

SESSION

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

Chair(s)

TBA

Integrative modeling of transcriptional regulatory networks in head and neck cancer

Bin Yan^{1*}, Huai Li², Zhong Chen³, Jiaofang Shao¹, Ming Zhan^{4*}

¹Department of Biology, Hong Kong Baptist University, Kln, Hong Kong; ²RRB, NIA, National Institutes of Health, Baltimore, MD, USA; ³Head and Neck Surgery Branch, NIDCD, National Institutes of Health, Bethesda, MD, USA ; ⁴Department of Systems Medicine and Bioengineering, Methodist Hospital Research Institute, Houston, TX, USA.

*Contact author, BY: bin1999@hkbu.edu.hk, MZ: mzhan@tmhs.org

Abstract *p53 is the most mutated tumor suppressor gene in cancers, which are usually inflammatory with aberrant NF- κ B activation. However, how NF- κ B family members and p53 interact to globally regulate genes expression is not yet fully understood. Using head and neck squamous cell carcinoma (HNSCC) lines as the model system, we developed a novel integrative model based on Regulatory Component Analysis, which combined mRNA expression profile with transcription factor and microRNA binding for integrated analyses through matrix decomposition. We observed that the majority of p53 targets are also co-regulated by NF- κ B in p53 wild-type or mutant subset of HNSCC cells. We further constructed regulatory networks of NF- κ B, p53 and microRNAs 21 and 34s. Our results unraveled the cross-regulations among NF- κ B, p53, and microRNAs, provided an insight into understanding of underlying regulatory mechanisms, and showed an efficient approach to inferring the regulatory programs in these datasets.*

Keywords: Regulatory networks, Integrative modeling, NF- κ B, p53, Head and neck cancer

1 Introduction

Reconstruction and modeling of gene regulatory networks are one of main challenges in computational biology. Various mathematical algorithms or computational methods have been developed for integrative analysis of microarray and transcription factor (TF) binding data for unraveling transcriptional regulatory modules. Several matrix decomposition methods, such as PSMF, ModulePro, NMF, have been recently presented for regulatory network reconstruction based on the constraints of sparseness, non-negativeness, or partial network connectivity information [1-3]. Although all these methods show an improved result in uncovering biologically meaningful regulatory networks than the decomposition methods without the constraints, they were conducted separately, and no integrative framework has been utilized that brings the sparseness and pre-knowledge of regulator-

target interactions together during matrix decomposition [1, 2, 4]. Here, we devised a new methodology, based on Regulatory Component Analysis (RCA), for inferring regulatory gene networks and uncovering transcriptional modules. The RCA-based model performs matrix decomposition under the joint constraints of sparseness and partial information of TF-target connectivity, and allows an integrated analysis of gene expression profile and regulator binding data. We used the new method in studies of the head and neck squamous cell carcinoma (HNSCC).

HNSCC is one of the most common human cancer worldwide. The development of HNSCC is associated with alterations in expression of a large set of genes, which could be underlined by shared TFs or regulators of key regulatory mechanisms that control transcriptional regulatory networks. Among these TFs, NF- κ B has been demonstrated to play a central role in the control of gene expression that mediates cellular proliferation, apoptosis, angiogenesis, immune and proinflammatory responses, and therapeutic resistance [5, 6]. In a systems biology study, we defined 748 NF- κ B target genes and their functional associations using an integrative model COGRIM in HNSCC, thus proposing that NF- κ B is one of the critical regulatory determinants of expression of multiple gene programs, interacting pathways and malignant phenotypes [7]. Followed by that study, some challenging questions would be further asked with regard to NF- κ B regulatory mechanisms. For example, whether NF- κ B functions are affected by other TFs or regulators? If so, how NF- κ B interacts with these TFs or regulators to modulate the gene programs of HNSCC? As a tumor suppressor, p53 is implicated as a master regulator of apoptosis, cell cycle and DNA repair, etc. Mutations of *TP53* have been observed in near half cases for all types of human cancer, including HNSCC. In previous studies, tumor suppressor TF p53 was reported to also regulate NF- κ B target genes [8-10]. However, the molecular basis underlying their interactions has not been adequately understood in HNSCC. In addition to p53, other cancer-related TFs such as AP1, STAT3, EGR1, CEBPB and SP1 were reported to be involved in

complex regulatory systems of NF- κ B [11-14]. Moreover, microRNAs are proposed to play as co-regulators that involve in modulation of gene expression at post-transcriptional level. Therefore, genome-wide investigation of significant interactions between NF- κ B and p53 or other regulators would enhance our understanding of transcriptional regulatory mechanisms associated with diverse HNSCC phenotypes.

In this study, we applied a newly developed method to identify transcriptional regulatory programs by combining TF and microRNA binding information with expression profiling in HNSCC cells with the wild type (wt) p53-deficient and the mutant (mt) p53 status. Our studies demonstrated that two master TFs NF- κ B and p53 have a wide impact on expression profile of gene programs in the tumor cells. Furthermore, our results revealed that NF- κ B, p53 and the microRNAs may form concerted regulatory modules for contributing to the gene programs in both wt p53-deficient and mt p53 phenotypes.

2 Methods

2.1 Microarray dataset

The gene expression data was collected from GEO database <http://www.ncbi.nlm.nih.gov/geo/> with an accession number GSE10774. The microarray data were generated from two subgroups: mt p53 and wt p53-deficient HNSCC cell lines [7, 15]. The microarray data of differentially expressed genes satisfying 2.0 fold and above change were used for the RCA-based analyses.

2.2 TF and microRNA binding data

NF- κ B and p53 binding data were extracted from available sources: 1) NF- κ B and p53 websites 2) previous publications. The binding information of other TFs (AP1, EGR1, CEBPB, STAT3, and SP1) and microRNAs was gained from 1) website <http://www.broadinstitute.org/gsea/msigdb/>; 2) curated from previous publications.

2.3 RCA-based method

The method was performing matrix decomposition under the joint constraints of sparseness and partial information of TF-target connectivity. The method allows an integrated analysis of gene expression profiles with binding data of a set of regulators, including TFs, microRNAs, etc.

The RCA-based method (see Figure 1) is a network structure-driven model for inferring gene regulatory networks and uncovering transcriptional modules. Given a microarray data matrix $\mathbf{X} \in \mathfrak{R}^{N \times M}$ with the sample size M and the numbers of genes N , our aim is to find $\mathbf{Y} \in \mathfrak{R}^{N \times L}$

and $\mathbf{Z} \in \mathfrak{R}^{L \times M}$ such that the square error (Euclidean distance) function:

$$E(\mathbf{Y}, \mathbf{Z}) = \|\mathbf{X} - \mathbf{YZ}\|^2 \quad (1)$$

is minimized under a desired degree of sparseness on the mixing matrix \mathbf{Y} . We defined a sparseness measure $\mathbf{S}(\mathbf{y}_l)$ based on the relationship between the L_1 norm and the L_2 norm [16]:

$$\mathbf{S}(\mathbf{y}_l) = \frac{\sqrt{N} - \sum_{i=1}^N |y_{il}| / \sqrt{\sum_{i=1}^N y_{il}^2}}{\sqrt{N} - 1} \quad (2)$$

where $\mathbf{y}_l = [y_{1l} y_{2l} \dots y_{Nl}]^T$ is the l th column of \mathbf{Y} , the superscript "T" means "transpose". The L_1 norm $\|\mathbf{y}_l\|_1$ and the L_2 norm $\|\mathbf{y}_l\|_2$ were defined as $\|\mathbf{y}_l\|_1 = \sum_{i=1}^N |y_{il}|$ and $\|\mathbf{y}_l\|_2 = \sqrt{\sum_{i=1}^N y_{il}^2}$, respectively. The sparseness evaluates to one if and only if \mathbf{y}_l contains a single non-zero element, and takes a value of zero if and only if all elements are equal. Here, there are two interpretations of the decomposition $\mathbf{X} \approx \mathbf{YZ}$. First, the rows of \mathbf{Z} represent the expression profiles of the L latent variables across samples. Second, the rows of \mathbf{Z} can be viewed as the activity profiles of the L regulators. Thus, we can cluster genes based on corresponding non-zero coefficients of \mathbf{Y} , which represent gene regulatory programs, i.e. transcriptional modules that are co-regulated by the L regulators.

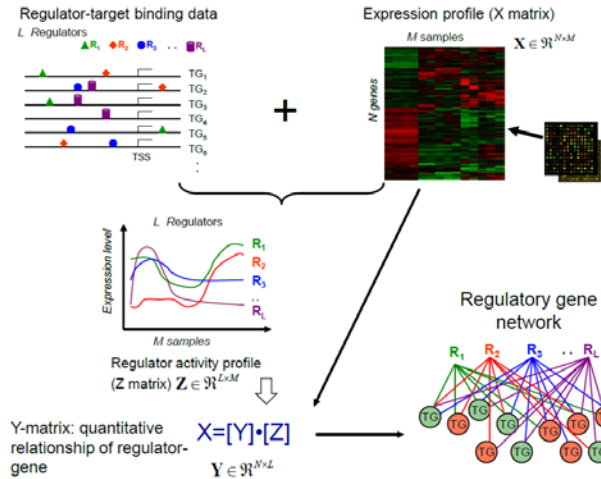


Figure 1. Overview of the Regulatory Component Analysis (RCA)-based method. R: regulators, such as transcription factors (TFs) and microRNAs. TG: target genes. TSS: transcriptional start site.

We devised an iterated learning algorithm that is capable of combining constraints of sparseness and limited information of regulator-target binding. The sparseness was used as a statistical parameter for modeling the regulatory components of regulators and their targets. The learning procedure was based on a projected gradient descent approach with sparseness constraints. The output (Y matrix) of the RCA procedure provided quantitative relationships between regulators (such as TFs and microRNAs in this study) and every gene from microarray dataset. The non-zero values stand for regulatory interactions or components which can be used to estimate how possible a gene is regulated by the regulators or whether a gene is target of the regulators.

3 Results

3.1 The RCA-based approach

In this study, we sought to unravel both TF- and microRNA-mediated regulatory gene networks responsible for the malignant phenotypes of HNSCC, and the analytic strategy is depicted in Figure 1. The integrative model exhibited several advantages in comparison with the COGRIM method, that was previously used in the same HNSCC gene profiling [7]. To justify our new method, we compared the results derived by using the RCA-based method in current study with those by COGRIM previously. The comparison was based on a Gene Ontology (GO) analysis of the target genes of the three NF- κ B subunits, RelA, NF κ B1 and cRel, predicted by using RCA and COGRIM, respectively, based on the same microarray dataset. We assessed the functional relevance of GO biological processes based on the enrichment analysis by Fisher's exact tests. Table 1 shows the statistical enrichment of biological processes among the target genes identified by the two methods. The enrichment level was calculated by transforming the enrichment P values after

Table 1. Comparison based on GO functional enrichment

HNSCC type	TFs	FDR by different methods	
		RCA	COGRIM
wt p53-deficient	RelA	1.92	1.81
	NF κ B1	1.66	1.70
	cRel	1.59	1.08
	Average of TFs	1.72	1.53
mt p53	RelA	1.80	0.89
	NF κ B1	1.65	1.55
	cRel	1.74	1.42
	Average of TFs	1.73	1.29

The enrichment level was calculated by transforming enrichment P values averaged over all GO processes with False Discovery Rate (FDR) corrected $P < 0.05$.

FDR correction to negative log10 values and averaged over all biological processes with corrected $P < 0.05$. Overall, our RCA-method showed advantages than COGRIM, where the averaged P values of FDR values were lower than COGRIM in both wt p53-deficient and the mt p53 datasets.

3.2 Prediction of HNSCC-specific target genes of TFs

Next, we intent to identify TF and microRNA regulatory modules controlling different gene expression programs in both malignant subgroups. Our analysis identified 248 and 418 target genes of NF- κ B, and putative 169 and 81 p53 target genes in the wt p53-deficient and mt p53 HNSCC cells, respectively. Then significant overlaps of target genes between NF- κ B and p53 was detected (Figure 2), that all p53 target genes predicted in the mt p53 cells overlapped with NF- κ B targets, whereas such overlap in the wt p53-deficient cells was 60% (overlapping P value = 2.56×10^{-18}). On the other hand, we noted that the fraction of the NF- κ B target genes that overlapped with p53 targets seemed different between the wt and mt p53 subgroups. Among the total NF- κ B target genes, 41% overlapped with the p53 targets in the wt p53-deficient, which was greater than those in the mt p53 (19%). This difference is mainly due to their different fractions observed in the underexpressed gene subsets (61% vs. 16%). We did not find such a difference in the overexpressed gene subsets (28% vs. 24%). Our analyses provide a set of common genes co-regulated by the two master TFs in HNSCC.

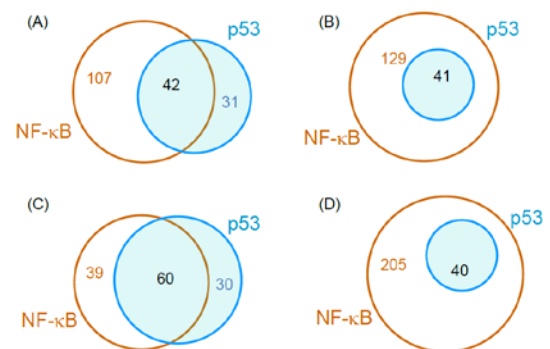


Figure 2. Overlaps between target genes of NF- κ B and p53 in wt p53-deficient HNSCC cells of overexpressed genes (A) and underexpressed genes (C), and mt p53 HNSCC cells of overexpressed genes (B) and underexpressed genes (D)

Additional TFs (AP1, EGR1, CEBPB, STAT3, and SP1) were also previously implicated as important regulators in the tumorigenesis. To identify regulatory programs co-regulated by NF- κ B and p53 with these TFs,

we first constructed two networks linking each TFs and their putative target genes predicted by the RCA-based method for the two tumor subgroups. Totally 298 and 232 genes were identified as targets of at least two TFs in the wt p53-deficient and the mt p53 cells, respectively (data not shown). Next, we identified two regulatory programs consisting of genes putatively co-targeted by all the seven TFs (Figure 3). The programs of the wt p53-deficient comprised 37 genes, where 17 and 12 genes are consistent with known NF- κ B and p53 targets based on previous publications, respectively. The percentage of known NF- κ B and p53 target genes in the program was greater than their total prediction (i.e. all of their predicted NF- κ B or p53 target genes), where NF- κ B is 46% vs. 19% and p53 is 34% vs. 14%. Similarly, 39 genes (including 12 known NF- κ B and 18 known p53 ones) formed the regulatory programs of the mt p53. The prediction of the known target genes in the network was also relatively accurate by compared with the total prediction for NF- κ B (31% vs. 15%) and p53 (46% vs. 29%). We further found that most genes in the TF regulatory programs were functionally classified to GO biological processes adhesion, angiogenesis, apoptosis, cell cycle, inflammatory and immune responses, proteolysis, regulation of transcription, etc.

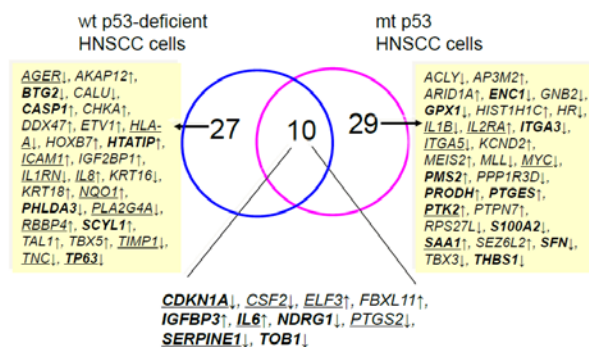


Figure 3. Gene programs co-regulated by all seven TFs (NF- κ B, p53, AP1, CEBPB, EGR1, SP1, and STAT3) in wt and mt p53 HNSCC cells. Genes in underlined, bold and bold-underlined refer to known targets of NF- κ B, p53 and NF- κ B/p53, respectively. $\downarrow\uparrow$ refer to genes differentially over- and underexpressed, respectively

3.3 microRNA target genes and their interaction with TFs

We applied the RCA-based approach to analyze target genes of microRNAs. Since oncogenic mir21 and tumor suppressor mir34s have been studied for their relationships with the p53 pathway [17, 18], we concentrated on their interaction with the NF- κ B regulatory network. mir34ac_449 was used to represent mir34s because both microRNAs mir34ac and mir449 share the same binding motif consensus from the available website (see method). In our analysis, 32% and 72% of the mir34ac_449 target

genes overlapped with NF- κ B ones in the wt p53-deficient and mt p53 cells, respectively, suggesting more interaction between mir34s and NF- κ B in gene regulation of mt p53 tumor cells. By the contrast, we did not observe such a difference of overlapping between mir21 and NF- κ B target gene sets (51-55% in the wt and mt p53 subgroups).

We then constructed two regulatory networks of NF- κ B, p53 and mir 21 or mir34s (Figure 4). The network of the wt p53-deficient comprised 49 common target genes of NF- κ B, p53, mir21 or mir34ac_449, respectively. Relatively, the network of the mt p53 was composed of a small number of 21 genes including 7 common targets of the two microRNAs. Even though most of common targets in the networks were underexpressed, we still detected several overexpressed ones, for example, *IL6* and *ELF3* (inflammatory), *PTGES* (proliferation), and *CASP4* (apoptosis) in the wt p53-deficient, and *MMP1* (proteolysis) and *PTK2* (angiogenesis and migration) in the mt p53 (Figure 4). This analysis highlights a considerable interaction of regulatory programs among NF- κ B, p53 and the two microRNAs.

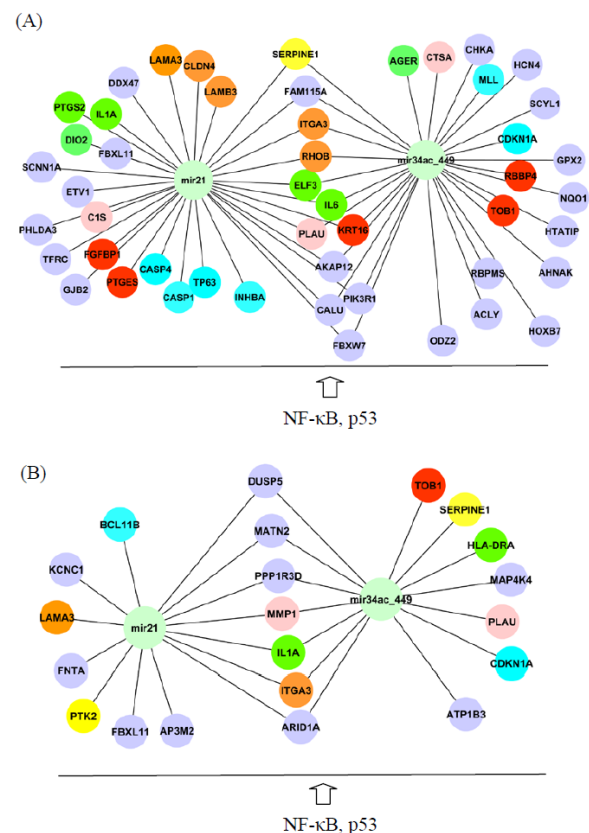


Figure 4. Regulatory gene networks of NF- κ B, p53 and microRNAs 21 and 34s in HNSCC cells. Every node represents a common target gene of NF- κ B, p53, mir21 or mir34ac_449, and was annotated to processes with different colors. (A) the wt p53-deficient. (B) the mt p53.

4 Discussion

Our integrative modeling is RCA-based and can capture the sparse structure existing in gene expression data for unraveling transcriptional regulatory networks. The efficiency of the RCA-based method is supported by its applicability in prediction of NF- κ B targets, in comparison with the analysis by other methods. We compared the RCA-based method with a similar method, COGRIM. Among the NF- κ B targets predicted by our method, 19% (in the wt p53-deficient) and 15% (in the mt p53) are consistent with known ones published previously. But the known NF- κ B genes predicted by COGRIM only reaches to 10% of the total prediction [7]. More importantly, the NF- κ B genes predicted by the RCA-based method are more functionally relevant than those by COGRIM (Table 1). Our method improved the efficiency and accuracy to identify regulatory associations between NF- κ B and their targets. The identified NF- κ B genes by the newly developed method are highly associated with biological processes, suggesting that they are biologically more meaningful than those by other methods.

A previous study has confirmed NF- κ B function in the tumor cells with both wt and mt p53 status [19]. By promoter analysis, p53 and NF- κ B were shown to play a reciprocal role in the two distinct over-expressed gene clusters of HNSCC [10, 15]. In the present study, we demonstrate a significant intersection of p53 and NF- κ B regulated genes in HNSCC. However, the p53 and NF- κ B interaction is different in the gene subsets underexpressed in the wt or mt p53 cells. In the wt p53-deficient cells, the two TFs can jointly regulate over 60% of the underexpressed genes. In contrast, all p53 targets were putatively regulated by NF- κ B in the mt p53 cells (Figure 2). This observation strongly suggests that a tight cooperation between NF- κ B family members and p53 controls the p53 network in the mt p53 cells, but to a less extent affects the p53 network of the wt p53-deficient cells. Moreover, our analyses showed a more accurate prediction of known NF- κ B and p53 target genes in the regulatory programs of the seven TFs (Figure 3) in comparison with those in the total prediction. Such predicted programs from both computational and literature search suggest that the malignant progression of HNSCC is likely as a result of co-regulation by a combinatorial cooperation of NF- κ B, p53, and the other five TFs.

