements E=(E1,...,En) in list Li do
            insertReport(checeRelaiablity(Ei);
        end
    end
end
insertReport(validateBySBMLOnlineValidator(m));
return report;
```

Figure 2. Algorithm for semantic analysis

The semantic analyzer takes each kind of component separately and identifies all ontologies and databases that are used to annotate it. For example, if any species is annotated with KEGG id, KEGG will be used to check the annotation of every remaining species. Then the percentage of KEGG annotation will be calculated. Species are considered to be more consistent/reliable if the percentage is high. In this way, percentage of every possible annotated ontologies/databases will be calculated. The maximum percentage will decide the best consistency level of the model element (e.g. species) in the resource (e.g. KEGG, MIRIAM Register)

Reliability score estimation

The consistency for the components (n_k) of kind *k* over n_{ok} ontologies and/or databases,

$$C_k = max(\forall_i, 0 \le i \le n_{ok}, percentage(O_i))$$

where O_i is an i^{th} ontology or database.

Finally, cumulative consistency is calculated by taking the average consistencies of each kind of component. Consistency of model m,

$$C_m = \frac{\sum_{\forall k} C_k n_k}{n}$$

where the number of components, $n = \sum_{\forall k} n_k$.

In addition to the consistency report, an error report is generated by combining the online SBML validator's error report with our own semantic check error report. Based on the quality checking, it will suggest to provide a valid model for any relevant applications such as model comparison and integration, but it can be skipped if they want.

Implementation of SBMLChecker

We used the IDE NetBeans (7.3) to fulfill everything related to coding, and the JSBML library (jsbml-0.8-with-depenedencies) was used to manipulate SBML files [8]. The JDK 1.7 java library were used for this development [9]. Furthermore, the library to handle excel file: apache poi, and any other relevant libraries were included.

The JSBML is a flexible and entirely java-based library for working with SBML. This library supports all SBML Levels and Versions through Level 3 Version 1, and it maintains the highest possible degree of compatibility with the popular library libSBML [10]. JSBML also supports modules that can facilitate the development of

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

plugins for end user applications, as well as ease migration from a libSBMLbased backend.

Validation

The model BIOMD0000000018 is examined for reliability by SBMLChecker. According to the results, it is syntactically valid and earned a semantic score of 79%. This semantic score comes from: compartment 100%, species 93%, and reaction 44%. The model's reliability can be improved by annotating it.

3. APPLICATION

SBMLChecker in a Workflow Management System (bioextract.org)

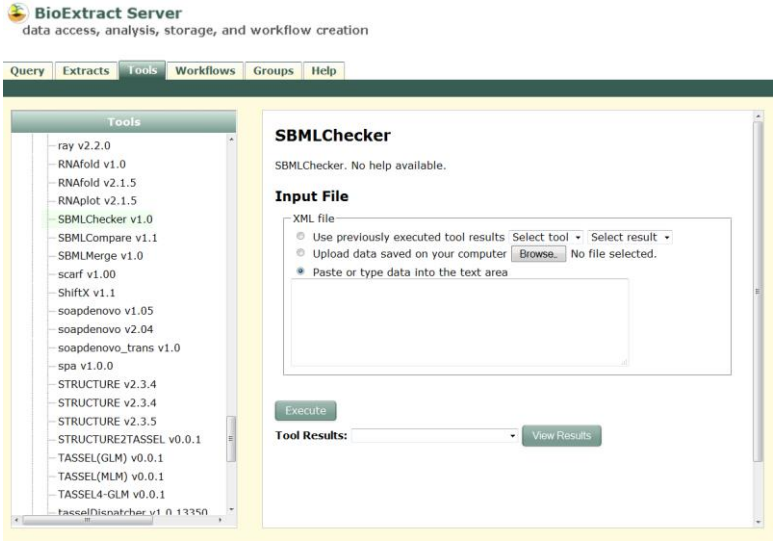


Figure 3 SBMLChecker on BioExtract server for the reliability checking of biomodels

A Java program named SBMLChecker.jar is designed for reliability checking. This can process SBML files received through command line parameter argument, and writes a generated report in excel, text, and xml formats. The SBMLChecker has been deployed on the HPC (High Performance Computing) infrastructure iPlant for availability on BioExtract server (Figure 3), and has been integrated on a web portal (Figure 4).

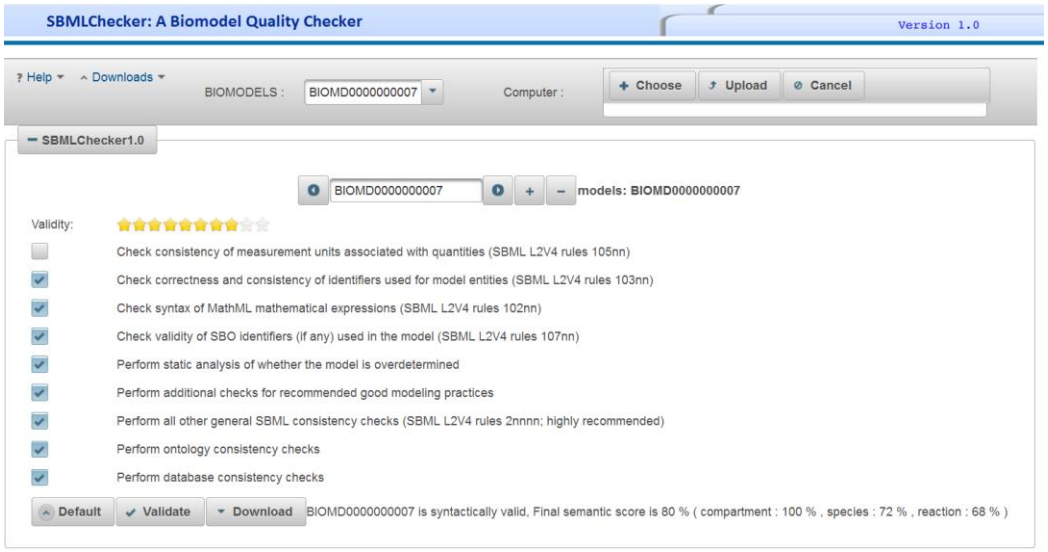


Figure 4 SBMLChecker on the web portal

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

4. CONCLUSION

SBMLChecker provide a novel approach in SBML model reliability measurement. Using a combination of the meaning (semantic) and the syntax, we generate a reliability score that can be used as indicator to interpret the output result from a given model in specific context. This approach also provides a unique identification to each model element (compound, reaction, and compartment) in order to facilitate any Systems Biology operation such as merging, splitting, simulation. Implemented in Java and connected to the model database BIOMODELS using Restful API, SBMLChecker is available online for small models and available on Bioextract.org for workflow design and big models.

Funding: This work was made possible by SD-INBRE Grant #P20RR016479-09 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

REFERENCES

[1] SBML.org, "SBML Online Validator." [Online]. Available: http://sbml.org/Facilities/Validator/. [Accessed: 10-Apr-2014].

[2] EBML-EBI, "BioModels Database," 2014. [Online]. Available: http://www.ebi.ac.uk/biomodels-main/. [Accessed: 10-Apr-2014].

[3] U. S. N. L. of Medicine, "NCBI." [Online]. Available: http://www.ncbi.nlm.nih.gov/. [Accessed: 10-Mar-2014].

[4] J. A. Blake, M. Dolan, H. Drabkin, D. P. Hill, N. Li, D. Sitnikov, S. Bridges, S. Burgess, T. Buza, F. McCarthy, D. Peddinti, L. Pillai, S. Carbon, H. Dietze, A. Ireland, S. E. Lewis, C. J. Mungall, P. Gaudet, R. L. Chrisholm, P. Fey, W. A. Kibbe, S. Basu, D. A. Siegele, B. K. McIntosh, D. P. Renfro, A. E. Zweifel, J. C. Hu, N. H. Brown, S. Tweedie, Y. Alam-Faruque, R. Apweiler, A. Auchinchloss, K. Axelsen, B. Bely, M.-C. Blatter, C. Bonilla, L. Bouguerleret, E. Boutet, L. Breuza, A. Bridge, W. M. Chan, G. Chavali, E. Coudert, E. Dimmer, A. Estreicher, L. Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, R. Hieta, C. Hinz, C. Hulo, R. Huntley, J. James, F. Jungo, G. Keller, K. Laiho, D. Legge, P. Lemercier, D. Lieberherr, M. Magrane, M. J. Martin, P. Masson, P. Mutowo-Muellenet, C. O'Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, P. Porras Millán, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, I. Xenarios, R. Foulgar, J. Lomax, P. Roncaglia, V. K. Khodiyar, R. C. Lovering, P. J. Talmud, M. Chibucos, M. G. Giglio, H.-Y. Chang, S. Hunter, C. McAnulla, A. Mitchell, A. Sangrador, R. Stephan, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P. J. Kersey, D. M. McDowall, D. M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, T. Hayman, S.-J. Wang, V. Petri, T. Lowry, P. D'Eustachio, L. Matthews, R. Balakrishnan, G. Binkley, J. M. Cherry, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, B. C. Hitz, E. L. Hong, K. Karra, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, S. Weng, E. D. Wong, T. Z. Berardini, E. Huala, H. Mi, P. D. Thomas, J. Chan, R. Kishore, P. Sternberg, K. Van Auken, D. Howe, and M. Westerfield, "Gene Ontology annotations and resources.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D530–5, Jan. 2013.

[5] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D109–14, Jan. 2012.

[6] EBML-EBI, "QuickGO." [Online]. Available: http://www.ebi.ac.uk/QuickGO/. [Accessed: 03-Oct-2013].

[7] M. Hucka, L. Smith, D. Wilkinson, F. Bergmann, S. Hoops, S. Keating, S. Sahle, and J. Schaff, "The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core," *Nat. Preced.*, Oct. 2010.

[8] A. Dräger, N. Rodriguez, M. Dumousseau, A. Dörr, C. Wrzodek, N. Le Novère, A. Zell, and M. Hucka, "JSBML: a flexible Java library for working with SBML.," *Bioinformatics*, vol. 27, no. 15, pp. 2167–8, Aug. 2011.

[9] O. Corporation, "Java™ Platform, Standard Edition 7 Development Kit." [Online]. Available: http://www.oracle.com/technetwork/java/javase/jdk-7-readme-429198.html. [Accessed: 01-Jun-2013].