Identification of TF-microRNA modules enhanced our understanding of complex transcriptional regulatory architectures in cancer cells. In this study, our results support that both mir21 and mir34s likely participate in transcriptional control of gene expression by NF- κ B and p53. Their functions may contribute to the progression or suppression of HNSCC cells. Several common genes were

downregulated by NF- κ B, p53 and the two microRNAs in both wt and mt p53 cells (Figure 4), such as *ITGA3* and *LAMA3* (adhesion), *SERPINE1* (angiogenesis), and *PLAU* (proteolysis). These are suggested to favor cancer metastasis [20, 21]. The microRNAs likely cooperate with p53 and NF- κ B to inhibit their expression so that repressing tumor progression of HNSCC. In contrast, several overexpressed genes in the networks may promote tumorigenesis of the wt p53-deficient cells by alterations in gene expression associated with inflammatory, proliferation, apoptosis and other processes, such as *IL6*, *ELF3*, *PTGES*, and *CASP4*, or trigger metastatic processes of the mt p53 cells.

5 Conclusions

In summary, our results provide a general view of cross-regulatory relationships among NF- κ B, p53 and the microRNAs in different malignant phenotypes. To our knowledge, this is the first investigation of TF-microRNA regulatory interactions by modeling diverse data sources and integrating constraints of sparseness in HNSCC. Within experimental validation of predicted microRNA targets, it would help in understanding of TF-microRNA regulatory mechanisms responsible for different cancer phenotypes and heterogeneity of HNSCC. Also, successful application of the RCA-based method in HNSCC showed it could serve as a useful approach to study on regulatory networks of regulators in other complex biological systems.

6 References

- [1] Dueck D, Morris QD, Frey BJ: Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* 2005, 21 Suppl 1:i144-i151.
- [2] Li H, Sun Y, Zhan M: The discovery of transcriptional modules by a two-stage matrix decomposition approach. *Bioinformatics* 2007, 23(4):473-479.
- [3] Pournara I, Wernisch L: Using temporal correlation in factor analysis for reconstructing transcription factor activities. *EURASIP J Bioinform Syst Biol* 2008:172840.
- [4] Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP: Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A* 2003, 100(26):15522-15527.
- [5] Hoffmann A, Natoli G, Ghosh G: Transcriptional regulation via the NF-kappaB signaling module. *Oncogene* 2006, 25(51):6706-6716.
- [6] Hayden MS, Ghosh S: Shared principles in NF-kappaB signaling. *Cell* 2008, 132(3):344-362.
- [7] Yan B, Chen G, Saigal K, Yang X, Jensen ST, Van Waes C, Stoeckert CJ, Chen Z: Systems biology-defined NF-kappaB regulons, interacting signal pathways and

- networks are implicated in the malignant phenotype of head and neck cancer cell lines differing in p53 status. *Genome Biol* 2008, 9(3):R53.
- [8] Ikeda A, Sun X, Li Y, Zhang Y, Eckner R, Doi TS, Takahashi T, Obata Y, Yoshioka K, Yamamoto K: p300/CBP-dependent and -independent transcriptional interference between NF-kappaB RelA and p53. *Biochem Biophys Res Commun* 2000, 272(2):375-379.
- [9] Park S, Hatanpaa KJ, Xie Y, Mickey BE, Madden CJ, Raisanen JM, Ramnarain DB, Xiao G, Saha D, Boothman DA et al: The receptor interacting protein 1 inhibits p53 induction through NF-kappaB activation and confers a worse prognosis in glioblastoma. *Cancer Res* 2009, 69(7):2809-2816.
- [10] Yan B, Yang X, Lee TL, Friedman J, Tang J, Van Waes C, Chen Z: Genome-wide identification of novel expression signatures reveal distinct patterns and prevalence of binding motifs for p53, nuclear factor-kappaB and other signal transcription factors in head and neck squamous cell carcinoma. *Genome Biol* 2007, 8(5):R78.
- [11] Ondrey FG, Dong G, Sunwoo J, Chen Z, Wolf JS, Crowl-Bancroft CV, Mukaida N, Van Waes C: Constitutive activation of transcription factors NF-(kappa)B, AP-1, and NF-IL6 in human head and neck squamous cell carcinoma cell lines that express pro-inflammatory and pro-angiogenic cytokines. *Mol Carcinog* 1999, 26(2):119-129.
- [12] Pensa S, Watson CJ, Poli V: Stat3 and the inflammation/acute phase response in involution and breast cancer. *J Mammary Gland Biol Neoplasia* 2009, 14(2):121-129.
- [13] Takahra T, Smart DE, Oakley F, Mann DA: Induction of myofibroblast MMP-9 transcription in three-dimensional collagen I gel cultures: regulation by NF-kappaB, AP-1 and Sp1. *Int J Biochem Cell Biol* 2004, 36(2):353-363.
- [14] Lv B, Wang H, Tang Y, Fan Z, Xiao X, Chen F: High-mobility group box 1 protein induces tissue factor expression in vascular endothelial cells via activation of NF-kappaB and Egr-1. *Thromb Haemost* 2009, 102(2):352-359.
- [15] Lee TL, Yang XP, Yan B, Friedman J, Duggal P, Bagain L, Dong G, Yeh NT, Wang J, Zhou J et al: A novel nuclear factor-kappaB gene signature is differentially expressed in head and neck squamous cell carcinomas in association with TP53 status. *Clin Cancer Res* 2007, 13(19):5680-5691.
- [16] Hoyer PO: Non-negative matrix factorization with sparseness constraints. *J Mach Learn Res* 2004, 5:1457-1469.
- [17] Hermeking H: The miR-34 family in cancer and apoptosis. *Cell Death Differ* 2009.
- [18] Si ML, Zhu S, Wu H, Lu Z, Wu F, Mo YY: miR-21-mediated tumor growth. *Oncogene* 2007, 26(19):2799-2803.
- [19] Duffey DC, Chen Z, Dong G, Ondrey FG, Wolf JS, Brown K, Siebenlist U, Van Waes C: Expression of a dominant-negative mutant inhibitor-kappaBalpha of nuclear factor-kappaB in human head and neck squamous cell carcinoma inhibits survival, proinflammatory cytokine expression, and tumor growth in vivo. *Cancer Res* 1999, 59(14):3468-3474.
- [20] Strojjan P, Budihna M, Smid L, Vrhovec I, Skrk J: Urokinase-type plasminogen activator (uPA) and plasminogen activator inhibitor type 1 (PAI-1) in tissue and serum of head and neck squamous cell carcinoma patients. *Eur J Cancer* 1998, 34(8):1193-1197.
- [21] Chen ZG: Exploration of metastasis-related proteins as biomarkers and therapeutic targets in the treatment of head and neck cancer. *Curr Cancer Drug Targets* 2007, 7(7):613-622.

Fuzzy Relational System for Identification of Gene Regulatory Network

Papia Das¹, Pratyusha Rakshit², Amit Konar², Mita Nasipuri¹, Atulya K. Nagar³

¹CSE Dept., ²ETCE, Jadavpur University, Kolkata, India

³Department of Math & Computer Science, Liverpool Hope University, Liverpool, UK

Abstract- *Generating inferences from a gene regulatory network is important to understand the fundamental cellular processes, involving gene functions, and their relations. The availability of time-series gene expression data makes it possible to investigate the gene activities of the whole genomes. Under this framework, gene interaction is explained through a set of fuzzy relational matrices. By transforming quantitative expression values into linguistic terms, the proposed technique defines a measure of fuzzy dependency among genes. Based on the fact that the measured time points are limited, we present an Artificial Bee Colony-based search algorithm to unveil potential genetic network constructions that fit well with the time-series data and explore possible gene interactions.*

Keywords- gene regulatory network; fuzzy relational system; fuzzy membership distribution; artificial bee colony optimization algorithm; differential evolution algorithm.

1 Introduction

Genes in living organisms form a virtual network through interaction with each other. This interaction mechanism is called gene regulatory network (GRN). GRNs form dynamic and distributed systems which control the expressions of the various genes in the cell. They explicitly represent the causality of developmental processes and explain exactly how genomic sequence encodes the regulation of expression of the sets of genes that progressively generate developmental patterns and execute the construction of multiple states of differentiation.

The complex control systems underlying development have probably been evolving for more than a billion years. These control systems consist of many thousands of modular DNA sequences. Each such module receives and integrates multiple inputs, in the form of regulatory proteins (activators and repressors) that recognize specific sequences within them. The end result is the precise transcriptional control of the associated genes. Some regulatory modules control the activities of the genes encoding regulatory proteins. Functional linkages between these particular genes, and their associated regulatory modules, define the core networks underlying development. This regulatory mechanism of genes provides an insight into the interaction between different genes.

With the rapid advancement of DNA microarray technologies, inferring genetic regulatory networks from time-series gene expression data has become critically important in revealing fundamental cellular processes, investigating functions of genes and proteins, and understanding complex relations and interactions between genes.

Several methods have been proposed to model maps of gene interaction, including Bayesian networks [1], dynamic Bayesian networks with hidden Markov model [2], and Boolean networks [3]. More recently, neural networks have also been applied to the problem of gene expression data analysis [4].

Boolean networks have been used to infer underlying GRN structures. In a Boolean network, the state of a gene is represented by a Boolean variable (ON or OFF) and interactions between genes are represented by Boolean functions. Boolean networks require that a number of assumptions be made to simplify analysis. Unfortunately, the validity of these assumptions has been questioned by many researchers, especially those in the biological community. To these researchers, there is a perceived lack of connection between simulation results and empirically testable hypotheses.

Instead of Boolean networks, Bayesian networks can also be used for GRN inferences. Bayesian network is a probabilistic model that describes the multivariate probability distribution of a set of genes whose interdependencies are known. A Bayesian network allows the conditional dependencies and independencies to be displayed by means of a directed acyclic graph. However, this approach to the learning of network structures is a NP-hard problem, especially for high-dimensional data such as gene expression data. Another problem that needs to be tackled when using the Bayesian network approaches for gene expression data analysis is concerned with the effect of small sample sizes.

A stochastic model of gene interactions capable of handling missing variables is proposed in [2]. It can be represented as a dynamic Bayesian network particularly well suited to tackle the stochastic nature of gene regulation and gene expression measurement. Parameters of the model are learned through a penalized likelihood maximization technique. The model referred to here is based on several strong assumptions, such as stationary or additive regulation. The model needs further improvement in order to represent more realistic phenomena, such as non-linear and combinatorial regulations.

Currently, with the advancements of the DNA micro array technology, it has become possible to simulate gene regulatory network from gene expression time-series data. In [6], a mathematical model for GRN has been proposed using fuzzy recurrent neural network to determine the numerical interaction values between genes. Due to the large number of model parameters and the small number of data sets available, the system of equations in GRN identification problem is highly underdetermined and ambiguous. GRN weights usually are multimodal functions of the gene expression time series

data. Hence, the solution sets of weights are non-unique, and naturally the solution does not guarantee the optimal selection of weights of the network.

In this context, it is necessary to propose models that attempt to get good predictions, reducing the need for prior knowledge. For one to infer the structure of a GRN, it is important to identify, for each gene in the GRN, whether other genes can affect its expression and how they can affect it. To better infer GRN structures, we propose a technique which is able to discover interesting fuzzy dependency relationships among genes. It can represent discovered fuzzy dependency relationships explicitly as “if a gene is highly expressed, its dependant gene is then lowly expressed” etc. These relationships can reveal biologically meaningful gene regulatory relationships that could be used to infer underlying GRN structures.

In this work, we present a fuzzy logic based algorithm for analyzing gene expression data, and employ an Artificial Bee Colony (ABC) optimization algorithm [7] to find the optimal membership function of normalized gene responses as well the fuzzy relation between genes. The membership function thus obtained are then defuzzified by centroidal defuzzification technique, and the results are found to be promising.

Using fuzzy logic, we have developed a technique to identify logical relationships between genes. The fuzzy logic has proved to be an important tool due to its ability to represent non-linear systems, its friendly language to express knowledge and the ability to incorporate and edit fuzzy rules. It can handle very noisy, high-dimensional time series gene expression data and can represent discovered fuzzy dependency relationships explicitly. These discovered relationships not only make hidden regularities easily interpretable, it also determines if a gene is supposed to be activated or inhibited and can be used to predict how a gene would be affected by other genes from an unseen sample (i.e., expression data that are not in the original database). The proposed technique has been tested with real expression data.

The performance of the current work is significantly better than the one reported in [6] considering root mean square error and convergence speed of the procedure. ABC seems to be promising for this optimization problem because of the following reasons: 1) providing better solution quality to find out fuzzy membership distribution of relation between genes in GRN, 2) combining local search methods with global search methods attempting to balance exploration and exploitation processes giving high speed of convergence, and 3) preventing the search technique from premature convergence problem providing global search ability with the help of scout unit.

The paper is organized as follows. First, the conventional concept of fuzzy sets and relations is described briefly in section 2. In section 3, we describe the fuzzy relational approach to solve GRN identification problem. The cost function used to determine the quality of a solution is proposed in section 4. In section 5, we describe the ABC optimization algorithm used to find the relational matrices between genes in the network and we explain the fuzzy technique to represent the membership values of gene response in section 6. In section 7, we present the simulated results and in section 8, we demonstrate the use of our model to simulate a gene regulatory network using real gene expression time series data. Section 9 concludes the paper.

2 An Overview of Fuzzy Sets and Relations

2.1 Definition 1

A *fuzzy set* A is a set of ordered pairs, given by

$$A = \{(x, \mu_A(x)) : x \in X\} \quad (1)$$

where X is a universal set of objects (also called the universe of discourse) and $\mu_A(x)$ is the grade of membership of the object x in A . Usually, $\mu_A(x)$ lies in the closed interval of $[0, 1]$.

2.2 Definition 2

A membership function $\mu_A(x)$ is characterized by the following mapping:

$$\mu_A(x) : x \rightarrow [0, 1], x \in X \quad (2)$$

where x is a real number describing an object or its attribute, X is the universe of discourse and A is a subset of X .

2.3 Definition 3

A fuzzy relation is a fuzzy set defined in the Cartesian product of crisp sets X_1, X_2, \dots, X_n . A fuzzy relation $R(x_1, x_2, \dots, x_n)$ thus is defined as

$$R(x_1, x_2, \dots, x_n) = \{\mu_R(x_1, x_2, \dots, x_n) / (x_1, x_2, \dots, x_n) \mid (x_1, x_2, \dots, x_n) \in X_1 \times X_2 \times \dots \times X_n\}$$

$$\text{where } \mu_R : X_1 \times X_2 \times \dots \times X_n \rightarrow [0, 1]. \quad (3)$$

In binary fuzzy relation instead of n universes we need only 2 universes.

2.4 Definition 4

A fuzzy implication relation for a given rule: IF x is A_i THEN y is B_i is formally denoted by

$$R_i(x, y) = \{\mu_{R_i}(x, y) / (x, y)\} \quad (4)$$

where the membership function $\mu_{R_i}(x, y)$ is constructed intuitively by many alternative ways. Here we have used Mamdani Implication. Mamdani proposed the following implication function:

$$\mu_{R_i}(x, y) = \min[\mu_{A_i}(x), \mu_{B_i}(y)] \quad (5)$$

2.5 Definition 5

Let us consider two fuzzy relations R_1 and R_2 defined on $X \times Y$ and $Y \times Z$ respectively. The *max-min composition* of R_1 and R_2 is a fuzzy set defined by

$$R_3 = R_1 \circ R_2 = \{\mu_{R_3}(x, z) / (x, z)\} \quad (6)$$

where

$$\mu_{R_3}(x, z) = \max_y \{\min(\mu_{R_1}(x, y), \mu_{R_2}(y, z)) \mid x \in X, y \in Y, z \in Z\}.$$

2.6 Definition 6

Let us consider a fuzzy production rule: IF x is A THEN y is B , and a fuzzy fact: x is A' . The Generalized Modus Ponens (GMP) inference rule then infers y is B' . Here A , B , A' , and B' are fuzzy sets such that A' is close to A , and B' is close to B . The inference rule also states that the closer the A' to A , the closer the B' to B . Symbolically, the GMP can be stated as follows:

Given: IF x is A THEN y is B .

Given: X is A' .

Inferred: y is B' .

For evaluation of membership distribution of y is B' , $\mu_{B'}(y)$, we need to know the membership distribution of x is

A' , $\mu_{A'}(x)$, and the membership of the fuzzy relation for the given IF-THEN rule, $\mu_R(x, y)$.

According to GMP

$$\mu_{B'}(y) = \mu_{A'}(x) \circ \mu_R(x, y) \quad (7)$$

where $\mu_{A'}(x)$ and $\mu_R(x, y)$ are row vector and matrices of compatible dimension respectively.

3 Solving the GRN Identification Problem by Fuzzy Relational Approach

To describe the proposed technique, let us assume that we are given a set of gene expression time series data $G = \{G_1, \dots, G_j, \dots, G_N\}$, consisting of N time series collected from experiments with N genes. Each of these N time series consists, in turn, of T data points collected at T different time instances.

Here we have considered that the response value of gene g_j at time instance t , $G_j(t)$ has a fuzzy membership distribution $\mu_A(G_j(t))$, and the corresponding fuzzy set A is given by the doublet $(G_j^k(t) | \mu_A(G_j^k(t)))$, where $j \in [1, N]$ and $k \in [1, F]$. The $G_j(t)$ is evaluated by centroidal defuzzification procedure given by

$$G_j(t) = \frac{\sum_{k=1}^F G_j^k(t) \times \mu_A(G_j^k(t))}{\sum_{k=1}^F \mu_A(G_j^k(t))} \quad (8)$$

As an example let $F=5$; so that a particular gene expression at time instance t can be represented as $\{0.2|0.35, 0.4|0.57, 0.6|0.62, 0.8|0.89, 1.0|0.93\}$, and after the de-fuzzification it becomes $(0.2 \times 0.35 + 0.4 \times 0.57 + 0.6 \times 0.62 + 0.8 \times 0.89 + 1.0 \times 0.93) / (0.35 + 0.57 + 0.62 + 0.89 + 0.93) = 0.68809$. At this point, we want the attention of the reader on the above fuzzy set A ; the members of fuzzy set A are 0.2, 0.4, 0.6, 0.8, and 1.0.

Now, gene expression can be described in two different states such as "highly expressed" and "lowly expressed" to a varying degree based on a set of membership functions. For our application here, we define two different states, "highly expressed" and "lowly expressed" in terms of two fuzzy sets as shown in Figure 1. In our proposed work, we are considering two fuzzy sets $A_1 = [0.1, 0.4]$ and $A_2 = [0.5, 1.0]$. Here $\mu_{A_1}(G_j(t))$ in fuzzy set A_1 indicates the degree of membership of $G_j(t)$ to be low and $\mu_{A_2}(G_j(t))$ in fuzzy set A_2 indicates the degree of membership of $G_j(t)$ to be high.

Let $A = A_1 \cup A_2$. Hence, gene expression is considered to be low with a high membership value of gene response within a range of 0.1 to 0.4 and otherwise gene expression is considered to be high.

From the membership distribution of $\mu_{A_1}(G_i(t=0))$, $\mu_{A_2}(G_i(t=0))$, $\mu_{A_1}(G_j(t=0))$ and $\mu_{A_2}(G_j(t=0))$ we can construct 4 fuzzy relational matrices for each pair of gene responses $G_i(t)$ and $G_j(t)$, $i, j \in [1, N]$ following Mamdani rule of Fuzzy implication.

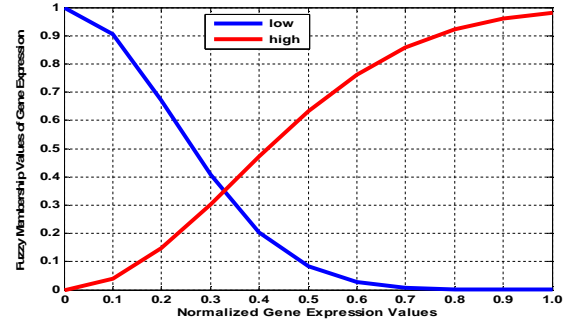


Figure 1. Fuzzy membership distribution of gene expression

The descriptions of four relational matrices are given as follows.

$$1) R_{i_low, j_low}(k, l) = \text{Min}(\mu_{A_1}(G_i^k(t=0)), \mu_{A_1}(G_j^l(t=0)))$$

$$2) R_{i_low, j_high}(k, l) = \text{Min}(\mu_{A_1}(G_i^k(t=0)), \mu_{A_2}(G_j^l(t=0)))$$

$$3) R_{i_high, j_low}(k, l) = \text{Min}(\mu_{A_2}(G_i^k(t=0)), \mu_{A_1}(G_j^l(t=0)))$$

$$4) R_{i_high, j_high}(k, l) = \text{Min}(\mu_{A_2}(G_i^k(t=0)), \mu_{A_2}(G_j^l(t=0)))$$

$\forall k, l \in [1, F]$.

The corresponding fuzzy production rules are given as follows.

PR1: IF g_i 's response is low
THEN g_j 's response is low.

PR2: IF g_i 's response is low
THEN g_j 's response is high.

PR3: IF g_i 's response is high
THEN g_j 's response is low.

PR4: IF g_i 's response is high
THEN g_j 's response is high.

Now, the entire fuzzy relational matrix between response of genes g_i and g_j is given by $R_{i,j}$ which is formed using 4 relational sub-matrices.

$$R_{i,j} = \begin{pmatrix} R_{i_low, j_low} & R_{i_low, j_high} \\ R_{i_high, j_low} & R_{i_high, j_high} \end{pmatrix} \quad (9)$$

Hence there will be such $N \times N$ relational matrices each of dimension $F \times F$.

Now our objective is to determine the membership distribution of gene g_i at next time instance $t+1$. Let this is denoted as $\mu_A(G_i(t+1))$. Once the relational matrix $R_{i,j}$ has been formed between two genes g_i and g_j , we can evaluate $\mu_A(G_i(t+1))$ by max-min composition between $R_{i,j}$ and $\mu_A(G_j(t))$, for $i, j \in [1, N]$, as given by GMP inference rule

$$\mu_A(G_i(t+1)) = \max_{j=1}^N [\mu_A(G_j(t)) \circ R_{j,i}], \forall i, j \in [1, N] \quad (10)$$

$$\text{where } \mu(G_j(t)) \circ R_{j,i} = \max_{k \in F} [\min\{\mu(G_j^k(t)), R_{j,i}(k, l)\}] \quad (11)$$

4 Proposed Cost Function

The proposed cost function in this work is designed keeping in mind the main issue of accurately identifying the existing relationship between genes in the network. Handling this issue is a tough job, since we do not have any knowledge except the available gene expression time series data. Therefore, a judicious choice of cost function can greatly influence the accuracy of the simulated network. To meet this issue, we evaluate the accuracy of the produced gene expression of our simulated network obtained using the fuzzy relational system by comparing it with the original gene expression with the hope that if the fuzzy relational matrices correctly identify the logical relationships between two genes then the difference (error) between the two set of gene expressions will be less. The error has been calculated by taking the squared difference between original gene expression, $G_{i_org}(t)$, and experimental gene expression, $G_{i_cal}(t)$, given by

$$\text{cost}_{fn} = \frac{1}{N \times T} \sum_{t=1}^T \sum_{i=1}^N (G_{i_org}(t) - G_{i_cal}(t))^2 \quad (12)$$

5 Artificial Bee Colony Optimization algorithm (ABC)

In ABC algorithm, the colony of artificial bees contains three groups of bees:

- Onlooker bee makes decision to choose a food source.
- Employed bee selects a food source.
- Scout bee carries out random search for food source.

Here, the position of a food source represents a possible solution of the optimization problem and the nectar amount of a food source corresponds to the fitness of the associated solution. The number of employed bees and onlooker bees is equal to the number of solutions in the population. ABC consists of following steps:

5.1 Initialization

ABC generates a randomly distributed initial population P ($G=0$) of N_p solutions (food source positions). Each solution X_i ($i=0, 1, 2, \dots, N_p - 1$) is a D dimensional vector.

5.2 Placement of employed bees on the food sources

An employed bee produces a modification on the position in her memory depending on the local information (visual information) as stated by equation (14) and tests the nectar amount of the new source. Provided that the nectar amount of the new one is higher than that of the previous one, the bee memorizes the new position and forgets the old one. Otherwise she keeps the position of the previous one in her memory.

5.3 Placement of onlooker bees on the food sources

An onlooker bee evaluates the nectar information from all employed bees and chooses a food source depending on the probability value associated with that food source, p_i , calculated by the following expression:

$$p_i = \frac{\text{fit}_i}{\sum_{j=0}^{N_p-1} \text{fit}_j} \quad (13)$$

where fit_i is the fitness value of the solution i evaluated by its employed bee. After that, as in case of employed bee, onlooker bee produces a modification on the position and checks the nectar amount of the candidate source. Onlooker bee memorizes the better position only.

In order to find a solution X'_i in the neighborhood of X_i , a solution parameter j and another solution X_k are selected on random basis. Except for the value of chosen parameter j , all other parameter values of X'_i are same as in the solution X_i , for example, $X'_i = (x_{i0}, x_{i1}, \dots, x_{i(j-1)}, x'_{ij}, x_{i(j+1)}, \dots, x_{i(D-1)})$. The value of x'_{ij} parameter in X'_i solution is computed using the following expression:

$$x'_{ij} = x_{ij} + u(x_{ij} - x_{kj}) \quad (14)$$

where u is a uniform variable in $[-1, 1]$ and k is any number between 0 to N_p-1 but not equal to i .

5.4 Send scouts for discovering the new food sources

In the ABC algorithm, if a position cannot be improved further through a predefined number of cycles called 'limit', the food source is abandoned. This abandoned food source is replaced by the scouts by randomly producing a position.

After that again steps (B), (C) and (D) will be repeated until the stopping criteria is met.

6 Extraction of Fuzzy Relationship between Genes Using ABC

In our paper, we have used the well known Artificial Bee Colony (ABC) optimization algorithm to find the simulated network. To spread the initial candidate solutions as far possible in the search space with the hope that some of the solutions may be close to the original solution we have used a chaos system [5] in ABC. The process of producing the chaos is as follows:

$$Z_{k+1} = \mu Z_k (1 - Z_k) \quad (15)$$

where $k = 0, 1, 2, 3, \dots, \Theta$, Θ is the number of chaotic iteration, μ is the control parameter. Z_k takes any value between 0 and 1; it is the selected value in the k th iteration. We indeed found that this initialization improve the overall convergence rate of the artificial bee colony optimization algorithm.

Each individual food source of ABC represents a complete solution. As an example one solution of the $N=4$ gene network contains $N \times F=4F$ data points where F is the number of elements in each of the $N=4$ fuzzy sets. These sets represent the membership values of gene responses in the network. We maintain a pop_size number of individual food sources all the time in the population pool. The population pool of the ABC optimization algorithm for the four gene network with $F=10$ can be represented pictorially as a two dimensional matrix as shown in Figure 2.

In Figure 2, $F=10$ and $\mu_A(G_i^k)$ represents the fuzzy membership values of the gene expression G_i of any individual food source, $k=1, 2, \dots, 10$ with the Fuzzy members as $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. At each step of ABC we evaluate fuzzy membership distribution of gene response, de-fuzzify each membership, calculate the cost function, and make the appropriate decision whether to keep that particular food source for the next generation or not.

$\mu_A(G_1^1)$	$\mu_A(G_1^2)$	$\mu_A(G_1^3)$	$\mu_A(G_1^4)$	$\mu_A(G_1^5)$	$\mu_A(G_1^6)$	$\mu_A(G_1^7)$	$\mu_A(G_1^8)$	$\mu_A(G_1^9)$	$\mu_A(G_1^{10})$
$\mu_A(G_2^1)$	$\mu_A(G_2^2)$	$\mu_A(G_2^3)$	$\mu_A(G_2^4)$	$\mu_A(G_2^5)$	$\mu_A(G_2^6)$	$\mu_A(G_2^7)$	$\mu_A(G_2^8)$	$\mu_A(G_2^9)$	$\mu_A(G_2^{10})$
$\mu_A(G_3^1)$	$\mu_A(G_3^2)$	$\mu_A(G_3^3)$	$\mu_A(G_3^4)$	$\mu_A(G_3^5)$	$\mu_A(G_3^6)$	$\mu_A(G_3^7)$	$\mu_A(G_3^8)$	$\mu_A(G_3^9)$	$\mu_A(G_3^{10})$
$\mu_A(G_4^1)$	$\mu_A(G_4^2)$	$\mu_A(G_4^3)$	$\mu_A(G_4^4)$	$\mu_A(G_4^5)$	$\mu_A(G_4^6)$	$\mu_A(G_4^7)$	$\mu_A(G_4^8)$	$\mu_A(G_4^9)$	$\mu_A(G_4^{10})$

Figure 2. Individual solution used in optimization algorithm

7 Simulation Results

The gene regulatory network identification problem is implemented in a Pentium processor. The results are generated with 4 time series data, one for each of 4 genes. The experiments are conducted for F=5, 10, and 20.

7.1 Experiment with Artificial Bee Colony

Figure 3 shows 16 fuzzy relational matrices R_{ij} between responses of genes g_i and g_j , $\forall i, j \in [1,4]$ with 1000 iterations for ABC algorithm with limit=100 and 300 iterations for chaotic initialization algorithm.

7.2 Experiment with Differential Evolution

Figure 4 shows 16 fuzzy relational matrices R_{ij} between responses of genes g_i and g_j , $\forall i, j \in [1,4]$ with 1000 iterations for DE algorithm with Cr=0.9 and 300 iterations for chaotic initialization algorithm.

7.3 Results on the time series data

Using the relational matrices obtained from ABC- and DE-based simulations, and de-fuzzifying values of gene responses at $t=1,2,\dots, 150$, we obtain the calculated gene expression time-series data. The relative performance of ABC-, DE-based simulations using our approach as well as the fuzzy recurrent neural approach proposed in [6], can be studied through the plot (Figure 5(i)-(iv)). Each plot consists of the gene expression levels at different time instances obtained by our approach (using ABC and DE), work proposed in[6] and the original time series data for a particular gene. Now we compare the derived time series plot with the original gene expression time series data. It is evident from the figures that ABC- based simulation using fuzzy relational system has outperformed the other two approaches.

7.4 Cost function evaluation

In order to compare the ability of ABC- and DE- based simulations to provide better solution with less cost function value, we plot the cost function value of the best solution obtained in each iteration of ABC- and DE-based simulations in Figure 6. It is apparent that for a fixed number of iteration ABC provides better solution than DE.

7.5 Performance analysis

To analyze the performance of the proposed approach for identification of gene regulatory network, we measure the following two parameters.

7.5.1 Root mean square error (RMSE)

The performance metric used here to determine how close the estimated gene responses are close to the original values of gene expressions is Root Mean Square Error (RMSE) given as

$$RMSE = \sqrt{\frac{1}{N \times T} \sum_{t=1}^T \sum_{i=1}^N (G_{i_org}(t) - G_{i_cal}(t))^2} \quad (16)$$

Here, T=150 and N=4. We obtain the following results from the plot of time series data in Fig.5.

- RMSE for ABC-based simulation with fuzzy relational system= 3.0239%
- RMSE for DE-based simulation with fuzzy relational system= 5.0735%
- RMSE for DE-based simulation with recurrent fuzzy neural model as in [6] = 15.6667%

7.5.2 Run time

After carrying out the experiment in a Pentium dual port computer using ABC optimization and DE algorithms, we find out

- Run_time_{ABC}=59 minutes
- Run_time_{DE}=32 minutes

ABC- based simulation takes more time than DE- based simulation due to complexity involved in ABC.

In Table-I, we represent the mean fuzzy relational matrix indicating relationship between expression of genes g_2 and g_1 obtained using ABC-based simulation after 25 runs with $F=10$. A close inspection of Table-I indicates that membership value of expression of gene g_2 is high (low) when that of gene g_1 is low (high). It indicates that gene g_1 regulates expression of gene g_2 by inhibiting its response.

8 Inferring GRN Using Real Data Set

We have used our model to infer the gene regulatory network of e.coli. Bacteria S.O.S DNA repair network consisting of nearly 30 genes regulated at the transcription level. Four experiments have been conducted with different UV light intensities and eight major genes have been documented. These genes are *uvrD*, *lexA*, *umuD*, *recA*, *uvrA*, *uvrY*, *ruvA*, *polB*. This data set is available in the website [<http://www.weizmann.ac.il/mcb/UriAlon>]. We have conducted same experiment as with the above artificial data.

The identified gene responses are represented in Figure 8.

TABLE-I: Fuzzy Relational Matrix between expression of genes g_1 and g_2

		g2 response										
		low						high				
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
g1 response	low	0.1	0.01	0.08	0.12	0.08	0.05	0.05	0.67	0.63	0.71	0.89
		0.2	0.01	0.27	0.45	0.19	0.19	0.04	0.69	0.71	0.67	0.86
		0.3	0.08	0.25	0.25	0.17	0.17	0.07	0.69	0.45	0.58	0.90
		0.4	0.27	0.05	0.36	0.17	0.19	0.17	0.55	0.45	0.55	0.71
		0.5	0.25	0.36	0.44	0.45	0.55	0.45	0.55	0.27	0.55	0.89
	high	0.6	0.71	0.63	0.58	0.55	0.55	0.44	0.55	0.25	0.55	0.67
		0.7	0.89	0.72	0.69	0.69	0.69	0.69	0.45	0.55	0.60	0.55
		0.8	0.86	0.67	0.72	0.72	0.63	0.45	0.44	0.55	0.27	0.07
		0.9	0.90	0.69	0.71	0.69	0.71	0.44	0.35	0.12	0.08	0.17
		1.0	0.80	0.83	0.72	0.72	0.69	0.58	0.15	0.45	0.25	0.05

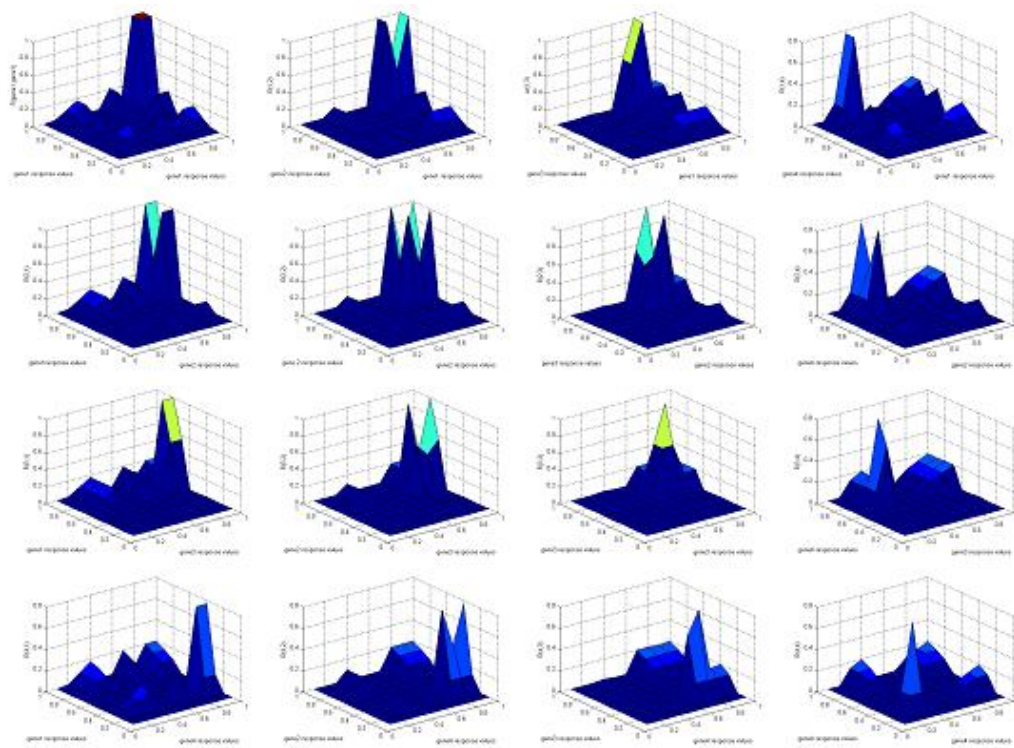


Figure 3. Fuzzy relational matrices $R_{i,j}$, $\forall i,j \in [1,4]$, obtained from ABC-based simulation

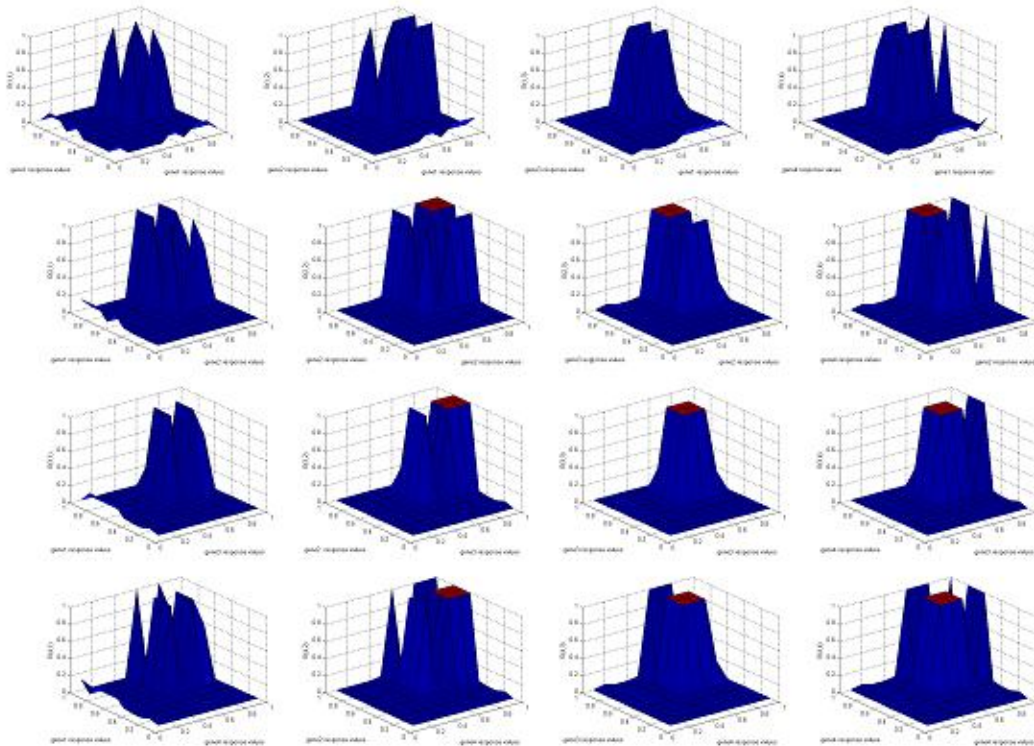


Figure 4. Fuzzy relational matrices $R_{i,j}$, $\forall i,j \in [1,4]$, obtained from DE-based simulation

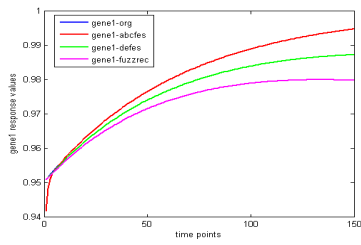


Figure 5(i). Plot of time series data for gene 1

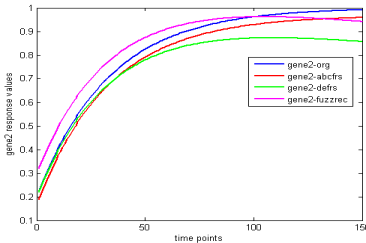


Figure 5(ii). Plot of time series data for gene 2

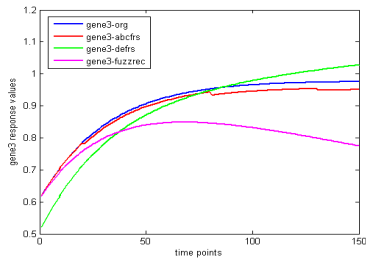


Figure 5(iii). Plot of time series data for gene 3

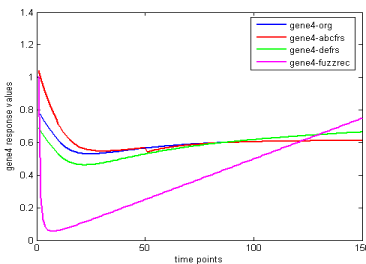


Figure 5(iv). Plot of time series data for gene 4

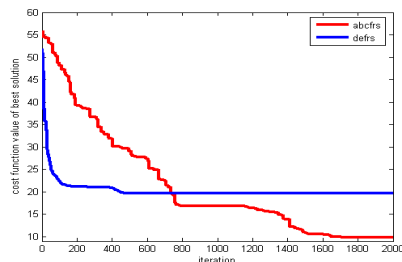


Figure 6. Minimum cost function value in each iteration of ABC- and DE-based simulation

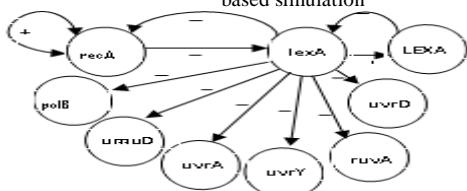


Figure 7. E.coli S.O.S. DNA repair network, activation is represented by '+' sign and inhibition by '-'

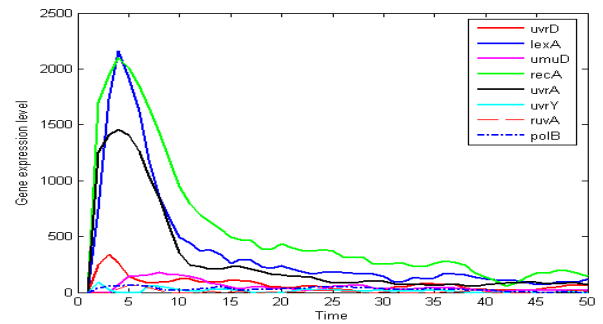


Figure 8. The measured gene expression profile of e. coli.