[10] B. J. Bornstein, S. M. Keating, A. Jouraku, and M. Hucka, "LibSBML: an API library for SBML.," *Bioinformatics*, vol. 24, no. 6, pp. 880–1, Mar. 2008.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Biological system analysis using integrated bioinformatics tools:

Levels of the central dogma in Folate Mediate One-Carbon Metabolism (FOCM)

Jordy Larson*, Abalo Chango**, Bill Conn*, Carol M. Lushbough*, Etienne Z. Gnimpieba*
*Computer Science Dept., University of South Dakota, 414 E. Clark St. Vermillion, SD 57069, USA,
**UPSP EGEAL, Institut Polytechnique LaSalle Beauvais, 19 Rue Pierre Waguet, 6000 Beauvais
Contact : Etienne.Gnimpieba@usd.edu

~~o~~

Abstract - *Bioinformatics has allowed for vast amounts of data, tools, knowledge, and technology to be created and collected. Large volumes of information have presented a problem with interpreting and using the majority of it. Bioinformatics is continuing to look for ways to utilize, connect, and condense all of this information for researchers to enable them to easily use and manipulate their research data, whether that is through data mining, manipulation, or using any other type of bioinformatics applications. This project provides a sampling of the types of information that can be obtained using specific tools and applications with a focus on the folate one-carbon system and breast cancer. At each step of our workflow we generated several results to better understand the FOCM behavior, supporting the relationship between the FOCM and the breast cancer disease, and new biological hypotheses that require experimental validations.*

Keywords: FOCM, Folate One-Carbon Metabolism, Bioinformatics, Inter-omics, Central Dogma, Databases

1 Introduction

With the growth of scientific quantitative and digital data, the problem is not so much obtaining the information, but analyzing it to extract new knowledge. This project provides an example of the type of information that can be obtained using bioinformatics tools and applications. We focused on the folate one-carbon metabolism (FOCM) system and the breast cancer development on all three levels of the central dogma; metabolomics, proteomics, and genomics.

Folate is the carbon donor in the one carbon (methyl) metabolism pathway. This molecule is better known as one of the water soluble B vitamins [2]. One of the functions of this folate and methionine metabolism is that it synthesizes the nucleotide bases that comprises the genome [4]. There have been multitudes of evidence that this metabolism is linked to an increased risk in cancer [1]. The folate receptors are known to be up-regulated by cancer cells in order to accelerate the production of nucleotides and support DNA synthesis and cell growth [3]. A particular link that has already been made is that breast cancer risk is inversely related to natural food folate intake [7]. However research methods used to investigate these links are generally focused in a single "omics" area; for example many paper focus solely on the metabolomics or genomics link between FOCM and breast cancer. In this study, however, we will be applying a much larger scale approach by using computational biology techniques. By utilizing various bioinformatics techniques and resources we were able to gain a bigger picture on the connections between FOCM and breast cancer as wells as consolidate information from similar experiments held in a multitude of databases.

This research provides a new workflow (set of steps) to connect the folate one-carbon metabolism through the different biological levels of the central dogma using an assortment of bioinformatics tools available online. The central dogma is the concept of the molecular thesis of inheritance (DNA to RNA, and RNA to Protein). The genomic and proteomic levels of FOCM were also used to find any links to a given biological disease (breast cancer) in order to show the tools available for finding connections between a specific metabolism or genome and a disease. By doing so this study will provide an example of how to thoroughly research a specific pathway as well as connect it to a specific disease. This work describes a handful of tools that can be utilized using an assortment of free online bioinformatics tools and applications available to any interested researcher.

2 Method

A large scale *in silico* approach had been taken in this thesis project by mining data from previously published research and utilizing various online bioinformatics tools. This process will be divided into

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

different biological areas transport of target enzymes into a cell, metabolic interactions within the cell, and proteomic interactions within a cell (**Fig 1**). These various levels will then be compared with genes, proteins, and metabolites related to breast cancer. Data were then collected and models created with each step [5]. At each step of the workflow we will come up with results to better understand the folate one-carbon pathway behavior (folate deficiency in DNA methylation) and to support the relationship between the folate one-carbon pathway and breast cancer (protein-protein interactions). This research should also lead to new biological hypotheses that require experimental validations.

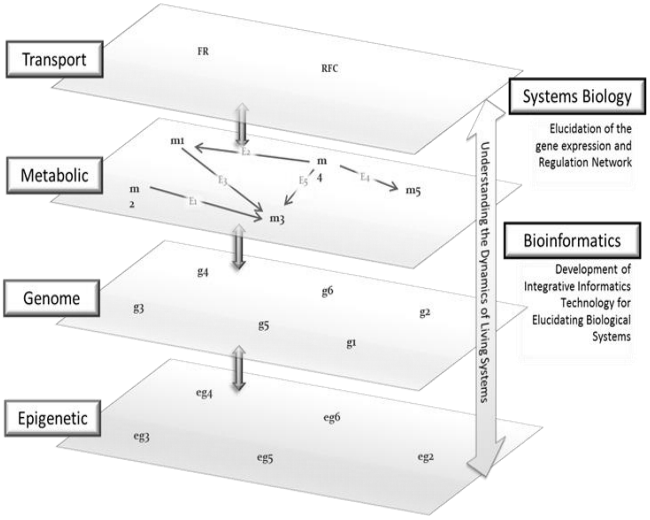


Fig 1. Biological level for workflow design based on systems integration (numerous biological systems) and data integration (numerous data types)

3 Application

Multiple databases were queried to extract each of the metabolites in the FOCM and its structural and functional information. The resulting data have been use to generate the FOCM mathematical metabolism network and pathogenesis, SMPDB metabolic pathway model, the breast cancer protein network model, and the Gene Expression Omnibus (GEO) breast cancer protein network model for analysis. At each step of our workflow, we provide good quality parameter for our result reliability (p-value, score).

For the metabolomics level, metabolomics databases (KEGG, BRENDA, SMPDB, REACTOME, SABIO-RK, BIOMODELS) were queried with the metabolites names to extract information on each metabolites in the folate one-carbon pathway; this includes reactions, enzymes, enzymes kinetics, enzyme parameters, and references. Results were used to design visual models of the folate one-carbon metabolism using SMPDB notations.

For the proteomics level, proteomics databases are queried to extract protein structural and functional information (PDB, SBKB, EXPASY) using the various names of the metabolites found in the FOCM (using reverse engineering technique). The results were used to build a protein-protein network using the STRING database and data learning tools.

On the genomics level, the folate one-carbon metabolism's gene name list was used to query existing profiles on Expression Atlas. Using Array Express we identified breast cancer datasets which hold information on experiments performed previously. Useful information obtained in those studies with acceptable quality measurements were compiled and analyzed in order to confirm additional links between the FOCM and breast cancer.

4 Results

This section is split up into the metabolic level, proteomic level, and genomic level for the results found in each level of the central dogma.

4.1 Metabolic Level

Metabolomics databases (KEGG, BRENDA, SMPDB, REACTOME, SABIO-RK, BIOMODELS) were queried to extract information on metabolite, pathways, reactions, enzymes, enzymes kinetics, enzyme parameters, and references. Multiple names and abbreviations for same metabolites were compiled and used for thorough querying. Results were then used to design the biological metabolism of FOCM using SMPDB notations (**Fig 2**, **Fig 3**).

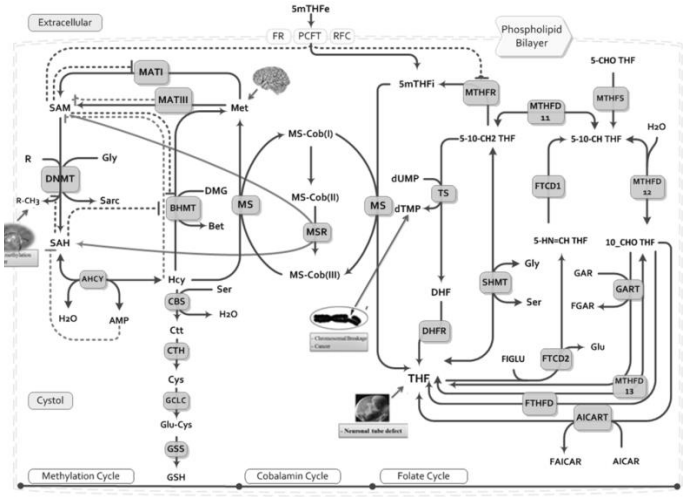


Fig 2 Folate Mediate One Carbon Metabolism (FOCM) network model.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

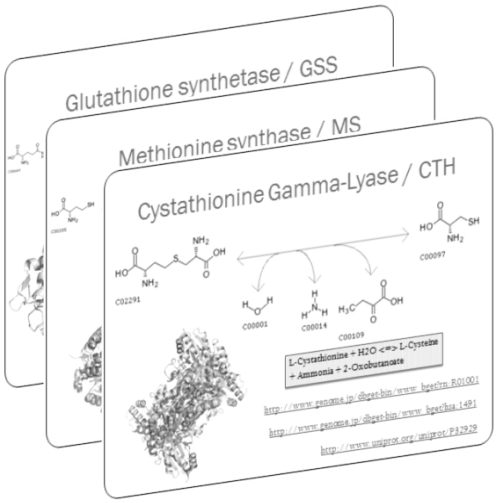


Fig 3 Sample of slides that compiled information on each metabolite

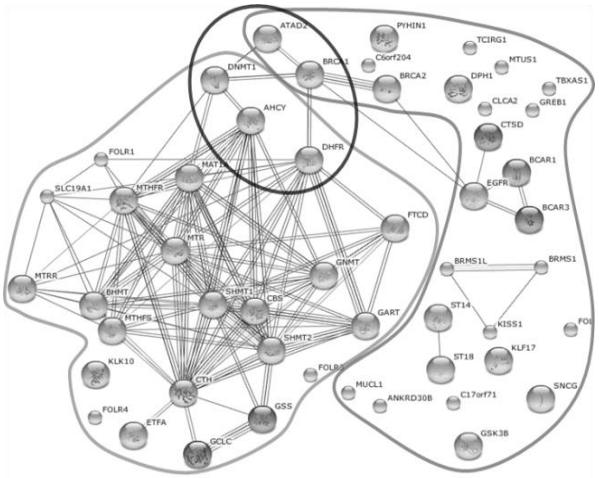


Fig 4 This models from STRING displays more than 80 protein interactions in FOCM, 30 in Breast Cancer and 8 interactions connecting the FOCM and Breast

The mathematical model in (Gnimpieba et al., 2011) [6] was redesigned using graph theory and Systems Biology Markup Language tools (SBML). Using Semantic SBML, we found 3 high score similar models in BIOMODELS database with score>0.1 and p-value<=1e-3. Additional results demonstrates how mathematical models could be involved in a given biological hypothesis studies using simulation tools (MatLab, Virtual Cell, Biochem).

4.2 Proteomic Level

Proteomics databases were queried with known FOCM metabolites and their various names to extract protein structural and functional information (PDB, SBKB, EXPASY). The resulting data was then used to build the protein-protein network using the STRING database and data learning tools (**Fig 4**). The following figures present several results including 18 over 33 structural models for FOCM enzymes, more than 80 protein interactions in FOCM, 30 in Breast Cancer and 8 interactions connecting the FOCM and breast cancer. Protein expression data analysis confirmed several known regulations that are linked to the breast cancer disease (MTHFR-MAT. CBS-AHCY) and provided new regulation hypotheses for possible additional connections (GART-{DNMT1, AHCY, MSR, CBS} or Folate regulating DNA methylation). The Protein's phylogenic tree was also obtained on STRING.