9 Conclusion

In this paper, we have presented an effective fuzzy technique for the discovery of GRNs from time series gene expression data. We design the fuzzy rules according to expressing level of gene, and fuzzy set theory. The proposed technique can discover fuzzy dependency relationships in high-dimensional and very noisy data. Based on the discovered fuzzy dependency relationships, the user can not only determine those genes affecting a target gene but also can identify whether or not the target gene is supposed to be activated or inhibited. The simulation results on both the artificial and the real data demonstrate that the proposed method is very promising in capturing the nonlinear dynamics of genetic regulatory systems and unveiling the potential gene interaction relation.

10 References

- [1] P. Spirtes, C. Glymour, R. Scheines, S. Kauffmann, V. Aimala and F. Wimberly, "Constructing Bayesian Network Models of Gene Expression Network from Microarray Data", Proc. Atlantic Symp. Computational Biology, Genome Information Systems and Technology, 2000.
- [2] E. Perrin, L. Rolaivola, A. Mazurie, S. Bottani, J. Mallet, and F. D'Alche-Buc, "Gene Network Inference using Dynamic Bayesian Networks", Bioinformatics, vol.19(2):138-148, 2003.
- [3] S. Laing, S. Fuhrman and R. Somogyi, "REVEAL, A general reverse engineering algorithm for inference of genetic network architectures", Proc. Pacific Symp. Biocomputing 3, 1998.
- [4] A. Nnarayanan, E.C. Keedwell, J. Gamalielsson and S. Tataneni, "Single Layer Artificial Neural Network for gene expression analysis", Proc. Neurocomputing Conf., vol.61:217-240, 2004.
- [5] C. Lng, S.Q. Li, "Chaotic spreading sequences with multiple access performance better than random sequences". IEEE transaction on Circuit and System-I, Fundamental Theory and Application, 47(3):394-397, 2000.
- [6] D. Datta, A. Konar, R. Janarthanan, "Extraction of interaction information among genes from gene expression time series data", NaBIC 2009.
- [7] B. Basturk, and Dervis Karaboga, "An Artificial Bee Colony (ABC) Algorithm for Numeric function Optimization" IEEE Swarm Intelligence Symposium 2006, May 12-14, 2006, Indianapolis, Indiana, USA.

CUDA-Accelerated Data-Mining for Putative Heteromeric Transcription Factors and Target Genes Using Microarray Gene Expression Profiles

Edward A. Salinas¹, Amitava Karmaker²

¹Independent Researcher, Rockville, Maryland 20852, USA

²University of Wisconsin-Stout, Menomonie, Wisconsin 54751, USA

Abstract - Understanding protein-protein and protein-DNA interactions is key to understanding the dynamics of gene regulation [3,17]. We here review a previously presented method [1,15,20], based on a variation of microarray expression profile correlation analysis, that seeks to identify interactions between a putative heteropolymeric transcription factor (TF) complex and DNA as well as some experimental results that bolster the argument for the method's validity. The method incorporates correlation coefficients between genes and transcription factors expression profiles, but also between genes and hypothetical TF co-factors, whose expression profiles are estimated by taking minima from constituent profiles. Second, we extend the technique to search for fourth-order protein interactions ($k=4$). Since a CPU-based analysis would require an execution time on the order of months, we have implemented the $k=4$ analysis on a CUDA-enabled NVIDIA GPU [16]. With CUDA, we achieved speedups of about 6-fold. Finally, we present the results of the higher order analysis and discuss those results as well as the implementation of the method using CUDA. To our knowledge CUDA has never been used to implement this particular algorithm for microarray gene expression profile analysis.

Keywords: Microarrays, Biological Data Mining, CUDA, correlation coefficients.

1 Introduction

Since the sequencing of the human genome [2] has been completed, the interpretation and biological connotation of sequences and the annotation of functional elements of the genome have been of great interest to researchers. Although a large number of human genes have been identified, their complete regulatory mechanisms are not wholly known at the transcriptional level [3]. To understand gene regulation, we need to identify regulatory elements and the transcription factor complexes that can regulate gene expression, allowing the construction of transcription regulatory networks (TRNs). To control the expression of genes, TF proteins bind to cis-elements in promoter regions and either facilitate or inhibit gene expression. Simply stated, trans-elements can be considered to be “keys”, cis-elements “locks”, and together “opened doorways” to transcription. By establishing whole

TRNs, we may be able to identify novel methods of gene regulation which could have applicability both in the laboratory and clinical settings.

In the post-genomic era, it has been a challenging task in functional genomics to construct TRNs from protein-DNA interactions. *In silico* discovery of transcription regulatory elements is quite effective for prokaryotes, like *Escherichia coli* [4], where genomes are more compact with many genes being regulated by a single operon. For higher multi-cellular eukaryotes, model-based approaches [3] that discover patterns among co-expressed genes with respect to regulating TFs have been proposed. The techniques involve finding over-represented DNA motifs and common transcriptional regulatory modules among co-expressed genes. A number of statistical and machine-learning algorithms have been used; they include position-weighted matrices, position-specific score matrices, Markov chains, artificial neural networks, and expectation maximization [5-11]. However, it has been reported that techniques incorporating a model-prediction-based approach are susceptible to a high false-positive prediction rate and that a majority of predicted TFBSs have no functional role *in vivo* [12].

Determining new ways to predict which proteins might participate in a heteromeric complex may facilitate the discovery of new TRNs. In this paper, we hypothesize that heteromeric TF complexes can be predicted *in silico* based on their constituent TF expression profiles. Using transcription factor expression profiles and gene expression profiles from microarray data, we review a technique that relies on combinations of TFs and correlation coefficients to predict TF-complexes [1,15,20]. Our dataset includes gene and TF expression profiles from a human female over 115 tissues samples [13]. The technique considers hypothetical TF-complex expression profiles in a given tissue which are estimated by taking minima from the constituent factors from the given tissue. By comparing these hypothetical profiles with each other and with the genuine expression profiles using correlation coefficients, we identify possible complexes. These proposed and hypothetical complexes are given a score based on the comparison. These scores are then compared with scores from other proposed and hypothetical complexes. This comparison leads to the identification of complexes that we believe are more likely to be genuine, and not hypothetical.

Our technique relies on a combinatorial approach selecting a gene, and tuples of TFs and computing many correlation coefficients. Due to extended program execution times, we decided to implement our algorithm using CUDA. We were able to achieve a speedup of approximately 6-fold. As a result of our analyses, we have been able to perform some validation of our technique as well as identify possible hetero-tetrameric transcription factor complexes. In the following sections of this paper, we describe our technique, its implementation and validation, our findings, and conclude with a brief discussion.

2 Methods and Materials

To carry out the analyses, we used publicly available microarray expression data [13]. The dataset covers a number of human genes and transcription factors expressions across 115 tissue samples, from adrenal tissue to uterine tissue. The dataset is essentially a matrix of expression values with genes indexed by row indices and tissues by column indices; each entry in the data matrix thus represents a gene's expression in a specific tissue. The data is different from typical microarrays in that the genomic DNA is used to estimate mRNA transcript abundance. A subset of 3166 gene transcripts, representing 2526 unique genes, of the data was selected and set aside. Additionally, 352 transcripts, based on information in the entrez-gene and TRANSFAC databases [23, 24] were tagged as transcription factors and also set aside. These data were used for all analyses.

Initial experimental correlation coefficients led to a distribution of coefficients that were weak and centered around zero [1]. This led to the development of a gene data pre-processing step where each gene's expression value was transformed with the equation $y' = ye^{ay}$ where a is a constant. For all experiments done for this paper, the value of a was set to 0.5. The graph in figure 1 demonstrates the motivation for the transformation.

Given a dataset of 1 row of microarray data for a gene g and a set of rows of N transcription factors TF_1, \dots, TF_N , our technique to assess the relationship between g and those transcription factors is as follows. First, the expression data for the gene is transformed with the previously described alpha transformation. Second, borrowing from previous techniques [12] N correlation coefficients are computed between the gene's expression values and the individual transcription factor expression values. The Pearson Correlation Coefficients are computed using the formula:

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (1)$$

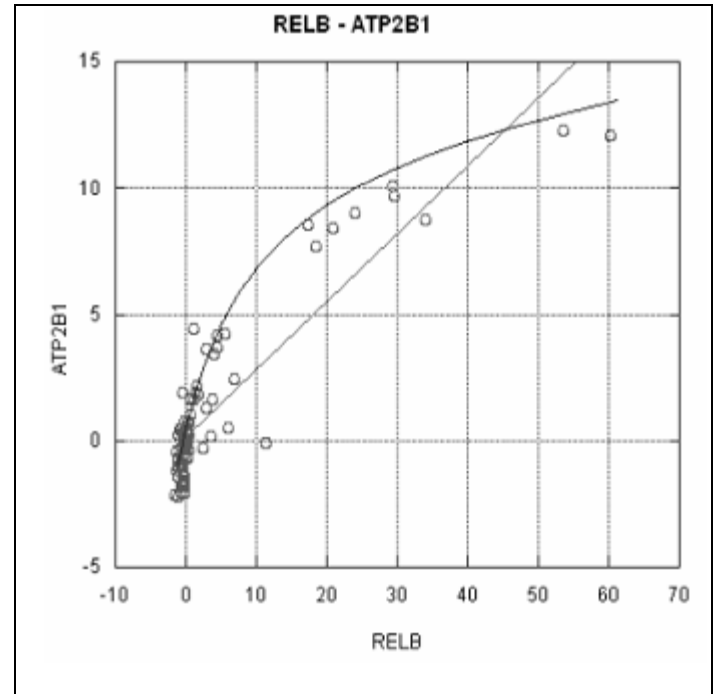


Fig. 1 Data such as depicted this chart helped motivate the α -transformation of the gene data.

Third, between all possible pairs, the hypothetical expression levels are computed and then as many correlation coefficients are computed. The hypothetical dimeric expression profiles are computed by taking the minimum expression value between the two constituent TFs expression values for a given tissue and assigning that value to the corresponding tissue expression for the hypothetical dimer. The same procedure is done for remaining $k=3, \dots, N$ expression profile triplets, quadruplets, etc. of the corresponding hypothetical trimers, tetramers, etc.

For example, for a hypothetical tetramer, its expression at tissue j would be $\min(TF_1, TF_2, TF_3, TF_4)$ where the TF_x is the x^{th} constituent factor at the j^{th} tissue. This way, altogether, the sum of $C(N, k)$ (" N choose k "), for $k=1, 2, \dots, N=k_{\max}$ correlation coefficients are computed between the gene expression profile and the real and hypothetical expression profiles; N are real and the remaining are hypothetical.

Fourth, the highest-order coefficient (k_{\max}), where the *minima* of N values for a given tissue was taken is compared with the remaining, lower-order coefficients. The value a , which we call the absolute improvement score is computed with the formula:

$$\min_{y \neq k_{\max}} (|\rho_{k_{\max}} - \rho_y|) \quad (2)$$

where the minimal absolute value between the highest order correlation and all other correlations is taken. This score we believe helps to distinguish any transcription regulatory signal from the highest-order hypothetical TF from the others. Fifth,

this procedure is carried out for all genes and for all k -tuples of transcription factors. In total,

$$c = g \left(\sum_{k=1}^{k_{max}} \binom{N}{k} \right) \quad (3)$$

where g is the number of genes, N is the number of transcription factors coefficients are computed, k is the different numbers of combinations of factors chosen, and k_{max} represents the highest-order polymerization under consideration. For the CFOS/CJUN example later, k_{max} is 2; for data-mining for heterotetramers, k_{max} is 4. Note that the sum over combinations is used in Eq. 3 because an analysis requires the computation of lower-order coefficients in the formula for computing the absolute improvement score. Finally, we rank the complexes by their scores. Figure 2 gives a schematic giving an overview of the technique.

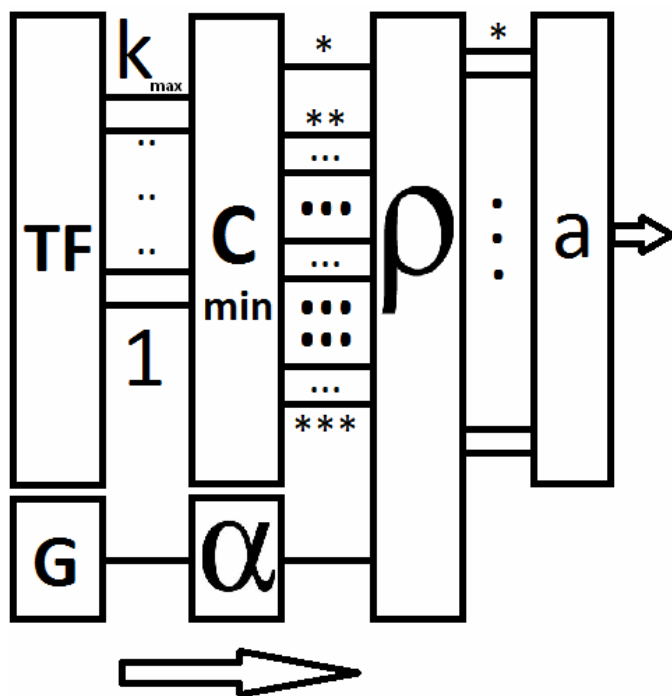


Fig. 2. A schematic shows data-flow and operations of the algorithm. TFs are chosen (k in total); a gene is chosen (box “G”) and then subjected to the alpha transformation (box “ α ”); 1-tuples, 2-tuples, ..., ($k_{max}-1$)-tuples, and k_{max} -tuples of TFs are chosen and minima are taken to form hypothetical expression profiles (boxes labeled “TF” & “C_{min}”). Finally, correlations are computed between the gene and all of the TF profiles (box “ ρ ”) (both genuine and hypothetical) and compared to generate an absolute improvement score for the highest-order putative heteropolymeric TF complex (box “a”). The scores are used for ranking hypothetical TFs as being likely transcription factor complexes. **Legend:** The “*” represents the highest-order coefficient, “***”, intermediates, and “****” the lowest.

When a gene shares a name with any of the transcription factors, or if any pair of the transcription factors share a name, then the corresponding coefficients and absolute improvement value are not computed. Such analyses are not carried out because we do not wish to consider polymerization involving self-regulating genes or any degree of homo-polymerization.

Central hypotheses of this project are that by taking the minima at a given tissue across expression profiles that we find the hypothetical expression profile of the corresponding polymeric TF and also that the computation and subsequent sorting of the absolute improvement scores may identify and distinguish a transcription regulatory signal from the transcription factors and their hypothetical joining to regulate the corresponding gene by binding to transcription factor binding sites on DNA.

All analyses were completed with a custom-written C/C++ computer program running on a 64-bit Ubuntu/Linux platform with an Intel core i7-960 processor. Perl and bash scripts played a role in loading data into our program as well. Our dataset was not free of missing values. Missing values were indicated with the value (-18). In computing the correlation coefficients, columns (tissues) with missing values were ignored and skipped over. In computing the hypothetical expression profiles, if any single component TF profile had a missing value in a given column, then the hypothetical profile was defined to also have a missing value in that column.

2.1 Methods Validation

To explore the validity of our technique we selected two well-known heterodimer-forming transcription factors CFOS and CJUN [26] from our dataset and applied our algorithm. The two transcription factors together form AP-1. Using the TRANSFAC and ENCODE [24, 25] databases we identified a total of 4 known target genes of the AP-1 TF complex in our gene dataset: TIMP1, GJA1, HMGA1, and MAP4K5. A perfect data-mining technique to identify TFs and their target genes would identify at least these four known target genes for AP-1. As described in the METHODS section, using every pair of transcripts in our dataset belonging to CFOS and CJUN, we carried out a $k_{max}=2$ analysis and computed correlation coefficients, hypothetical expression profiles, absolute improvement scores, and then sorted. After sorting our list and discounting the reported target genes CFOS, and CJUN (the components of AP-1 itself), we found two of the known target genes (HMGA1 and MAP4K5) among the top ten rows of the sorted list of absolute improvement scores and corresponding genes and TFs. Using the hypergeometric distribution, similarly as elsewhere [21, 22], based on the null hypothesis that the four known positives are distributed in the list of 2526 genes at random, we computed that there is a P-value of $8.4 \cdot 10^{-5}$ for finding 2 or more of the known target genes in the top 10 of the list sorted by the absolute improvement scores. This indicates that we may reject that null hypothesis, H_0 , that the four target genes are randomly distributed in the sorted list at the $\alpha=1\%$ significance threshold. The results are displayed in table 1.

Table 1. Genes and correlations (between CFOS, CJUN and the supposed but genuine AP-1 complex. Known targets of the AP-1 complex are underlined. AI is the absolute improvement score, used for ranking.

	Gene	C1	C2	CC	AI
1	VAR52	0.43	-0.32	0.07	0.36
2	RNU3IP2	0.50	-0.23	0.15	0.35
3	ZFX	-0.40	0.37	-0.06	0.34
4	AP2S1	0.46	-0.21	0.12	0.34
5	LRP6	-0.37	0.37	-0.05	0.32
6	<u>MAP4K5</u>	-0.35	0.32	-0.04	0.32
7	LOC56902	0.42	-0.22	0.09	0.31
8	ERG1	-0.04	0.61	0.30	0.31
9	<u>HMGAI</u>	-0.25	-0.25	0.05	0.30
10	TAPBP	-0.22	-0.22	0.08	0.30

2.2 Data Mining for Heterotetrameric Transcription Factors

To search for putative heterotetrameric transcription factors, we decided to carry out our algorithm with $k_{max}=4$; we wrote a C/C++ computer program and ran four time trials. Using a quad-core i7 Pentium processor and the OpenMP API for multi-threaded computer programming, our $k_{max}=4$ analysis was over a single gene running 1, 2, 3, and 4 OpenMP threads at a time. Our time trials were done not to analyze the results, but solely to acquire execution-time data. With four essentially identical time-trials with 1..4 OpenMP threads we saw average execution times of 3090, 1550, 1036, and 780 seconds. For analyzing all 3166 gene transcripts (including loading the data and printing results), this would be about 113, 57, 38, and 29 days. Preferring shorting execution times, we deemed such running times too long; in fact a previous analysis never completed [20]. Figure 3, along with some power curves generated with Excel, shows the timing data for the time trials of a single gene.

For these reasons we decided to explore computing the correlation coefficients using C/C++ and NVIDIA's CUDA architecture. CUDA is a specialized GPU parallel computing architecture implemented on NVIDIA GPUs. CUDA-enabled NVIDIA GPUs allow the parallel execution of threads on the GPU within logically organized grids. The organization is known as an execution configuration. The precise parameters for the execution configuration are set up by the program. A complete description of CUDA is beyond the scope of this paper, but it suffices to say that it enables programs that run on the GPU, called "kernels" to run many threads in a parallel at a time and that CUDA is optimized for arithmetically intense compute-bound programs which have a high ratio of computation operations to I/O operations. More information can be found about CUDA elsewhere [16, 19].

2.3 CUDA-accelerated Data Mining for Heterotetrameric Transcription Factors

Our C/C++ CUDA-based implementation of the $k_{max}=4$ analysis incorporated 42,875 grid blocks (a $35 \times 35 \times 35$ cube of blocks) with each block composed of 1000 threads (a

$10 \times 10 \times 10$ cube of threads) for its kernel execution configuration. This way, each CUDA-kernel invocation led to the execution of $(10 \times 35)^3 = 350^3 = 42,875,000$ CUDA threads. For a given kernel invocation, the gene is fixed. In the large grid-cube of threads, the row, column, and height indices correspond to row indices in our dataset table of transcription factors. This way, at cell x, y, z in the cube, the correlation coefficient with the hypothetical trimeric transcription factor made of the three transcription factors indexed by $x, y,$ and z is computed by a thread; control flow, partitioning the cube in two by the inequality $x > y > z$, prevents redundant computation of some coefficients however. If any of the indices are equal, then the correlation is between the gene and either a dimer (if two are equal) or a monomer (if all three are equal). This is because

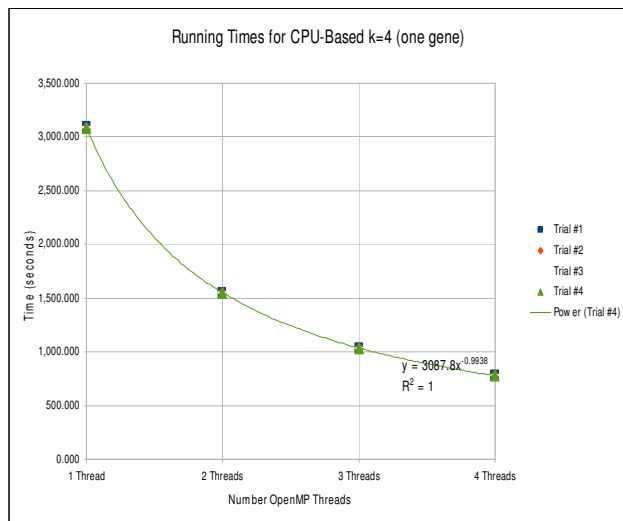


Fig. 3. Four essentially indistinguishable execution time data and power curves for a $k=4$ analysis with one gene using 1,2,3, & 4 OpenMP threads

of the property of the minimum function: $\min(a_1, a_2, \dots, a_N) = a_1$ if $a_1 = a_2 = \dots = a_N$. This way, a single kernel invocation computes correlations with $k=1$, $k=2$, and $k=3$ which yielded great computational efficiency.

We chose the dimension 350 because of the maximal 1000 threads per block limitation of the CUDA compute capability of the NVIDIA GTX 590 GPU we employed. Because of the 352 TFs in the dataset, the $350 \times 350 \times 350$ grid could not accommodate the computation of all the coefficients. To address this issue we introduced grid offsets into our code. This way, the grid always computes the aforementioned 42 million coefficients, but across different indices. By changing the offsets we are essentially moving a 3-D window that offers views into a 3-D correlation space. To compute correlations for $k=4$, we also held an index for one of four TFs constant in addition to holding a gene's index constant.

By varying a gene index, a single transcription factor index, and adjusting the kernel's grid offsets we were able to compute a total of 2,013,884,773,648 (≈ 2 trillion) correlation coefficients. Both the CUDA/GPU-based and CPU-based

analyses were done using single-precision floating-point numbers and calculations. Our architecture uses 4 bytes of memory to store a single-precision floating point number; the total number of correlation coefficients computed therefore corresponds to about 8 terabytes of data. Because we do not have such memory capacity available, as we analyzed combinations of genes and TFs, we recorded combinations (as well as the correlations) that had improvement scores above an arbitrary threshold of 0.35. The top 65,536 combinations were recorded; the remaining data points were not recorded.

CUDA kernels, because of the specialized GPU architecture on which they execute, run faster with fewer control flow statements. Our code could have computed a hypothetical expression profile conditioned on checks that none of the component expression profiles had missing values by looking to see if any of the values was missing (-18). To avoid such control flow statements, a *boolean* flag was created from logical AND and in-equality testing operations on the values. In the code, the flag was used as an indicator whose value was interpreted as zero or one. The indicator was incorporated into computations within a *for* loop; intermediate values and counters were adjusted accordingly. During development, this change led to a dramatic speedup.

3 Results

Our C/C++ $k=4$ CUDA-based analysis led to two results: a) putative heterotetrameric TF complexes and target genes along with the corresponding coefficients sorted by their improvement scores and b) timing data for comparison with the CPU-based implementation.

Table 2 presents the top 10 genes and putative TF-tetramers of our analysis results. The CUDA-accelerated program ran in approximately 4.6 days. Figure 4 displays GPU speedup against estimated OpenMP thread running times (1, 2, 3, and 4 OpenMP threads).

Table 2. The top-scoring genes and hypothetical transcription factors from the CUDA-based $k=4$ analysis. Legend: AI "Abs. improvement"

	AI	GENE	TF1	TF2	TF3	TF4
1	0.73	SERPINA6	EPC1	PLAGL1	WT1	ZNF10
2	0.73	FGB	IRF1	MGA	PAPOLA	SNAPC3
3	0.73	FGB	PRKAR1A	TWISTNB	ZNF155	ZNF83
4	0.72	FGB	EPC1	HMGB2	ITGB3BP	SP110
5	0.72	FGB	EPC1	ITGB3BP	PAPOLA	SP110
6	0.71	AFP	BCL6	ID4	SIAH2	ZNF212
7	0.70	FGB	E2F5	MHGB2	MGA	SP110
8	0.70	FGB	HMGB2	MGA	SP110	ZNF83
9	0.70	FGB	TWISTNB	ZNF155	ZNF198	ZNF83
10	0.70	FGB	E2F5	HMGB2	ITGB3BP	SP110

Interestingly, we note that in the top 10 results from the CUDA-based $k=4$ analysis that the FGB gene is seemingly overrepresented as well as the SP110 transcription factor. FGB forms the beta portion of fibrinogen. The protein helps form blood clots. The SP110 transcription factor plays a role forming a part of a leukocyte-specific nuclear-body [14, 23].

We submit these top results to the body of scientific literature as candidates for subjects of further research and inquiry. In addition, the complete list of over 65,000 putative target genes, correlations, and tetrameric TF complexes, dataset and source code are available from the corresponding author of this paper as well.

4 Discussion

We here discuss the efficacy of our algorithm, the role of CUDA in it, its execution, and ways to possibly improve it by parameter adjustment and tuning. We also discuss further ways to test the technique. Finally, we discuss its role of the in a greater bioinformatics context.

Regarding efficacy we note how the program detected two out of four known target genes for the AP-1 complex in the top ten listed target genes (out of 3166 transcripts total). This outcome suggests that the algorithm has some value, but that to be more precise, it needs to be improved. To further explore the algorithm's efficacy, other known dimers and their target genes could be considered and the program's output could be analyzed similarly as was done in this paper.

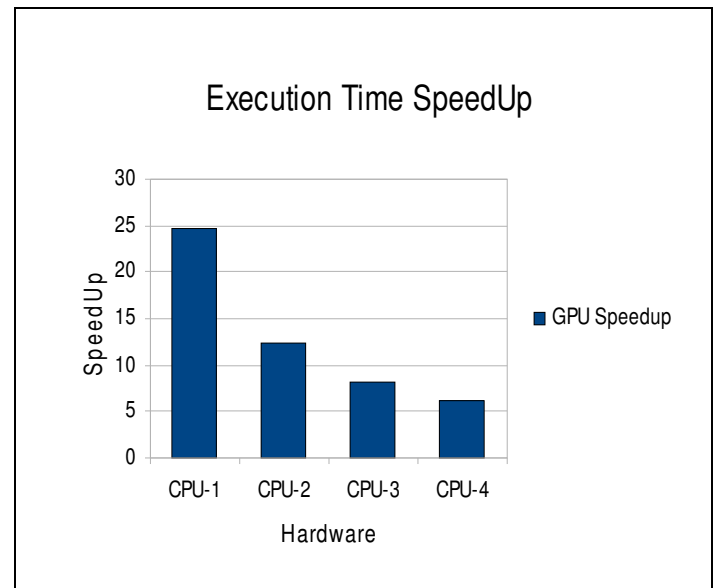


Fig. 4 The speedup (GPU vs. CPU) achieved by the CUDA-based implementation of our algorithm compared with OpenMP threads (1...4).

All of the gene expression profiles were subjected to an alpha transform. The parameterization of alpha leaves a place for experimentation, adjustment, and hopefully improvement. In our analyses for this paper, α was set to 0.5. Perhaps, known heterodimers and their target genes could be set to vary, so that alpha, on a per-dataset basis, could be variable and calibrated or optimized to reveal the most known heteropolymeric transcription factors and their target genes as possible.

Our program computed all of the correlation coefficients with the GPU. Their computations and subsequent comparisons of the absolute improvement score for sorting were carried out by the CPU. The majority of the program's execution time was spent doing such things. This suggests that having the GPU carry out such calculations presents a future avenue to expand the use of the GPU and further contract the running time of the program. Such use of the hardware will require further and continued use of the CUDA API to coordinate kernel calling and data transfers between GPU memory and host memory.

Our dataset set included a total of 44,886 missing values (40,080 in the gene dataset, 4806 in the TF dataset). Thus, with a total of $3166+352=3518$ expression profiles, there is an average of approximately 11.4 missing values per expression profile. From this point of view, every composite hypothetical expression profile of two TFs would have approximately at least that many missing values. Thus, for a given tissue (of 115 total), the probability that its expression value is missing is about 9.93%. Using the binomial distribution, for a hypothetical dimer there is a nearly $19\% = P(X \geq 1 | n=2; p=0.0993)$ chance that a given tissue's data is missing. For three TFs this value is just over 25%. For four, it is nearly 35%. The more TFs that are under consideration for a given tissue, then the more likely that at least one component TF expression value is missing increases at that tissue. Thus, for higher order composite expression profiles, many tissue expression values would be missing. Thus, for $k=4$, any results should perhaps be used with some caution. To make such analysis more meaningful, missing values could be estimated, but any results from analyses with such imputed values we believe should similarly be used with some, but perhaps less caution. In addition, as k increases, because there are more missing values, the signal-to-noise ratio also increases and that is a further reminder for using results from higher-order analyses with some caution.

Such ideas cause us to remember the fact that the "gold standard" techniques to definitively tell whether or not two or more proteins heteropolymerize include standard "wet lab" molecular biology techniques. Such techniques include crystallography and co-immunoprecipitation (co-IP) [30]. Crystallography [29] involves actual structures, crystallized and examined as 3D structures; co-IP extracts protein-protein-DNA complexes from a solution using antibodies. Such techniques however, are relatively time-consuming and expensive. Moreover, as the number of combinations of proteins whose polymerization is considered increases, more experiments and procedures are necessary to determine whether they bind or not. This means more time and money is needed to make such determinations. Thus our technique explored in this paper may have some value in saving time and money.

To our knowledge, CUDA has never been used to implement this particular technique for microarray data-mining for TF complexes. A somewhat related work for microarray analysis,

the TSP algorithm has also been ported to CUDA [27]. Another exists as well that computes correlations and is integrated into the R package for statistical computing [28]. We note that our CUDA kernels' correlation coefficients here are distinct from other CUDA kernels' coefficients in that here minima are taken.

5 Conclusion and Summary

In conclusion, we have presented a set techniques used to analyze a microarray dataset by computing correlation coefficients between gene expression profiles and transcription factor expression profiles across tissues. Its goal is to find multiple transcription factors that bind together and have a target gene whose transcription is modulated. The technique involves hypothetical heteromeric transcription factor profiles whose expressions are estimated by taking minima for each tissue. A scoring function based on a comparison among the correlation coefficients is used to sort and prioritize combinations of genes and transcription factors. The higher scoring combinations are though to be more likely to form transcription factor complexes for the gene. We presented some test data showing the efficacy of our program; it gave interesting results in revealing some 2 out of 4 true positives with a P -value of $8.4 \cdot 10^{-5}$. To consider 4 TFs at a time, the computational demands are high, so we explored using CUDA-enabled NVIDIA GPUs to speed up the computations. We achieved speedups of about 6x. For analyzing whether or not four TFs bind, we completed an analysis and have presented some the results from that analysis. Finally, we discussed some of the strengths and weaknesses of the algorithm and our CUDA-implemented technique to speed it up; we also mentioned some ways that the technique could be further improved.

6 Acknowledgements

We acknowledge Dr. Michael Allan for providing ideas for validating the technique and biological insights too. We also acknowledge Dr. Stephen Kwek (Medio Systems) for guidance in implementing the algorithm. All programming was done by Edward A. Salinas.

Funding: All funding for the computer hardware was provided by Edward A. Salinas.

7 References

- [1] A. Karmaker, E. Salinas, S. E. Harris and S. Kwek, *Identifying Correlations between Genes and Transcription Co-factors using Expression Profile.*, JCIS, 2007.
- [2] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, et al., *Initial sequencing and*

- analysis of the human genome*, Nature, 409, pp. 860-921, 2001.
- [3] J. W. Fickett and W. W. Wasserman, *Discovery and modeling of transcriptional regulatory regions*, Curr Opin Biotechnol, 11, pp. 19-24, 2000.
- [4] L. A. McCue, W. Thompson, C. S. Carmack and C. E. Lawrence, *Factors influencing the identification of transcription factor binding sites by cross-species comparison*, Genome Res, 12, pp. 1523-32, 2002.
- [5] M. Defrance and H. Touzet, *Predicting transcription factor binding sites using local over-representation and comparative genomics*, BMC Bioinformatics, 7, pp. 396, 2006.
- [6] A. E. Kel, E. Gossling, I. Reuter, E. Chermushkin, O. V. Kel-Margoulis and E. Wingender, *MATCH: A tool for searching transcription factor binding sites in DNA sequences*, Nucleic Acids Res, 31, pp. 3576-9, 2003.
- [7] M. C. Frith, M. C. Li and Z. Weng, *Cluster-Buster: Finding dense clusters of motifs in DNA sequences*, Nucleic Acids Res, 31, pp. 3666-8, 2003.
- [8] C. T. Workman and G. D. Stormo, *ANN-Spec: a method for discovering transcription factor binding sites with improved specificity*, Pac Symp Biocomput, pp. 467-78, 2000.
- [9] M. C. Frith, U. Hansen, J. L. Spouge and Z. Weng, *Finding functional sequence elements by multiple local alignment*, Nucleic Acids Res, 32, pp. 189-200, 2004.
- [10] K. Ellrott, C. Yang, F. M. Sladek and T. Jiang, *Identifying transcription factor binding sites through Markov chain optimization*, Bioinformatics, 18 Suppl 2, pp. S100-9, 2002.
- [11] W. Ao, J. Gaudet, W. J. Kent, S. Muttumu and S. E. Mango, *Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR*, Science, 305, pp. 1743-6, 2004.
- [12] W. B. Alkema, O. Johansson, J. Lagergren and W. W. Wasserman, *MSCAN: identification of functional clusters of transcription factor binding sites*, Nucleic Acids Res, 32, pp. W195-8, 2004.
- [13] R. Shyamsundar, Y. H. Kim, J. P. Higgins, K. Montgomery, M. Jorden, A. Sethuraman, et al., *A DNA microarray survey of gene expression in normal human tissues*, Genome Biol, 6, pp. R22, 2005.
- [14] Entrez *Gene*
<http://www.ncbi.nlm.nih.gov/entrez/http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene>,
- [15] E. Salinas, A. Karmaker, BioComp 2009 Analysis of Correlations between Genes and Triads of Transcription Factors Using Microarray Expression Profiles.
- [16] The NVIDIA CUDA Programming Guide
http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf
- [17] Watson, et. al., Mol. Biology of the Gene, 6th Edition, 2008
Microarray Expression Profiles.
- [18] Lee, et. al., Coexpression Analysis of Human Genes Across Many Microarray Data Sets, Genome Res. 2004 June; 14(6): 1085-1094 .
- [19] Farber, Rob; CUDA Application and Development, MK Press, 2011
- [20] E. Salinas, A. Karmaker, Analysis of Correlations Between Genes and Tetrads of Transcription Factors Using Microarray Expression Profiles, Proc. Of BioComp 2010, Las Vegas, NV, USA
- [21] S. Falcon and R. Gentleman Using GOSTats to test gene lists for GO term association Bioinformatics (2007) 23(2): 257-258
- [22] W. Ewens, G Grant, Statistical Methods in Bioinformatics, an Introduction, 2nd Edition, Springer, 2005
- [23] Sayers et. al., Database Resources of the National Center for Biotechnology Information, Nucleic Acids Res. (2009) 37(suppl 1): D5-D15
- [24] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites, Nucl. Acids Res., (1996) 24(1): 238-241
- [25] D. Thomas, et al., The ENCODE Project at UC Santa Cruz, Nucl. Acids Res.(2007) 35(suppl 1): D663-D667
- [26] Halazonetis TD et al., CJUN Dimerizes with CFOS, Forming Complexes of different DNA Binding Affinities, Cell. 1998 Dec. 2; 55(5):917-924
- [27] Magis A., et al., Graphics processing unit implementations of relative expression analysis algorithms enable dramatic computational speedup Bioinformatics (2011) 27(6): 872-873
- [28] Buckner, et. al., The gputools package enables GPU computing in R Bioinformatics (2010) 26(1): 134-135
- [29] Park, Young-Jun, et. al., Crystal structure of a heterodimer of editosome interaction proteins in complex with two copies of a cross-reacting nanobody; Nucl. Acids Res. (2011) doi: 10.1093/nar/gkr867
- [30] Zhang L., et. al., Successful co-immunoprecipitation of Oct4 and Nanog using cross-linking, Biochem Biophys Res Commun. 2007 September 28; 361(3): 611-614

BIOCAMP

Study of Classification Accuracy of Microarray Data for Cancer Classification using Hybrid, Wrapper and Filter Feature Selection Method

Sujata Dash¹, Bichitrananda Patra²

¹ Department of CSE, Gandhi Institute for Technology, Bhubaneswar, Orissa, India

² Department of CSE, KMBB College of Engineering and Technology, Bhubaneswar, Orissa, India

Abstract - *Microarray analysis are becoming a powerful tool for clinical diagnosis, as they have the potential to discover gene expression patterns that are characteristic for a particular disease. This problem has received increased attention in the context of cancer research, especially in tumor classification. Various feature selection methods and classifier design strategies also have been used and compared. Feature selection is an important pre-processing method for any classification process. Selecting a useful gene subset as a classifier not only decreases the computational time and cost, but also increases classification accuracy. In this study, we applied the correlation-based feature selection method (CFS), which evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them as a filter approach, and three wrappers (J48, Random Forest and Random Trees) to implement feature selection; selected gene subsets were used to evaluate the performance of classification. Experimental results show that by employing the proposed method fewer gene subsets are need to be selected to achieve better classification accuracy.*

Key Words: Microarrays, Hybrid Method, Filter Method, Wrapper Method, Correlation Based Feature Selection

1 Introduction

DNA microarray technology allows simultaneous monitoring and measuring of thousands of gene expression activation levels in a single experiment. This technology is currently used in medical diagnosis and gene analysis. Many microarray research projects focus on clustering analysis and classification accuracy. In clustering analysis, the purpose of clustering is to analyze the gene groups that show a correlated pattern of the gene expression data and provide insight into gene interactions and function. Research on classification accuracy is aimed at building an efficient model for predicting the class membership of data, produce a correct label on training data, and predict the label for any unknown data correctly.

Typically, gene expression data possess a high dimension and a small sample size, which makes testing and training of general classification methods difficult. In general, only a relatively small number of gene expression data out of the total number of genes investigated shows a significant correlation with a certain phenotype. In other

words, even though thousands of genes are usually investigated, only a very small number of these genes show a correlation with the phenotype in question. Thus, in order to analyze gene expression profiles correctly, feature selection (also called gene selection) is crucial for the classification process. Methods used for data reduction, or more specifically for feature selection in the context of microarray data analysis, can be classified into two major groups: filter and wrapper model approaches [28].

In the filter model approach a filtering process precedes the actual classification process. For each feature a weight value is calculated, and features with better weight values are chosen to represent the original data set. However, the filter approach does not account for interactions between features. The wrapper model approach depends on feature addition or deletion to compose subset features, and uses evaluation function with a learning algorithm to estimate the subset features. This kind of approach is similar to an optimal algorithm that searches

for optimal results in a dimension space. The wrapper approach usually conducts a subset search with the optimal algorithm, and then a classification algorithm is used to evaluate the subset.

Several machine learning algorithms have already been applied to classifying tumors using microarray data. Voting machines and self-organising maps (SOM) were used to analyse acute leukemia [10]. Support vector machines (SVMs) were applied to multi-class cancer diagnosis by [21]. Hierarchical clustering was used to analyse colon tumor [1]. The best classification results are reported by Li et al.[17] and Antonov et al. [2]. Li et al [17]. employed a rule discovery method and Antonov et al.[2] maximal margin linear programming (MAMA). Given the nature of cancer microarray data, which usually consists of a few hundred samples with thousands of genes as features, the analysis has to be carried out carefully. Work in such a high dimensional space is extremely difficult if not impossible. One straightforward approach to select relevant genes is the application of standard parametric tests such as the *t*-test [24][25] and a nonparametric test such as the Wilcoxon score test[24][3]. Wilks's Lambda score was proposed by [13] to access the discriminatory power of individual genes. A new procedure [2] was designed to detect groups of genes that are strongly associated with a particular cancer type.

In this paper we have applied two general approaches of feature subset selection, more specifically, wrapper and filter approaches and then created a new model called hybrid model by combining the characteristics of the two specified models for gene selection. We compared the gene selection performance of the filter model, wrapper model and hybrid model. Wrappers and filters differ in how they evaluate feature subsets. Filter approaches remove irrelevant features according to general characteristics of the data. Wrapper approaches, by contrast, apply machine learning algorithms to feature subsets and use cross-validation to evaluate the score of feature subsets. Most methods of gene selection for microarray data analysis focus on filter approaches, although there are a few publications on applying wrapper approaches[14] [29] [28]. Nevertheless, in theory, wrappers should provide more accurate classification results than filters [15]. Wrappers use classifiers to estimate the usefulness of feature subsets. The use of “tailor-made” feature subsets should provide a better classification accuracy for the corresponding classifiers, since the features are selected according to their contribution to the classification accuracy of the classifiers. The disadvantage of the wrapper approach is its computational requirement when combined with sophisticated algorithms such as support vector machines.

As a filter approach, correlation-based feature selection (CFS) was proposed by Hall[12]. The rationale behind this algorithm is “a good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other.” It has been shown in Hall [12] that CFS gave comparable results to the wrapper and executes many times faster.

To evaluate and compare the proposed method to other feature selection methods, we used two classification algorithm namely, the K-nearest neighbour (KNN) and a Support Vector Machine (SVM) to evaluate the selected features, and to establish the influence on classification accuracy. The results indicate that in terms of the number of genes that need to be selected and classification accuracy of the proposed method is superior to other methods in the literature. The rest of this paper is organised as follows. We begin with a brief overview introducing the methods presented in Section 2. The experimental framework and settings are described in Section 3. Section 4 consists of the results and a theoretical discussion thereof. Finally, the conclusion and future work is presented in Section 5.