4.3 Genomic Level

On the genomics level, the FOCM gene name list was used to query existing profiles on Expression Atlas. Using Array Express we identified Breast Cancer datasets which hold information on experiments performed previously as well as its results so as to be used again or referenced by others. We chose the experiment, GSE36683 - Gene Regulation by Estrogen Signaling and DNA Methylation in MCF7 Breast Cancer Cells. We then went to the given Gene Expression Omnibus (GEO) link to analyze the data using GEO2R. This experiment was selected since the value distribution obtained was acceptable (close to the same values in all results and close to zero). Groups were defined according to the treatments given to the samples; DAC and ETOH treatment; E2 treatment; ETOH treatment only. The gene DNMT was then queried within the results and the gene profile was obtained). The P value was also obtained for various DNMT genes. Gene-Gene interactions were collected and modeled using GENEMANIA by entering the known gene names involved in FOCM (**Fig 5**).

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

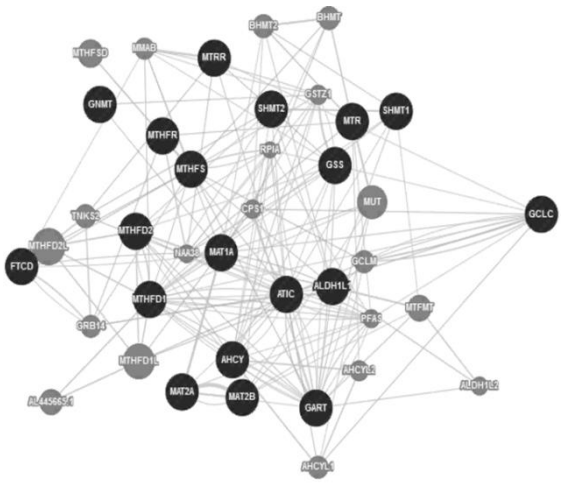


Fig 5 Gene interactions for the FOCM genes using the GENEMANIA tool.

5 Conclusion

Bioinformatics and computational biology tools development presents innovative opportunities to improve life science research and contribute in its numerous challenges. This work presents a workflow for bioinformatics and systems biology tools usage in inter-OMICS (metabolomics, proteomics, genomics) studying of a given biological systems. This workflow has been applied to the study of the Folate Mediate One Carbone Metabolism (FOCM) as well as its relation to breast cancer. At each step of our workflow we came up with several results to better understand the FOCM behavior (folate deficiency in DNA methylation), supporting the relationship between the FOCM and breast cancer (protein-protein interaction such as BRCA1-DHFR, ATAD2-DNMT1, BRCA1-DNMT1), and new biological hypotheses that requires experimental validations (such as GART can control the amino acid cycle by regulating AHCY-CBS-DNMT1, GART which may play an important role in breast cancer through the regulation of BRCA1-DNMT and Parkinson Disease through CBS regulation). This work has some limitations because it was performed exclusively using the *in silico* approach. However it does demonstrate the applications of multiple bioinformatics tools throughout the different levels of the central dogma that are available free online. This is a valuable ability when trying to look at the larger picture of a specific gene, protein or metabolic process and its connection to a certain disease.

6 References

[1] C. M. Ulrich, "Folate and cancer prevention: a closer look at a complex picture.," *Am. J. Clin. Nutr.*, vol. 86, no. 2, pp. 271–3, Aug. 2007.

[2] J. B. Mason, "Biomarkers of nutrient exposure and status in one-carbon (methyl) metabolism.," *J. Nutr.*, vol. 133 Suppl, p. 941S–947S, Mar. 2003.

[3] J. F. Ross, P. K. Chaudhuri, and M. Ratnam, "Differential regulation of folate receptor isoforms in normal and malignant tissues in vivo and in established cell lines. Physiologic and clinical implications," *Cancer*, vol. 73, no. 9, pp. 2432–2443, May 1994.

[4] J. W. Locasale, "Serine, glycine and one-carbon units: cancer metabolism in full circle.," *Nat. Rev. Cancer*, vol. 13, no. 8, pp. 572–83, Aug. 2013.

[5] E. Z. Gnimpieba, B. S. Anderson, A. Chango, and C. M. Lushbough, "RESTful API in life science research systems and data integration challenges : linking metabolic pathway , metabolic network , gene and publication .," *Jopurnal Comput. Commun. ISSN 1930-1553*, no. 10, p. xx, 2013.

[6] E. Z. Gnimpieba, D. Eveillard, J.-L. Guéant, and A. Chango, "Using logic programming for modeling the one-carbon metabolism network to study the impact of folate deficiency on methylation processes.," *Mol. Biosyst.*, vol. 7, no. 8, pp. 2508–21, Aug. 2011.