2 Related Methods

2.1 Feature subset selection

We now define the basic notions used in the paper. Given a microarray cancer data set D , which contains n samples from different cancer types or subtypes, we have to build a mathematical model which can map the samples to their classes. Each sample has m genes as its features. The assumption here is that not all genes measured by a microarray are related to cancer classification. Some genes are irrelevant and some are redundant from the machine learning point of view. It is well-known that the inclusion of irrelevant and redundant information may harm performance of some machine learning algorithms. Feature

subset selection can be seen as a search through the space of feature subsets. One major problem of *filters* that score individual features is the selection of a threshold by which to discard features. Although all the features will be given a score by the filter algorithm, it is not clear how to determine the optimal threshold for the data. One heuristic approach (the so called $n - 1$ rule) in microarray cancer analysis chooses the top $n - 1$ genes to start the analysis[16]. Golub et al. [11] chose 50 genes most closely correlated with leukemia subtypes. Nevertheless, ranking genes by filters does present an overall picture of the microarray data.

In general, *filters* are much faster than *wrappers* [31]. However, as far as the final classification accuracy is concerned, *wrappers* normally provide better results. The general argument is that the classifier that will be built from the feature subset should provide a better estimate of accuracy than a separate measure that may have an entirely different classification bias. The main disadvantage of *wrapper* approaches is that during the feature selection process, the classifier must be repeatedly called to evaluate a subset. For some computationally expensive algorithms such as SVMs or artificial neural networks, wrappers can be impractical.

2.2 The choice of filter algorithms and classifiers

2.2.1 Correlation-based feature selection

CFS evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them [12].

$$CFS_S = \frac{k \bar{r}_{cf}}{\sqrt{k + k(k-1)r_{ff}}} \quad (1)$$

where CFS_S is the score of a feature subset S containing k features, \bar{r}_{cf} is the average feature to class correlation ($f \in S$), and \bar{r}_{ff} is the average feature to feature correlation. The distinction between normal filter algorithms and CFS is that while normal filters provide scores for each feature independently, CFS presents a heuristic “merit” of a feature subset and reports the best subset it finds.

2.2.2 Support Vector Machines (SVMs)

SVMs are relatively new types of classification algorithms. An SVM expects a training data set with positive and negative classes as an input (i.e. a binary labelled training data set). It then creates a decision boundary (the maximal-margin separating boundary) between the two classes and selects the most relevant examples involved in the decision process (the so-called support vectors). The construction of the linear boundary is always possible as long as the data is linearly separable. If this is not the case, SVMs can use kernels, which provide a nonlinear mapping to a higher dimensional feature space. The dot product has the following formula:

$$K(x, y) = (x \cdot y + 1)^d \quad (2)$$

where x and y are the vectors of the gene expression data. The parameter d is an integer which decides the rough shape of a separator. In the case where d is equals to 1, a

linear classification algorithm is generated, and in the case where d is more than 1, a nonlinear classification algorithm is generated. In this paper, when d is equals to 1, it is called the SVM dot product, when d is equals to 2, it is called the SVM quadratic dot product and when d is equals to 3, it is called the SVM cubic dot product. The radial basis kernel is as follows,

$$K(x, y) = \exp\left(\frac{-|x - y|^2}{2\sigma^2}\right) \quad (3)$$

where σ is the median of the Euclidean distances between the members and non-members of the class. The main advantages of SVMs are that they are robust to outliers, converge quickly, and find the optimal decision boundary if the data is separable [7]. Another advantage is that the input space can be mapped into an arbitrary high dimensional working space where the linear decision boundary can be drawn. This mapping allows for higher order interactions between the examples and can also find correlations between examples. SVMs are also very flexible as they allow for a big variety of kernel functions. Sequential minimal optimization (SMO) [20] is used in this paper to train an SVM. SVMs have been shown to work well for high dimensional microarray data sets [10]. However, due to the high computational cost it is not very practical to use the wrapper method to select genes for SVMs, as will be shown in our experimental results section.

2.2.3 k-nearest Neighbour

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

2.2.4 Decision Trees- J48, Random Forest, Random Trees

In decision tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. There are advantages with decision tree algorithms: they are easily converted to a set of production rules, they can classify both categorical and numerical data, and there is no need to have a priori assumptions about the nature of the data. However multiple output attributes are not allowed in decision tree and algorithms are unstable. Slight variations in the training data can result it different attribute selections at each choice point within the tree. The effect can be significant since attribute choices affect all descendent sub-trees [27]. ID3 (Iterative Dichotomiser 3)

is an algorithm used to generate a decision tree. Developed by J. Ross Quinlan [21], ID3 is based on the Concept Learning System (CLS) algorithm [19].

J48 is an improved version of ID3 algorithm. It contains several improvements, including: choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, and handling continuous attributes [21]. Random forest is another classifier that consists of many decision trees. It outputs the class that is the mode of the classes output by individual trees [6][8].

3 Experimental procedure

The experiments were performed with the Weka machine learning package [26]. We used the following three general strategies to identify predictive features.

3.1 Selecting genes using CFS

- a) Choose a search algorithm.
- b) Perform the search, keeping track of the best subset encountered according to CFS.
- c) Output the best subset encountered.

3.2 Selecting genes using a wrapper method

- a) Choose a machine learning algorithm to evaluate the score of a feature subset.
- b) Choose a search algorithm.
- c) Perform the search, keeping track of the best subset encountered.
- d) Output the best subset encountered.

The search algorithm we used was best-first with forward selection, which starts with the empty set of genes. In this paper we report accuracy estimates for classifiers built from the best subset found during the search. Once the best subset has been determined, then a classifier evaluates the performance of the subset selected.

4 The Proposed Hybrid Method

In this study, we hybrid the filter and wrapper model methods to select feature genes in microarrays, and used two different classification algorithms to evaluate the performance of the proposed method[18]. Figure 1 depicts the process of the hybrid filter and wrapper model feature selection method.

For example, let a microarray data set have 10 gene numbers (10 feature numbers which can be represented by $f1 f2 f3 f4 f5 f6 f7 f8 f9 f10$). If only 5 genes ($f1, f2, f4, f7$ and $f10$) conform to the CFS selection, only these 5 genes ($f1 f2 f4 f7 f10$) are used for the wrapper procedure to implement the selection process. However, when using the filter model selection, the feature number could be reduced dramatically. In order to remove more effectively unwanted features, we used wrappers namely, J48, Random Forest and Random Trees after the initial filter model selection to select features again, and then applied KNN and SVM algorithm to measure the classification performance.

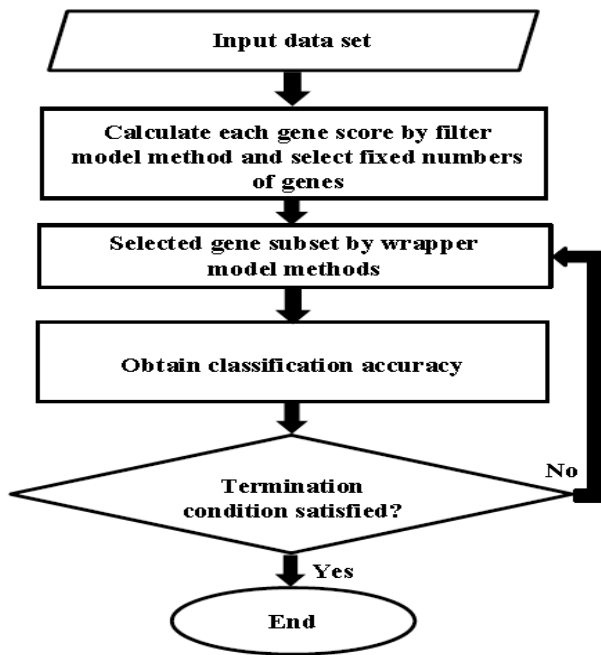


Figure1. Hybrid filter and wrapper model feature selection method

5 Experimental Results and Comparison

In this section, we perform comprehensive experiments to compare the CFS-J48, CFS-Random Forest and CFS-Random Tree selection algorithm with CFS filter algorithm and the wrapper algorithms (J48, Random Forest and Random Tree) on three different datasets using two different classifiers SVM and KNN.

5.1 Datasets description and pre-processing

To evaluate the usefulness of the CFS-J48, CFS-Random Forest and CFS-Random approaches, we carried out experiments on three datasets of gene expression profiles. The datasets and their characteristics are summarized in Table 1. The data is taken from <http://sdmc.lit.org.sg/GEDatasets/Datasets.html>.

- The Colon tumor dataset consists of 62 microarray experiments collected from colon-cancer patients with 2000 gene expression levels. Among them, 40 tumor biopsies are from tumors and 22 (normal) biopsies are from healthy parts of the colons of the same patients.
- The Leukemia dataset consists of 72 microarray experiments with 7129 gene expression levels. Two classes for distinguishing: Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). The complete dataset contains 25 AML and 47 ALL samples.
- The Lung cancer dataset involves 181 microarray

experiments with 12533 gene expression levels. Classification occurs between Malignant Pleural Mesothelioma (MPM) and Adenocarcinoma (ADCA) of the lung. In tissue samples there are 31 MPM and 150 ADCA.

Note that in these datasets, the samples in each class is generally small, and unevenly distributed. This, together with the large number of classes makes the classification task more complex. The original gene expression data are continuous values. We pre-processed the data so each gene has zero mean value and unit variance. We also discretized the data into categorical data to reduce noise.

We discretized the observations of each gene expression variable using the respective σ (standard deviation) and μ (mean) for this gene's samples: any data larger than $\mu + \sigma/2$ were transformed to state 1; any data between $\mu + \sigma/2$ and $\mu - \sigma/2$ were transformed to state 0; any data smaller than $\mu - \sigma/2$ were transformed to state -1. These three states correspond to the over expression, baseline, and under-expression of genes.

5.2 Parameter Settings

We used Weka, a well known comprehensive toolset for machine learning and data mining [4], as our main experimental platform. We evaluated the performance of feature selection methods in Weka environment with two classifiers, using 10-fold Cross Validation .

To evaluate the performance of the proposed method, the selected feature subsets were evaluated by K-fold cross validation (K-fold) for KNN and SVM classifiers. For K-fold cross validation, we set K=10 in this study.

During K-fold cross-validation, the data was separated into 10 parts $\{D_1, D_2, \dots, D_{10}\}$, and training and testing was carried out a total of 10 times. When any part D_n , $n = 1, 2, \dots, 10$ is processed as a test set, the other 9 parts will be training sets. Following 10 times of training and testing, 10 classification accuracies are produced, and the averages of these 10 accuracies are used as the classification accuracy for the data set. We assumed that the obtained classification accuracy is an adaptive functional value.

5.3 Results and Comparison

- We started experiment by evaluating performance accuracies of both the classifiers, SVM and KNN on the three datasets using 10-fold Cross Validation (CV) without using feature selection algorithms. The result of the 10-fold CV accuracy for the two classifiers are shown in table 5.
- After feature selection, the selected feature subsets were evaluated using two common classification algorithms SVM and KNN using 10-fold CV method. Table 2 and Table 3 show the accuracies achieved by the filter (CFS with a best-first search), wrapper (J48, RF, RT using best-first search) and hybrid model (wrapper method

Table 1. Cancer related human gene expression datasets

Dataset	# of genes	# of samples	# of classes	# of positive samples	# of negative samples
Leukemia	7129	72	2	47(ALL)	25(AML)
Lung Cancer	12533	181	2	31(MPM)	150(ADCA)
Colon Cancer	2000	62	2	22	40

Table 2. KNN Accuracy performance of three microarray data sets for the Filter, Wrapper and Hybrid feature selection method.

KNN (Statnikov et al)[22]		Filter	Wrapper			Hybrid		
Dataset		CFS	J48	RF	RT	CFS+J48	CFS+RF	CFS+RT
Colon		87.10	95.16	82.26	82.26	85.48	87.10	82.26
Leukemia	83.57	98.61	93.06	88.89	90.28	95.83	98.61	94.44
Lung Cancer		99.45	99.45	99.45	96.13	99.45	99.45	98.43

Table 3. SVM Accuracy performance of three microarray data sets for the Filter, Wrapper and Hybrid feature selection method.

	Filter	Wrapper			Hybrid			SVM (NO FS)	
Dataset	CFS	J48	RF	RT	CFS+J48	CFS+RF	CFS+RT	Akadi et al [23]	Statnikov et al[22]
Colon Cancer	75.48	87.10	79.03	75.81	89.03	87.10	85.48	85.48	
Leukemia	87.22	91.67	95.83	90.28	95.83	97.22	93.06	98.61	97.50
Lung Cancer	95.45	99.45	97.45	96.13	100	99.24	98.34	87.67	

and CFS in conjunction with a best-first search) feature selection methods individually. In Table 2, the classification accuracy is evaluated by KNN and in Table 3 by SVM.

- The experimental results show that the accuracy of microarray data which had feature selection implemented was better than without feature selection. Comparing filter and wrapper selection methods, the accuracy of the wrapper model was better than for the filter model, and the number of selected feature was smaller for the wrapper model than for the filter model which can be observed from Table 4.

- The J48, Random Forest (RF) and Random Tree (RT) wrapper models differ from the filter model in that it is dependent on a classifier and evaluates the combination of feature subsets using 10-fold CV internally. The wrapper model can identify interaction amongst all features simultaneously. However, how many gene subsets are truly necessary to identify cancer categories is still a question under debate [21].

- But filter selection does not reduce the number of features very much; hence another method is needed to reduce the number of features further. In order to select more effective feature subsets, we used wrapper models namely, J48, Random Forest(RF) and Random Tree(RT) algorithms after implementing the filter approach.

- Again, we can observed from Table 2 and Table 3 that the proposed method effectively increases classification accuracy and selects a smaller number of feature subsets. During the wrapper phase of the proposed method, we have implemented the same wrapper model and this method returns very small sets of genes compared to alternative variable selection methods, while retaining predictive performance. Our method of gene selection will not return sets of genes that are highly correlated, because they are redundant. This method will be most useful under two scenarios:

- when considering the design of diagnostic tools, where having a small set of probes is often desirable;
- to help understand the results from other gene selection approaches that return many genes, so as to understand which ones of those genes have the largest signal to noise ratio and could be used as surrogates for complex processes involving many correlated genes.

A best first search with forward direction, searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.

Table4. Number of feature selected for the three microarray datasets using Filter, Wrapper and Hybrid feature selection method.

	Filter	Wrapper			Hybrid		
Dataset	CFS	J48	RF	RT	CFS + J48	CFS + RF	CFS+ RT
Colon Cancer	26	3	4	3	2	9	5
Leukemia	81	2	2	4	2	3	3
Lung Cancer	161	2	2	2	2	2	2

Table 5. 10-fold cross validation accuracy (%) with all features

Dataset	SVM Accuracy	KNN Accuracy
Leukemia	68.06	80.56
Lung Cancer	76.24	92.82
Colon Cancer	80.65	82.26

The experiment showed that the combination of decision tree wrapper model with a correlation based filter method achieves a better performance than CFS or single wrapper model.

Compared to previous works, it should be noted that

without using feature selection Statnikov et al.[4] have obtained 83.57% accuracy for Leukemia dataset using KNN classifier. Whereas, our result is 80.56% without using selection method, 98.61% using CFS filter, 90.73% average classification performance of all three wrappers and 96.29% average accuracy using proposed hybrid method.

For multi class SVM with no feature selection, they obtained 2.50% error in Leukemia data classification and 2.39% by Akadi et al.,[5]. On the other hand with binary SVM classifier the rate of error of our result using CFS was 2.72%, 7.41% average error of all three wrappers and 4.67% average error of all three hybrid filter methods. For Colon dataset, our result obtained for hybrid filter CFS-J48, CFS-RF and CFS-RT were better than Akadi et al.,[5]. For Lung dataset, we obtained 100% result for J48 wrapper and CFS-J48 hybrid filter and almost 98% for rest of the methods. Whereas Akadi et al.,[5] obtained only 87.67% classification accuracy in their work.

We believe that our results will motivate more microarray practitioners to use wrappers and hybrid using CFS as their analysis tools. These machine learning algorithms are implemented in WEKA, a publicly available open-source software package. This software can be used both by experienced and novice users. WEKA has been already applied in a number of bioinformatics studies as reviewed elsewhere [9].

6 Conclusion

In this paper, we hybrid the filter and wrapper model methods for microarray classification to implement a feature selection process, and then used KNN and SVM to evaluate the classification performance. Experimental results showed that the proposed method simplified gene selection and the total number of parameters needed effectively, thereby obtaining a higher classification accuracy compared to other feature selection methods. The classification accuracy obtained by the proposed method was comparatively higher than other methods for all three test problems. In the future, the proposed method can assist in further research where feature selection needs to be implemented. It can potentially be applied to problems in other areas as well.

7 References

- [1] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natl. Acad. Sci.*, 1999, 96 (12), 6745–6750.
- [2] Antonov, A.V., Tetko, I.V., Mader, M.T., Budczies, J., Mewes, H.W., "Optimization models for cancer classification: extracting gene interaction information from microarray expression data", *Bioinformatics* 20, 2004, 644–652.
- [3] Antoniadis, A., Lambert-Lacroix, S., Leblanc, F., "Effective dimension reduction methods for tumor classification using gene expression data", *Bioinformatics* 19,2003, 563–570.
- [4] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, Vol. 21, 2005, No. 5, pp 631–643.
- [5] A. E. Akadi, A. Amine, A. E. Ouardighi, D.Aboutajdine, , "Feature selection for Genomic data by combining filter and wrapper approaches", *INFOCOMP Journal of computer science*,2009, vol. 8, no. 4, pp. 28-36.
- [6] Breiman Leo, Cutler Adele,, "Random Forest", *Machine Learning Conference Paper for ECE591Q*,2010, 25 Apr.
- [7] Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D.Haussler., *Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines.*, *Proc. Natl. Acad. Sci. USA.*,1999, 97: 262-267.
- [8] Diaz-Uriarte R, Alvarez de Andres S,, " Gene selection and classification of microarray data using random forest", *BMC Bioinformatics* ,2006, 7:3.
- [9] Frank E, Hall M, Trigg L, Holmes G, Witten IH: *Data mining in bioinformatics using Weka. Bioinformatic* 20(15),2004,:2479-2481.
- [10] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M.,Haussler, D., "Support vector machine classification and validation of cancer tissue samples using microarray expression data", *Bioinformatics* 16,2000, 906–914.
- [11] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.H.H.C., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E., " Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science* 286,1999, 531–537.
- [12] Hall, M.A.," Correlation-based feature selection for machine learning", *Ph.D. Thesis. Department of Computer Science, University of Waikato*, 1999.
- [13] Hwang, D., Schmitt, W.A., Stephanopoulos, G., Stephanopoulos, G., "Determination of minimum sample size and discriminatory expression patterns in microarray data", *Bioinformatics* 18,2002, 1184–1193.
- [14] Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A., " Filter versus wrapper gene selection approaches in DNA microarray domains", *Artif. Intell. Med.*,2004, 31 (2), 91–103.
- [15] Langley, P., " Selection of relevant features in machine learning", *Proceedings of AAAI Fall Symposium on Relevance*,1994, pp. 140–144.
- [16] Li, W., Yang, Y., " How many genes are needed for a discriminant microarray data analysis. *Methods of Microarray Data Analysis*", *Kluwer Academic Publishers*,2002, pp. 137–150.
- [17] Li, J., Liu, H., Ng, S.-K., Wong, L., *Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics* 19,2003 93ii–102ii.

[18] Li-Yeh Chuang, Chao-Hsuan Ke, and Cheng-Hong Yang, Member, *IAENG*, “A hybrid both filter and wrapper feature selection method for microarray classification”, Proceedings of the International MultiConference of Engineers and Computer Scientists ,2008,Vol I, 19-21 March, Hong Kong.

[19] Mitchell Tom M, “ Machine Learning”, *McGraw-Hill* 1997.

[20] Platt, J.,” Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods–Support Vector Learning*”,1998, MIT Press.

[21] Quinlan J.R.,, “C4.5: Programs for Machine Learning”,1993, *Morgan Kaufmann Publishers*.

[22] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo,M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., Golub, T., Multiclass cancer diagnosis using tumor gene expression signatures.,2001, *Proc. Natl. Acad. Sci.* 98 (26), 15149–15154.

[23] Shi et al., “ Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction”, *BMC Bioinformatics* 2011., <http://www.biomedcentral.com/1471-2105/12/375>.

[24] Thomas, J.G., Olson, J.M., Tapscott, S.J., Zhao, L.P. “ An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles”, *Genome Res.*,2001, 11, 1227–1236.

[25] Tsai, C.-A., Chen,Y.-J., Chen, J.J., “ Testing for differentially expressed genes with microarray data”, *Nucl. Acids Res.*,2003 31, e52.

[26] Witten, I.H., Frank, E., “ Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”,1999, Morgan Kaufmann.

[27] Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW:, “Gene selection from microarray data for cancer classification--a machine learning approach”, *Comput Biol Chem* , 2005,29(1):37-46.

[28] Xing, E., Jordan, M., Karp, R., “ Feature selection for high-dimensional genomic microarray data”, Proceedings of the 18th International Conference on Machine Learning, 2001.

[29] Xiong, M., Fang, X., Zhao, J., “ Biomarker identification by feature wrapper”, *Genome Res.*,2001, 11 (11), 1878–1887.