[7] Z. Gong, C. B. Ambrosone, S. E. McCann, G. Zirpoli, U. Chandran, C.-C. Hong, D. H. Bovbjerg, L. Jandorf, G. Ciupak, K. Pawlish, Q. Lu, H. Hwang, T. Khoury, B. Wiam, and E. V Bandera, "Associations of dietary folate, Vitamins B6 and B12 and methionine intake with risk of breast cancer among African American and European American women.," *Int. J. Cancer*, vol. 134, no. 6, pp. 1422–35, Mar. 2014.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Discovering Association Rules and Classification for Biological Data using Data Mining Methods

J. Tsiligaridis[1], **M. Pagela**[2]
[1,2]Math & Computer Science Department, Heritage University , Toppenish, USA

Abstract - *This project presents a set of algorithms and their efficiency for discovering association rules and classification using Genetics Algorithm(GA), Decision Trees (DT) and Neural Networks (NN).*
A GA generates a large set of possible solutions to a given problem. Apriori is the basic algorithm for association rules. A GA is developed for finding the frequent conditions. The proposed GA based on encoding and generation construction method (GA_EN) can mine association rules with improved performance using appropriate generation of the rules. For GA classification (GA_CL) algorithm, rules are classified using predefined constraints. A Decision Tree algorithm (DTA) is created from data using probabilities, and the goal is to create on-demand an accurate decision tree (DT).
Based on the rules produced from GA_CL, a Neural Network classifier (NNC_GA) is created. For learning a backpropagation neural network algorithm is used to adjust the weights. Simulation results are provided.

Keywords: Genetic Algorithm, Decision Trees, Neural Network, Data Mining

1 Introduction

GA is based on biological principles of natural selection. The key idea of Apriori is to find the frequent conditions constructed by possible values of attributes in any data set. This idea can also be used to find the frequent conditions constructed by possible values of attributes in any
data set. The GA finds all the possible associations between conditions constructed by attribute values under a given constraint (e.g. support and confidence). The association rules mining, integrated with classification, creates the association classification [1],[2],[3]. The objective of Classification Association rules (CAR) is to generate a set of class association rules that satisfy the min support (msup), and minimum confidence (mconf) constraints and to build a classifier from the class association rule set. The GA_EN, based on an encoding method and construction of generations has advantages over the Apriori because it includes GA mining techniques that improve performance. The GA_CL discovers rules with minimum class support (mcsupp) and less than the maximum classification error (maxclerror). There are two types of a DT [3]; the complete and the incomplete. In the incomplete one there are subtrees where the repetition and replication are included. A DT represents a procedure for classifying objects based on their attributes. The rule set can be created by running the tree. Decision tree is used to find predictive rules combining numeric and categorical attributes. The splitting process is recursively repeated until the end of data. The DTA creates a DT using the criterion of maximum probability. There are two phases. In order to avoid repetition or replication of subtrees a new criterion, the elimination of a branch (CEB), is applied.
The criterion of elimination a branch (CEB), eliminates redundant branches. The pruning of decision rules case is also examined with consequences on the accuracy.
A NN is a collection of units that are connected in some pattern to allow communication between the units. The back propagation algorithm is in wide use because it learns by adapting its weight using the generalized delta rule which attempts to minimize the squared error between what is the desired network output and the actual network output.
The NNC_GA is a classifier using NN methodology and having as input the rules created from GA_CL.
The article is organized as follows. Section 2 ,3 introduces definitions for association rules and CAR. Section 4 deals with GA_EN and GA_CL. Section 5 includes the description of DTA. Section 6 contains the NN and NNC_GA. Simulation results appear in Section 7.

2 Association rules

Let D = {T1, T2, . . . ,Tn} be a set of n transactions and let I be a set of items, I = {i1, i2 . . . im}. Each transaction [5] is a set of items, i.e. Ti $\subseteq$ I. An association rule is an implication of the form X $\Rightarrow$ Y, where X, Y $\subset$ I, and X $\cap$ Y = $\varnothing$; X is called the antecedent and Y is called the consequent of the rule. In general, a set of items, such as X or Y , is called an itemset. For an itemset X $\subseteq$ I, support(X) is defined as the fraction of transactions Ti $\in$ D such that X $\subseteq$ Ti. That is, P(X) = support(X). The support of a rule X $\Rightarrow$ Y is defined as support(X $\Rightarrow$ Y) = P(X $\cup$ Y). An association rule X $\Rightarrow$ Y has a measure of reliability called confidence(X $\Rightarrow$ Y) defined as P(Y |X) = P(X $\cup$ Y)/P (X) = support(X $\cup$ Y)/support(X). For the CAR it is supposed that data samples are given with n attributes (A1,A2,..,An) and for each sample there is a class label C (C={c1,c2,..ccm}). A pattern P (P={a1,a2,..,ak})is a set of attribute values for different attributes (1≤k≤n). For rule R: P-> c the number of data samples matching pattern P with class label c is called the support of rule R. The ratio of the number of samples

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

matching pattern P and having class label c versus the total number of samples matching pattern P is called confidence of R.

3 CAR

The CAR[3] contains methods for associative classification. CBA is one of the methods. It uses an iterative approach to frequent itemset mining which is similar to Apriori. CBA construct the classifier where the rules are ordered according to the precedence based on the rule confidence and support. More details appear in [3].

4 GA_EN ,GA_CL

A GA is an iterative procedure, which works with a *population* of individuals represented by a finite string of characters or binary values. The traditional method usually searches a very large space for the solution using a large number of iterations, where a GA minimizes searches by adopting a fitness function. Each iteration consists of a created evolved population of genomes and a new *generation*. There are three operations: selection, crossover, mutation. For GA_EN, the number of conditions determines the construction of the chromosome and the population size. The next generations can be created from other either one or two previous generations using the "or" operation for crossover of chromosomes. This provides less memory operations and fewer offspring chromosomes. The msupp and mconf is the constraint set by user. The GA_CL working with the next generations under conditions and the crossover of the chromosomes and considering the predefined classification error can discover the classification rules. The conditions: msup and maxclerror reduces the number of undesired rules in the mining process. For the GA_CL only the rules from chromosomes are extracted which the classification error does not surpass the maxclerror. The next generation includes initially chromosomes with support greater than the minsup (gr_mins_chrom). The 'or' operation among gr_mins_chrom can create the chromosomes of the next generation. Chromosomes with class error less than the maxclerror can provide the classification rule. There are no classification rules for any attribute if there are no chromosomes with acceptance classification error. The GA mines rules of our interest by defining the msupp and mconf . With low constraints the number of discovered association rules becomes extremely large.