[30] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, 23(19), 2007, pp. 2507-2517.

[31] Y. Wang et al.,” Gene selection from microarray data for cancer classification—a machine learning approach”, *Computational Biology and Chemistry*,2005 29 , 37–46.

SESSION

BIOINFORMATICS DATABASES, DATA MINING, AND PATTERN DISCOVERY TECHNIQUES

Chair(s)

TBA

A Cluster-based Approach for Biological Hypothesis Testing and its Application

Jiwon Park, Jin Soung Yoo[†], and Ahmed Mustafa[‡]

[†]Department of Computer Science, and [‡]Department of Biology
Indiana University-Purdue University, Fort Wayne, Indiana, USA
{parkj01,yooj,mustafaa@ipfw.edu}

Abstract—This interdisciplinary study investigated computational analytic methods used for biological hypothesis testing, and applied the methods for the validation of the effects of nutraceuticals on growth and immune response of Nile tilapia, *Oreochromis niloticus*, in cool water. Farmers in cooler regions face problems with cultivating tilapia, one of the most popular cultivated fish species, due to poor survival rates at suboptimal temperatures. We hypothesized that two nutraceuticals, phosphatidylcholine and β -carotene, help tilapia adapt to cooler water temperatures, and benefit tilapia's growth and immune response. This hypothesis testing problem was managed using an unsupervised learning technique in data mining and statistics called cluster analysis. The significance of clustering results are often computed using external indexes and internal indexes. We show, in particular, that the external index can be used for testing the biological hypothesis by formulating the level of agreement between two different partitions of samples: experimental groups and clusters based on the similarity of features. Contrary to the findings of previous studies which showed the beneficial effects of phosphatidylcholine and β -carotene supplementation in a range of fish including tilapias, our test result shows no significant difference among the fish reared in cool water and fed with either the basal diet or diets supplemented with the nutraceutical. This study also shows our computational approach is a promising analytic tool for similar hypothesis testing in biology domain including fish biology.

Index Terms—tilapia, temperature, stress, phosphatidylcholine, beta-carotene, data mining, clustering

I. INTRODUCTION

Tilapia has quickly become one of the most popular cultivated fish species around the world [1]. Originating from the African Great Lakes, this hardy, warm water fish is raised in many regions of the world including indoor and outdoor ponds, tanks, and waterways. However, farmers in cooler regions face problems with low survival rates of tilapia in suboptimal temperatures [2]. Although tilapia species have proved to be very hardy, the decrease in growth rate and increased rates of disease at low temperatures are the main factors preventing cultivation at cooler temperatures, or in cooler regions.

The stress experienced by tilapia when subjected to suboptimal water temperatures elicits similar effects as other common stressors present in aquaculture such as handling, sorting, grading, and transporting [3], [4]. In recent years dietary supplements termed “nutraceuticals” have been used in an effort to combat the stress-induced effects in aquaculture. Previous studies [5], [6] have shown positive effects of phosphatidylcholine (PC) and β -carotene (BC) supplementation in many fish species. We investigated that two nutraceuticals, PC

and BC help tilapia adapt to cooler water temperatures via improved membrane fluidity, and improve growth and immune response.

Experiments were conducted in two different environment: warm water ($28 \pm 1^\circ\text{C}$) and cool water ($16 \pm 1^\circ\text{C}$), with differing the nutraceutical supplement in a basal diet, over an 8-week period. The sample data was collected with a set of features such as length, weight, condition factor, plasma concentration of glucose, hematocrit and phagocytic capacity of macrophage cells, subject to measurements in different experimental conditions and times.

Traditionally, statistical analysis plays critical roles in the interpretation of experimental data across the life sciences, including fish biology. A scientific question creates a set of hypothesis tests to conclude statistically significant differences among samples from different treatment groups. Each hypothesis test is conducted in a series of formal stages. The first stage is to state the *null* hypothesis (H_0). The null hypothesis states that a biological treatment has no effect, or there is no difference between treated and untreated populations. For example, in the case of two groups, “control” and “treated” samples, the null hypothesis can be $\theta_1 - \theta_2 = 0$, where θ is a “statistic” (e.g., mean) on the data. The second stage states the alternative hypothesis (H_A) which proposes the opposite of the null hypothesis. In the third stage, the statistics derived from the data are compared with the statement of the null hypothesis. Various methods exist to test whether there is a difference in the treatment, for example, between control and treated samples. Statistical exploratory methods [7] such as descriptive statistics, correlations and t-Test are typically used in fish biology [8], [9]. Statistically significant differences between samples (typically, $\rho < 0.05$) indicate that the observed difference is unlikely to have occurred by chance, suggesting a “real” difference between control and treated samples.

The null hypothesis is tested against each feature or a set of features in the data for the scientific question, but each result may not be of primary interest. Multiple hypothesis testing eventually confirms the original scientific question. Our work shows that the multiple hypothesis testing problem can be managed with an unsupervised learning technique in data mining and statistics, i.e., *cluster analysis*, and the cluster-based approach is applied to test the effect of the two nutraceuticals (PC and BC) on the growth and immune response of Nile tilapia in cool water.

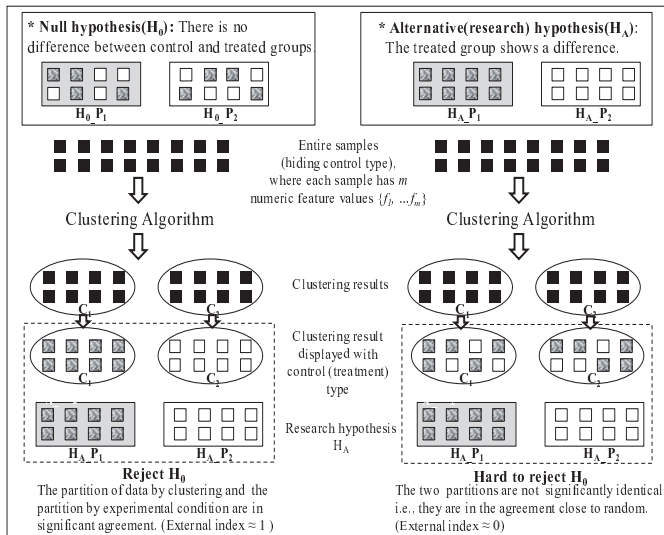


Fig. 1. Our methodology for hypothesis testing

In Section 2, we explain our methodology for hypothesis testing in detail. Section 3 presents our design decision and procedure. The test results with real experimental tilapia data are presented in Section 4. Section 5 describes related work, and ends with conclusion.

II. OUR METHODOLOGY FOR HYPOTHESIS TESTING

Fig. 1 illustrates our methodology for hypothesis testing. A sample in the experimental data $D = \{s_1, \dots, s_n\}$ can be formalized as a numerical vector $s_i = (f_{i1}, \dots, f_{im})$, where f_{ij} is the value of the j th feature for sample s_i where $1 \leq i \leq n$ and $1 \leq j \leq m$. Each s_i can be represented as a data point in m -dimensional space. In Fig. 1, for simplicity, samples are displayed in the 2-dimensional space. Clustering algorithms seek to partition a given data set into groups so that data objects within a group are more similar to each other than data objects in different groups [10], [11], [12].

Groups based on the characteristics data possesses are called *clusters*. The objects in each cluster are “similar” according to the value of a distance or similarity function on their features. Two sample data s_i and s_j are in the same cluster if and only if $(f_{i1}, \dots, f_{im}) \simeq (f_{j1}, \dots, f_{jm})$. In Fig. 1, samples are divided into two clusters, C_1 and C_2 .

Clustering is called to *unsupervised classification* because clustering does not rely on predefined classes (or labels) and training examples while grouping the data objects, and aims to separate the data into a finite and discrete set of “natural” structures. Therefore, samples which show similar responses to a biological treatment are expected to be in the same cluster. Our main idea of hypothesis testing with this unsupervised learning technique is to first divide samples based on common features and properties through clustering processing, and then compare the clustering result (i.e., clusters) with experimental groups.

Cluster validation is the process of assessing the quality and reliability of the cluster sets that are derived from various

Partition/Cluster	C_1	C_2	...	C_k	Sums
P_1	$n_{1,1}$	$n_{1,2}$...	$n_{1,k}$	$n_{1.}$
P_2	$n_{2,1}$	$n_{2,2}$...	$n_{2,k}$	$n_{2.}$
...
P_k	$n_{k,1}$	$n_{k,2}$...	$n_{k,k}$	$n_{k.}$
Sums	$n_{.1}$	$n_{.2}$...	$n_{.k}$	$n_{..=n}$

TABLE I
CONTINGENCY TABLE

clustering processes. Cluster validation techniques have the potential to provide an analytical assessment of the amount and type of data distribution captured by grouping. External indexes and internal indexes are commonly used for cluster validation [13], [14], [15], [11], [16]. We use the external index for formulating the level of agreement between two different partitions of sample data: experimental groups and clusters based on the similarity of features.

External indexes are usually defined via a *contingency table*. Let $C = \{C_1, \dots, C_k\}$ be a partition of samples in D into k clusters, and $P = \{P_1, \dots, P_k\}$ be another partition of D into k groups. P is an external partition of the data, derived from biological experimental controls, while C is a partition obtained by a clustering algorithm. Let $n_{i,j}$ be the number of samples in both C_i and P_j , $1 \leq i \leq k$ and $1 \leq j \leq k$. Moreover, let $|C_i| = n_{i.}$ and $|P_j| = n_{.j}$. These values can be conveniently arranged in a contingency table as shown in Table I. An external index E is a function, which takes as input, the contingency table values, and returns a value p to assess how close the cluster C is to the reference partition P . The external index value p is 1 when the two partitions are identical, and 0 when they are selected at random. That means p close to 0 indicates a level of significance in the agreement close to random, and the null hypothesis of no difference (H_0) cannot be rejected.

III. DESIGN DECISION AND PROCEDURE

Our hypothesis testing procedure is summarized in Fig. 2. This section describes our design decision for each step in the procedure.

A. Data collection

For the experiment, fingerling Nile tilapias (mean weight 7.5 grams and mean length 6.8 cm) *Oreochromis niloticus* were randomly stocked in eight tanks containing dechlorinated water (four groups \times two replicates) at a density of 25 fish per tank. After the acclimation period, the room temperature was lowered to achieve the cool water temperature of 16°C, while 100W heaters were used to maintain the optimal warm water treatment at 28°C. A basal diet which is a commercial floating pelleted feed, and L- α -phosphatidylcholine (PC), and β -carotene(BC) supplements were used. The four fish groups were: *warm water control* (28 \pm 1°C), *cool water control* (16 \pm 1°C), *cool water with PC*, and *cool water with BC*. In this paper, the terms *warm water* and *cool water* are used to describe fish reared at 28°C and 16°C, respectively.

For the collection of experimental data, six fish per treatment group (three fish per tank \times two replicates) were randomly sampled during the experimental phase at weeks 0, 2, 4, 6, and 8. The health and stress levels of these fish were determined by the measures of following features: length, weight, condition factor, plasma glucose, blood hematocrit, and phagocytic capacity of macrophage. The condition factor from length and weight was calculated according to [17]. The equation is $Condition\ factor = (weight \times 100)/length^3$. To determine blood glucose levels, methods were followed as given by [18], [3]. Isolation of macrophages and assessment of their phagocytic activity was accomplished by the technique described in [9].

B. Data preprocessing and exploration

The first step in data analysis is data preprocessing [19]. Sample data may have noisy, missing, and inconsistent values. There are a number of data preprocessing tasks: data cleaning, data integration, data transformation, and data reduction [20], [21]. For data cleaning, missing values in our samples were filled with the median value of the feature of the missing value. Task relevant data is selected and transformed into an input data format which can be fed to data mining algorithms.

Exploratory data analysis is used for general insight into data. Descriptive statistics such as count and average are used to quantitatively describe the main features in a collection of data [22].

C. Clustering

Clustering methods are applied to the pre-processed data. The experimental condition type of each sample is hidden in this stage.

1) *Clustering algorithms*: There are many clustering methods: partition-based methods, hierarchical-based methods, density-based methods and grid-based methods [20], [23], [24], [25]. We limit ourselves to the class of clustering algorithms that take as input D and an integer k , and return the k clusters of D , since the clustering result should be compared with pre-defined experimental groups. Partition-based method and hierarchical-based methods are included in this category.

The K-means algorithm [46] is a typical partition-based clustering method. Given a specified number k , the algorithm partitions a data set into k disjointed subsets in which each data object belongs to the subgroup with the nearest centroid. The K-means algorithm is known for being simple and fast. In contrast to partition-based clustering, hierarchical clustering generates a hierarchical series of nested clusters that can be graphically represented by a tree, called a dendrogram. The branches of a dendrogram not only record the formation of the clusters, but indicate the similarity between the clusters as well. By cutting the dendrogram at a certain level, we can obtain a specified number of clusters.

We used clustering algorithms in WEKA (Waikato Environment for Knowledge Analysis), which is a popular suite of machine learning software [26], [27]. WEKA has *KMeans* for partition-based clustering and *HierarchicalClusterer* for hierarchical-based clustering.

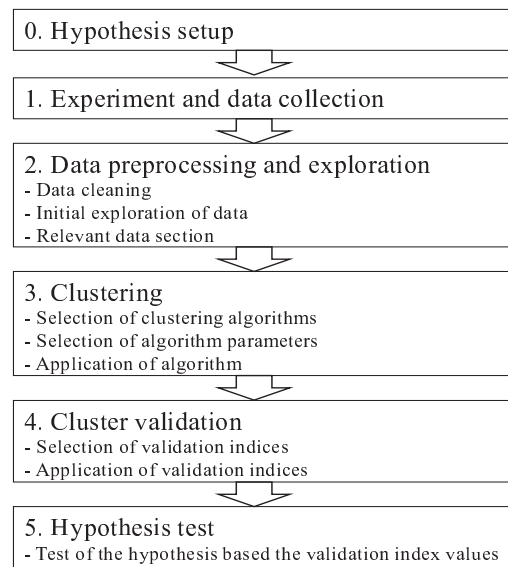


Fig. 2. Our procedure for hypothesis testing

2) *Similarity measures*: Inter object similarity is a measure of the correspondence or resemblance between objects to be clustered. The specification and formalization of similarity between data objects, depend heavily on the application domain [14], [23]. Distance measures of similarity are often used to compare objects whose characteristics are measured with quantitative variables. There are several popular proximity functions such as Euclidean distance, Pearson's correlation coefficient, and Spearman's rank-order correlation coefficient. The most commonly used distance measure is Euclidean distance. On the other hand, if object characteristics are measured with qualitative variables, association measures of similarity are used. Since all features in our sample data are quantitative, we used the Euclidean distance to measure the similarity of two samples s_i and s_j .

$$Euclidean(s_i, s_j) = \sqrt{\sum_{l=1}^m (f_{il} - f_{jl})^2}, \text{ where } s_i = \{f_{i1}, \dots, f_{im}\}, \text{ where } f_{il} \text{ is the value of the } l\text{th feature for the } i\text{th sample, } 1 \leq l \leq m$$

D. Cluster validation

Most clustering algorithms do not provide estimates of the significance of the cluster results returned. The verification of clustering results is therefore based on a manual, lengthy and subjective exploration process. Generally, cluster validity has three aspects. First, the quality of clusters is measured in terms of *homogeneity* and *separation* based on the definition of a cluster: "Objects within one cluster are similar to each other, while objects in different clusters are dissimilar with each other." The second aspect relies on a given "ground truth" of the clusters. The ground truth could come from domain knowledge. The third aspect of cluster validity focuses on the reliability of the clusters or the likelihood that the cluster structure is not formed by chance. The first aspect is often

validated using an internal index, and the second and third aspects are examined using an external index.

1) *Internal index*: Internal indexes are used to measure how well the data partitioned by a clustering algorithm corresponds to the natural cluster structure of data. They are internal because the quality of the partition is measured according to information contained in the dataset without resorting to external knowledge. We use internal indexes to estimate the quality of clustering solutions generated from different clustering algorithms. The representative internal indexes are Within Cluster Sum of Squares (WCSS) [20], [19], [11], KL: the Krzanowski and Lai index [15], Gap statistics [28], and so on. WCSS was used in our work since KL is based on WCSS, and Gap statistics is for a special case of clustering, i.e., single cluster. WCSS is defined as following: let $C = \{C_1, \dots, C_k\}$ be a clustering solution, with k clusters.

$$WCSS(k) = \sum_{i=1}^k \sum_{s \in C_i} |s - \mu_i|^2, \text{ where } s \text{ is a data object}$$

in cluster C_i and μ_i is the centroid of cluster C_i .

2) *External index*: External indexes are used to formulate the level of agreement between two different partitions. We used external indexes to formulate the level of agreement between a partition of samples by the experimental control, and another partition of the sample as the output of a clustering algorithm. Rand index [29], Adjusted Rand index [16], Fowlkes and Mallows (FM) index [30], and Fowlkes (F)-index [31] are representative external indexes. Adjusted Rand index (R_A) was used in this study since it is a statistic often recommended in the classification literature [32]. Adjusted Rand index can be defined with the contingency table values in Table I as:

$$R_A = \frac{\sum_{ij} \binom{n_{i,j}}{2} - \frac{[\sum_i \binom{n_{i,j}}{2}] \sum_j \binom{n_{i,j}}{2}}{\sum_i \binom{n}{2}}}{\frac{1}{2}[\sum_i \binom{n_{i,j}}{2}] + \sum_j \binom{n_{i,j}}{2} - \frac{[\sum_i \binom{n_{i,j}}{2}] + \sum_j \binom{n_{i,j}}{2}}{\binom{n}{2}}]}$$

Rand index R is similar with adjusted Rand index R_A , but it does not tell how significant is the concordance between two partitions as measured by the value of R . Therefore, R_A is often used instead of R .

E. Hypothesis test

The adjust Rand index R_A has a maximum value of 1, indicating a perfect agreement between the two partitions, while its expected value of 0 indicated a level of agreement due to chance. When a biological hypothesis is tested with adjusted Rand index R_A , R_A must be a non negative value substantially away from 0 so that the null hypothesis can be rejected.

IV. TEST RESULTS

The clustering-based approach described in the previous sections was used to validate the effect of two nutraceuticals on growth and immune response of tilapia in cool water. This section presents the results.

We first compare the performance of two different clustering algorithms, K-means and Hierarchical clustering on our experimental data, and then present the hypothesis test results with all features, and with selected features.

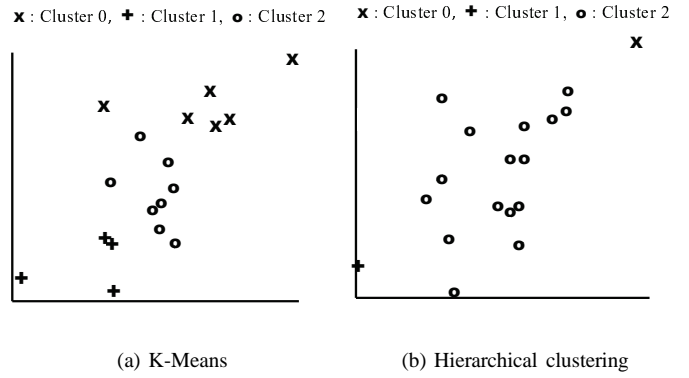


Fig. 3. Clustering algorithm performance

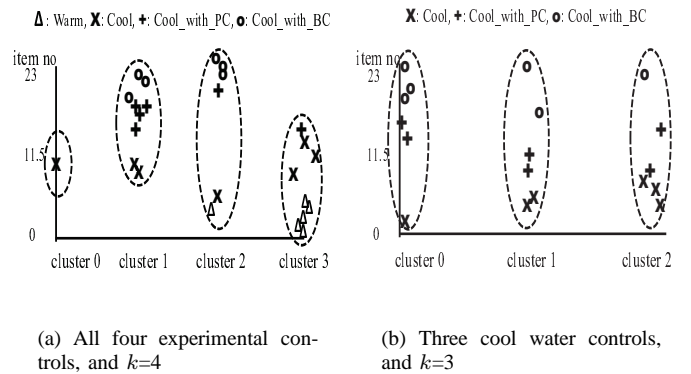


Fig. 4. Clustering results of all features

1) *Algorithm performance*: The quality of clustering results generated by K-means and Hierarchical clustering was measured using an internal index, WCSS. Fig. 3 shows the clustering results of growth related data. In this experiment, the number of clusters was 3. The WCSS of K-means was 75.68, and the WCSS of Hierarchical clustering was 247.24. K-means showed better performance than Hierarchical clustering showing a much smaller WCSS value. Clustering results with other feature data showed similar performances. Since K-means showed better clustering performance than Hierarchical clustering, we chose K-means for our hypothesis test.