5 DTA

In decision trees [4],[5] the input data set has one attribute called class C that takes a value from K discrete values 1, . . . , K, and a set of numeric and categorical attributesA1, .. . , Ap. The goal is to predict C given A1, . . . , Ap. Decision tree algorithms automatically split numeric attributes Ai into two ranges and they split categorical attributes Aj into two subsets at each node. Split the records based on an attribute test that optimizes certain criterion (Greedy strategy). The multi-way split is used, where as many partitions as distinct values. Nodes with homogeneous class distribution are preferred.

In the incomplete DT there are subtrees where the repetition and replication are included. Repetition is where an attribute is repeatedly tested along a given branch of three , e.i. age, and replication where duplicate subtrees exist within a tree , such as the subtree headed by the node "credit_rating". The DTA uses the criterion of maximum probability and can be created in the following phases:

Phase 1: Discover the root (i) (from all the attributes)

$$P(E_{Ai}) = \sum_{Ci} \sum_{Ai} p(Ai) * p(Ci / Ai) \quad ,$$

where Ai : the attributes of the tuples and Ci the classes (attribute test). MP = max (P(EAi)) //max attribute test criterion

Phase 2: split the data into smaller subsets, so that the partition to be as pure as possible using the same formula. The measure of nodes impurity is the MP. Continue until the end of the attributes.

There is a stopping criterion for expanding a node when all records belong to the same class.

DTA : Input : training data
Output: decision tree
1. define root node (phase 1)
2. discover the branches from root
while (! end of the attributes)
{3. splitting the attribute (phase 2) }

Example:

Weather	parents	money	decision (Example)
Sunny	yes	rich	cinema
Sunny	no	rich	Tennis
…..			
Windy	no	rich	cinema
…..			

yes no

Parents: class, P(E) = (5/10)*(5/10) + (5/10)*(1/10) = 0.3 (phase 1)

Weather: class, P(E) = (3/10)*(1/10) + (4/10)*(3/10) + (3/10)*(2/10) = 0.21

The CEB, is used to eliminate redundant branches. It is a prepruning approach [3].

For an attribute (attr1) with value v1 , if there are tuples from attr2 that have all the values in relation with v1 (of attr1) then the attr2 is named as: *do n't care* attribute Example: R1.

$P_{CEB} = P(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = ci) = \prod_{i=1}^{|A|} p(A_i = a_i \mid C = c_j)$

A branch is eliminated when the $P_{CEB} \neq 0$

The criterion of Elimination of Branch (CEB):: if the $P_{CEB} = 0$, between two attributes (A1, A2) then A2 is don't care attribute. The CEB criterion is valid when $P_{CEB} \neq 0$. CEB

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

is to develop the DT so that to avoid the repetition or replication.

Theorem: The CEB criterion can determine the existence of a small DT with the best accuracy (100%, or complete) avoiding repetitions and replications. Proof: Because if CEB criterion is valid discourage the repetition

Example:

Age	Has_job	Own-house	Credit-rating	class (Example)
Young	false	False	fair	No
Young	false	False	Good	No
Young	True	False	Good	Yes
Young	True	True	fair	Yes
Middle	True	True	Good	Yes
Old	false	true	Excellent	Yes

……..

Of the DT (with own_house as a root) it is not necessary to have more extension with the attribute of age for all the probable partitions ("young"," middle", "old"). P_{CEB} = P(A1 =a1,…, A|A| = a|A| | C=ci) = P(age =young , own_house="y" | C="yes") =P(age =young | C="yes") * P(own_house="y" | C="yes") = 2/5 * 6/9 ≠ 0

DTs provide less rules than the CAR. DT can find rules with very low support (like medical rules). CAR requires discrete attributes. DT learning uses discrete and continues attributes.

The ID3 is a method for discovering DT and uses the information gain as its attribute selection measure [3].

6 Neural Network

For supervising learning it is necessary to have: data that have a known classification, sufficient data to represent all aspects of the problem beign solved, sufficient data to allow for testing. The back propagation algorithm learns by adapting its weight using the generalized delta rule which attempts to minimize the squared error between what is the desired network output and the actual network output. During learning it continually cycles through the data until the error is at a low enough value for the task to be considered solved.

There are two activation functions one from input layer to hidden layer and the other from hidden layer to output layer. Both of them are the logistic functions.

Example: for tennis participation with tuples: "outlook , temperature, humidity, windy, class". The distributed coding has been used. The logistic function is applied to hidden layer and output layer. A set of five input processing units and five hidden layer units has been used. The weight matrix is 5x5 dimension with bias input z0 =1. Parameters are: m: #of input vectors with length of 5 (including bias input). Array x[i] is input values , d[i] the desired output for each input x[i]. Two activation (logistic) functions fh (input layer to hidden layer) and fo (hidden layer to output layer) are assumed. Test of convergence is achieved by checking the output error function to see if its magnitude is below some given threshold. Some disadvantage of NN: (a) Does not explain the solution is derived, need more examples, need appropriate examples that matches the real world situation. (b) it can be a lengthy and computationally expensive process to train a NN using a large number of high dimensional training examples. The advantage of DT over NN is that it will require much less training time than NN.

If we compare decision trees and neural networks we can see that their advantages and drawbacks are almost complementary. For instance humans easily understand knowledge representation of decision trees, which is not the case for neural networks. Decision trees have trouble dealing with noise in training data, which is again not the case for neural networks, decision trees learn very fast and neural networks learn relatively slow, etc.

The classification rules created by GA_CL will be used as input to the construction of the NNC_GA. The NNC_GA architecture has three layers. First, the input layer that has the input nodes where each one is represented by one characteristic of the rules. Second , the hidden layer in which each node will be connected with the characteristic (the attribute value in the rule is called characteristic) of each rule. The number of rules is equal to the number of hidden nodes.

Third, the output layer , that has the classes nodes (i.e. c1,..cn). In the NNC_GA for learning the input vector can be created in binary format by considering '1' for activated input node and '0' for the non activated one. The gradient descent would proceed in infinitesimal steps along the direction established by the gradient. For the learning rate, n., is selected large enough to cause the network to converge quickly without oscillations.

7 Simulation

Two are the scenarios for the simulation

1. Apriori vs GA_EN: The GA_EN using the particular way to perform the next generations has better performance than the Apriori for the mining of the rules of the "heart" data set.

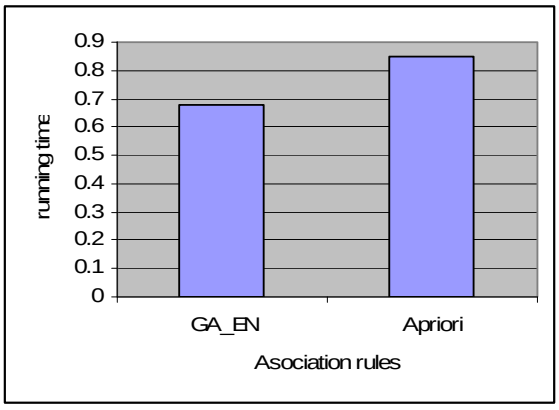


Fig.1 GA_EN vs Apriori

2. CBA vs NNC_GA: Using data set "hepatitis" there is better accuracy for NNC_GA because NNC_GA follows the NN method for learning the process and adjusting the weight.

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

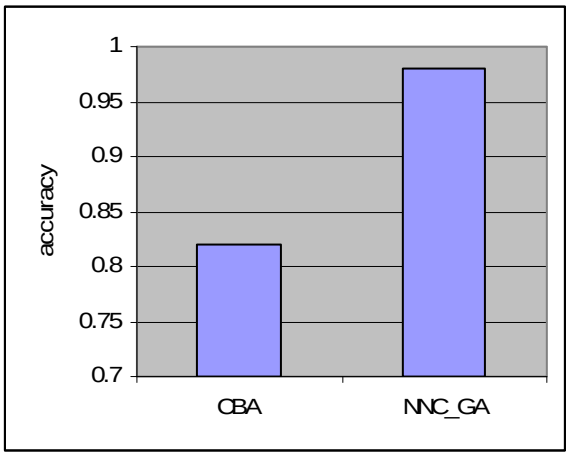


Fig. 2 CBA vs NNC_GA

4. ID3 vs DTA: using the data set "iris". DTA results are slightly better than the ones of ID3.

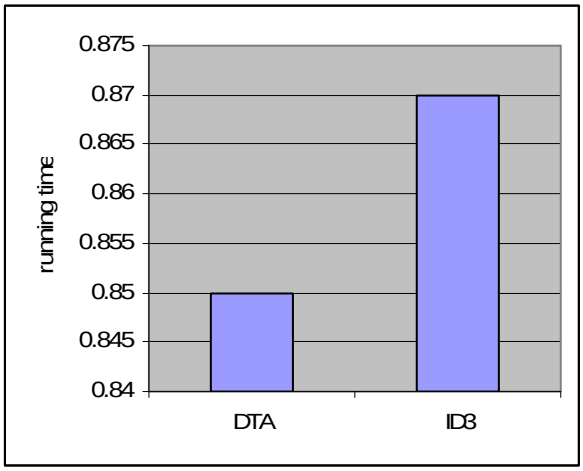


Fig. 3 ID3 vs DTA

8 Conclusions

In this project a new framework of algorithms is developed based on the ability of Genetic Algorithms for discovering Association rules and Classification on biological data.

Certain advantages apply depending on the algorithms' particular way of operation. Future work will focus on NN.

9 References

[1] B. Bringmann, S. Nijssen and A. Zimmermann , "Pattern-based classification: a unifying perspective", In Proceedings of ECML-PKDD workshop on from local patterns to global models, pp. 36-50, 2009

[2] W. Li, J. Han and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules", *In Data Mining '01*, Proceedings IEEE International Conference on, pp. 369-376, Nov. 2001.

[3] J. Han, ,M. Kamber, J. Pei, " Data Mining Concepts and Techniques", Morgan Kaufman, 3 ed, 2012

[4] U. Fayyad and G. Piateski-Shapiro, " From Data Mining to Knowledge Discovery. MIT Press, 1995.

[5] C. Ordonez, "Comparing Association Rules and Decision Trees for Disease Prediction", HIKM 2006, Nov 11, Virginia.

[6] M. Karntardzic, "Data Mining: Concepts, Models, Methods, and Algorithms", IEEE Press, 2003

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America

Copyright © 2014 CSREA Press, ISBN: 1-60132-265-8; Printed in the United States of America