2) *Test with all features*: In this experiment, we conducted the clustering process with all feature data of samples at week 8. Fig. 4 (a) shows the result. We can notice that samples in four experimental groups are distributed over four different clusters although most of the warm control samples are included in the cluster 3. Although fish reared at the optimal water temperature (i.e., warm) make a distinct group, it is hard to find a difference in fish reared in cool water. The adjusted Rand index between the two partitions, {Warm, Cool, Cool_with_PC, Cool_with_BC} by the experimental control and {cluster0, cluster1, cluster2, cluster3} by clustering, was 0.15. We can also notice some samples from cool water are

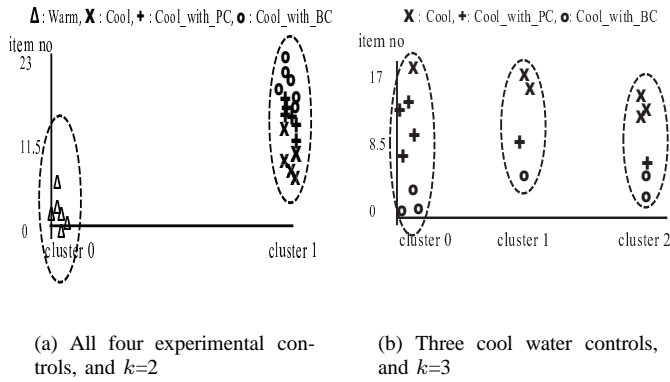


Fig. 5. Clustering results of growth related features

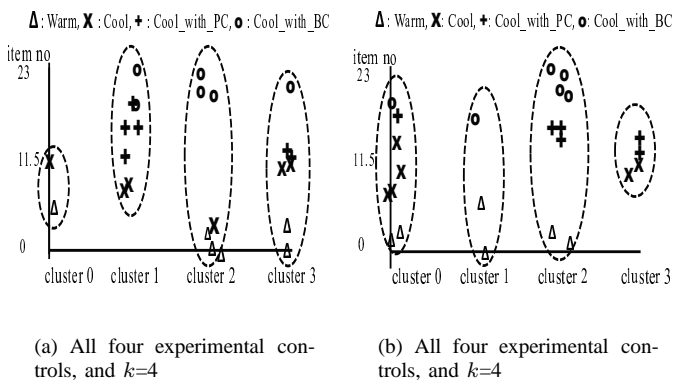


Fig. 6. Clustering results of stress and immunity related features

in the same cluster with warm control samples. That indicates that fish reared in 16°C adapt to lower temperature, and remain healthy.

To closely examine the effect of PC or BC in cool water groups, we conducted the clustering with samples from only cool water. Fig. 4 (b) shows the clustering result. The adjust Rand index between the two partitions, {Cool, Cool_with_PC, Cool_with_BC} and {cluster0, cluster1, cluster2} was -0.07 which is far from 1. With this index value, our null hypothesis cannot be rejected. That means there is no evidence of a beneficial effect of PC or BC to tilapia in cool water.

3) *Test with growth features:* In this experiment, the effect of the two nutraceuticals on growth was examined. Fig. 5 (a) shows the clustering result based on growth related features when $k=2$. One cluster, cluster0, includes all warm water samples, and cluster1 includes other samples. The adjusted Rand index between two clusters and {Warm, Cool} was 1, which shows a perfect agreement between the two partitions. The growth of fish reared in warm water was significantly different than the fish reared in cool water. On the other hand, we compared samples from cool water with different diets. Fig. 5 (b) shows the clustering result. The adjusted Rand index between {Cool, Cool_with_PC, Cool_with_BC} and {cluster0,

cluster1, cluster2} was -0.127. No significant differences in growth were found when comparing the fish reared in cool water regardless of the diet.

4) *Test with stress and immunity features:* Fig. 6 (a) shows the clusters based on stress related features. The adjust Rand index was 0.011. It is hard to reject our null hypothesis with this value. The stress measurements were not significantly different for fish irrespective of the water temperature and diet. The result with the immunity related feature data is shown in Fig. 6 (b). The adjust Rand index was 0.08. No significant differences in phagocytic capacity were seen among the different groups regardless of temperature and supplement.

V. DISCUSSION

We first describe a brief review of related work in biology domain and computer science domain, and end with the summary of this study.

A. Related Work

In aquaculture biology literature, several works [33], [34], [35], [36], [5] have demonstrated the beneficial effects of PC supplementation in a range of fish species including tilapias. Farkas et al. [37] demonstrated involvement of PC in increasing membrane fluidity in fish during adaption to lower water temperature. Hu et al. [6] showed β -carotene, a dimer of vitamin A, increases growth rate in Nile tilapia. Blazer et al. [38] showed β -carotene has a role in disease resistance in fish. However, there is no previous study on the effect of PC or BC to tilapia in suboptimal water temperatures.

A very rich literature on cluster analysis has developed in statistics and data mining over the past decades [39], [23], [40], [10], [11], [12]. Xu et al. [23] conducted a survey of clustering algorithms for data sets appearing in statistics, computer science, and machine learning. In the bioinformatics area, Jian et al. [41] gave a survey of cluster analysis for gene expression data. Since clustering algorithms have been proved useful for identifying biologically relevant groups of genes, many conventional clustering algorithms have been adapted or directly applied to gene expression data. Handl et al. [42] explored computational cluster validation methods in post-genomic data analysis. Kerr et al. [43] showed bootstrapping cluster analysis for assessing the reliability of conclusions from microarray experiments. However, as we know, there is no work which uses clustering techniques for hypothesis testing in fish biology domain.

B. Conclusion

This interdisciplinary work explored clustering methods and clustering validation measures, and used them for the validation of the effects of two nutraceuticals on the growth and immune response of tilapia in cool water. Contrary to the findings of previous studies which showed the beneficial effects of PC and BC supplementation in a range of fish including tilapias, our test results showed no significant difference of growth and immune response in the tilapia fed the nutraceutical supplemented diet in cool water, showing external index values

close to 0. Thus, it is clear that there is no need to incur the expenses of using BC or PC to enhance growth during cool weather. It can be concluded that tilapia adapt to temperature stress without deleterious physiological consequences. The effect of PC of BC supplementation has no significant effect on physiological or immunological response.

Our conclusion is consistent with the results proven by traditional statistical methods in [8]. In [8], the means and their standard errors of each feature in four different experiment controls were compared, and tested with one-way analysis of variance (ANOVA) and $\rho < 0.05$. Our cluster-based approach is particularly good for a multivariate hypothesis test where all features are simultaneously considered. Our study shows that the cluster-based approach is a promising analytic tool for similar hypothesis tests in fish biology. In the future, to show the scalability and effectiveness of our approach, we plan to include many other biomarker features like plasma cortisol, plasma glucose, hematocrit, leukocrit, RBC numbers, hemoglobin, plasma protein, internal cell size, metabolic rate, and hypoxia.

REFERENCES

- [1] C. Kohler, "A White Paper on the Status and Needs of Tilapia Aquaculture in the North Central Region," north Central Regional Aquaculture Center.
- [2] R. Shell, "Farming Tilapia," 1993, <http://www.thefishsite.com/articles113/farming-tilapia>.
- [3] C. B. Schreck and P. B. Moyle, *Methods for Fish Biology*. MD: American Fisheries Society, 1990.
- [4] B. A. Barton and G. K. Iwama, "Physiological Changes in Fish from Stress in Aquaculture with Emphasis on the Response and Effects of Corticosteroids," *Annual Review Fish Dis.*, vol. 1, pp. 3–26, 1991.
- [5] C. S. Kasper and P. B. Brown, "Growth Improved in Juvenile Nile Tilapia Fed Phosphatidylcholine," *North American Journal of Aquaculture*, vol. 65, pp. 39–43, 2003.
- [6] C. J. Hu, S. M. Chen, C. H. Pan, and C. H. Huang, "Effects of Dietary Vitamin A or β -carotene Concentrations on Growth of Juvenile Nile Hybrid Tilapia, *Oreochromis niloticus* x *Oreochromis aureus*," *Aquaculture*, vol. 253, pp. 602–607, 2006.
- [7] J. W. Tukey, *Exploratory Data Analysis*. Addison Wesley, 1977.
- [8] A. Mustafa, L. Randolph, and S. Dhawale, "Effect of Phosphatidylcholine and Beta-Carotene Supplementation on Growth and Immune Response of Nile Tilapia, *Oreochromis niloticus*, in Cool Water," *Applied Aquaculture*, vol. 2, pp. 136–146, 2011.
- [9] A. Mustafa, C. MacWilliams, N. Fernandez, K. Matchett, G. Conboy, and J. Burka, "Effect of sea lice (*Lepeophtheirus salmonis* Kröyer, 1837) infestation on macrophage functions in Atlantic salmon (*Salmo salar* L.)" *Fish Shellfish Immunol.*, vol. 10, pp. 47–59, 2000.
- [10] B. Everitt, S. Landau, and M. Leese, *Cluster Analysis*. London: Arnold, 2001.
- [11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2003.
- [12] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.
- [13] R. Dubes, "How Many Clusters are Best?-An Experiment," *Pattern Recognition*, vol. 20, no. 6, pp. 645–663, 1987.
- [14] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.
- [15] W. Krzanowski and Y. Lai, "A Criterion for Determining the Number of Groups in a Dataset using Sum of Squares Clustering," *Biometrics*, vol. 44, pp. 23–34, 1985.
- [16] L. Hubert and P. Arabie, "Comparing Partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.
- [17] G. P. Busacker, I. R. Adelman, and E. M. Goolish, "Stress and acclimation," *In Methods for fish biology, edited by C. Schreck & P.B. Moyle, American Fisheries Society*, pp. 363–388, 1990.
- [18] M. Gensic, R. Wissing, and A. Mustafa, "Effects of iodized feed on stress modulation in steelhead trout, *Oncorhynchus mykiss* (Walbaum)," *Aquaculture Research*, vol. 35, pp. 1117–1121, 2004.
- [19] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, Second Edition*. Morgan Kaufmann, 2005.
- [20] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison Wesley, ISBN 0321321367, 2005.
- [21] J. R. Quilan, "Unknown Attribute Values in Induction," in *Proc. of International Workshop on Machine Learning*, 1989.
- [22] D. Freedman, R. Pisani, and R. Purves, *Statistics*. W. W. Norton & Company, 1997.
- [23] R. Xu and D. W. II, "Survey of Clustering Algorithm," *IEEE Transactions on Neural Networks*, vol. 15, no. 3, pp. 645–678, 2005.
- [24] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1988.
- [25] A. D. Gordon, *Clustering Algorithms and Cluster Validation*. Computational Statistics (P. Dirschedl and R. Ostermann, editors), 1996.
- [26] I. H. Witten and M. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.
- [27] "Weka (Waikato Environment for Knowledge Analysis)," <http://www.cs.waikato.ac.nz/ml/weka/>.
- [28] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the Number of Clusters in a Dataset via the Gap Statistics," *Journal of Royal Statistical Society B*, vol. 2, pp. 411–423, 2001.
- [29] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.
- [30] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association*, vol. 78, pp. 553–584, 1983.
- [31] C. V. Rijsbergen, *Information Retrieval*. Butterworths, London, UK, 1979.
- [32] G. W. Milligan and M. C. Cooper, "A study of the Comparability of External Criteria for Hierarchical Cluster Analysis," *Multivariate Behavioral Research*, vol. 21, pp. 441–458, 1986.
- [33] S. Craig and D. M. Gatlin, "Growth and Body Composition of Juvenile red drum (*Sciaenops ocellatus*) Fed Diets Containing Phosphatidylcholine and Supplemental Choline," *Aquaculture*, vol. 151, pp. 259–268, 1997.
- [34] A. Kanazawa, *Puffer fish Fugu rubripes: Handbook of Nutrient Requirements of Finfish*. CRC Press, 1991.
- [35] H. A. Poston, "Performance of Rainbow Trout Fry Fed Supplemental Soy Lecithin and Choline," *Fish-Culturist*, vol. 52, pp. 218–225, 1990.
- [36] H. Poston, "Effect of Body Size on Growth, Survival, and Chemical Composition of Atlantic Salmon Fed Soy Lecithin and Choline," *Fish-Culturist*, vol. 52, pp. 226–230, 1990.
- [37] T. Farkas, E. Fodor, K. Kitajka, and J. E. Halver, "Response of fish membranes to environmental temperature," *Aquaculture Research*, vol. 32, pp. 645–655, 2001.
- [38] V. S. Blazer, "Nutrition and Disease Resistance in Fish," *Annual Review of Fish Diseases*, vol. 2, pp. 309–323, 1992.
- [39] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," in *Proc. of International Conference on Data Mining*, 2004.
- [40] S. M. Savareasi and D. Boley, "A Comparative Analysis on the Bisecting K-Means and the PDDP Clustering Algorithms," *Intelligent Data Analysis*, vol. 8, no. 4, pp. 345–362, 2004.
- [41] D. Jian, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.
- [42] J. Handl, J. Knowles, and D. B. Kell, "Computational Cluster Validation in Post-Genomic Data Analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.
- [43] M. K. Kerr and G. A. Churchill, "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments," *National Academy of Sciences*, vol. 98, pp. 8961–8965, 2001.

Prediction of Proteins' Molecular Surfaces from their Corresponding Amino Acid Sequences

J. Zhao¹, E. Paquet^{1,2} and H. L. Viktor¹

¹School of Electrical Engineering and Computer Science, University of Ottawa, 800 King Edward, Ottawa, Ontario, K1N 6N5, Canada

²National Research Council, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada

Abstract - *This paper presents a new approach for the prediction of proteins' molecular surfaces from their amino acid sequences, using a multilayer artificial neural network. Our novel approach allows one to learn, and to predict, a pose-invariant shape index describing the molecular surface of a protein, by starting from a descriptor of the amino acid sequence. The input layer of the neural network is formed by a set of probability density functions associated with the correlation in between the constituent residues, while the output layer is formed by a pose invariant shape index describing the molecular surface. Once the neural network is trained, our molecular surface search engine allows one to retrieve the closest known molecular surface associated with an unknown, so-called query protein. We test our method against a database of more than 45,000 molecular surfaces. The neural network is evaluated for various topologies and optimization methods, and yield promising results.*

Keywords: Amino Acid, Correlation, Indexing, Invariant, Macromolecule, Molecular Surface, Neural Network, Protein, Structure

1 Introduction

The prediction of a protein structure from the corresponding amino acid sequence is of great importance in bioinformatics and yet, it is a demanding task for machine learning algorithms. The knowledge of the macromolecular structure is essential in order to understand the protein functions and the biological processes [1]. Proteins are macromolecules formed by one or many chains of amino acids. The chemical properties of their residues, their mutual interaction as well as their interaction with the surrounding environment determine their three-dimensional structure through a process called folding. There are many repositories from which the amino acid sequence of proteins may be obtained: for instance, the Universal Protein Resource [2]. While the amino acid sequences are known for a vast number of proteins, a relatively small number of three-dimensional structures have been experimentally determined. This is due to the fact that experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) are time consuming and labour intensive. In contrast, amino acid sequences may be

obtained through highly efficient automated high throughput experimental methods. For all these reasons, it is important to bridge the gap in between the two approaches. That is, there is an urgent need for solutions to aid domain experts to predict proteins' three-dimensional structures solely from their corresponding amino acid sequences.

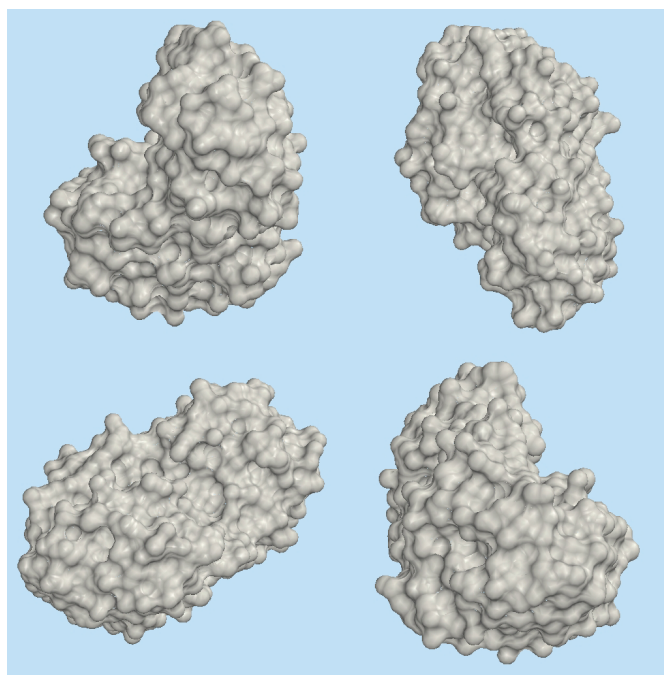


Figure 1. Four views of the molecular surface of protein phage T4 lysozyme from bacteriophage T4 (142l).

A number of approaches have been proposed in the literature, which rely on an artificial neural network in order to predict the structure of a protein from its amino acid sequence [3, 4]. The network is first trained with a set of known amino acid sequences and their corresponding macromolecular structures. Once the training phase is completed, unknown three-dimensional structures may be predicted from known amino acid sequences. However, most of these approaches are limited to the prediction of the secondary structures which is a composite structure constituted of three basic elements namely the helix, the beta-sheet, and the coil [5]. More complex structures may be

obtained through a process known as homology modelling or threading [6] which exploits the fact that two proteins whose sequences are evolutionarily connected display similar structural features.

In this paper, we propose a new approach that aims to predict directly the molecular surface of a given protein from its amino acid sequence. Our motivation is as follows. The molecular surface is directly responsible for the interactions between proteins and, consequently, is of prime importance in order to better understand their functions and their mutual interaction. The prediction of the molecular surface or envelope is a complex task. Let us consider Fig. 1 which shows four views of protein phage T4 lysozyme from bacteriophage T4 (PDB: 1421). One immediately notices that the shape is highly complex and, as opposed to the secondary structures, there are no obvious basic geometrical elements, for instance the helix, from which the surface may be constructed.

In order to address this problem, we propose an approach based on a feed-forward artificial neural network. Because the molecular surface has a complex shape, we first compute a descriptor or index that characterizes the entire three-dimensional shape of the molecular surface. The shape index, which shall be described later, is formed of the probability density functions (PDF) associated with the radial and angular distributions of the surface elements forming the molecular surface. The structure and the size of the index are independent on the underlying shape, which makes it a perfect candidate to train a neural network. That is, the number and the meaning of the neurons are the same for each protein. In addition, this index is invariant under rotation and translation, which implies that the underlying macromolecule may have any spatial orientation or pose. That means that it is not required to align the proteins into a common orientation, prior to training the neural network.

The input of the neural network is formed from the probability density functions associated with the correlation in between adjacent amino acid: for instance, the first nearest neighbour, the second nearest neighbour, etc. As for the shape index, the structure and size of the probability density functions do not depend on the underlying amino acid sequence. This fact makes them suitable candidates in order to train a neural network. Once the network has been trained, it is possible to predict the descriptor associated with the amino acid sequence. The molecular surface is then obtained by searching through a database of shape indexes for the closest known molecular surface. We have created such a database by analyzing more than 45,000 proteins from the Protein Data Bank [7].

The paper is organized as follows. In Section 2, we present our approach for the description of amino acid sequences. Then, in Section 3, we describe our shape index or descriptor. This is followed, in Section 4, with a depiction of the neural network and the training process. We analyze the relative merit of various configurations for the neural network in addition to evaluate the effectiveness of many optimization methods for calculating the weights of the connexions.

Finally, in Section 5, we describe our approach for molecular surface prediction and present some preliminary experimental results. Our main conclusions are presented as well as some potentially promising research directions.

2 Amino Acid Sequences Description

Approaches for the prediction of macromolecular structures are based on the assumption that there exist a strong correlation in between the three-dimensional structure and the underlying amino acid sequence. This is one of the main motivations for using neural networks. That is, one may train the network with known pairs of amino acid sequences and three-dimensional structures and then determine an unknown structure from the corresponding amino acid sequence and the trained neural network.

Multilayer neural networks have a fix number of input and output neurons. Consequently, in order to train efficiently the neural network, the structure of the input and output layer should not depend, for instance, on the length of the amino acid sequence or on the complexity of the corresponding molecular surface. That situation potentially constitutes a problem. Although similar three-dimensional structures tend to have similar amino acid sequences, it does not mean that they are identical. Indeed, similarity here means that they share a high number of similar subsequences. For various reasons, including mutation and evolution, uncorrelated outliers may appear in between the otherwise similar subsequences. The compositions of the outliers, in terms of amino acid types, their number and their localization in the main sequence are practically unpredictable. Furthermore, similar subsequences may have, for a given position in the subsequence, distinct amino acids which are nevertheless highly similar from their chemical properties point of view. Consequently, we need a metric in order to compare two amino acids which quantify their degree of similarity from a physicochemical point of view. Various metric have been introduced for such a characterization. In the present work, we use the BLOSUM (BLOCKS of Amino Acid SUBstitution Matrix) [8] metric but another metric that characterized the similarity in terms of physicochemical properties may be substituted in the algorithm.

For the description of the amino acid sequences, we have introduced an approach based on two-points of correlation. For each amino acid of a given sequence, we compute the similarity, using a physicochemical metric, in between that particular amino acid and its two closest neighbours, or its closest neighbour if the amino acid is located at the end of the sequence. Then, from all the similarity measures, we compute their corresponding probability density function. The whole process is repeated for various distances in between amino acid pairs: i and $i \pm 2$, i and $i \pm 3$ up to i and $i \pm i_{\max} \forall i$.

The similarity probability density functions associated with the various distances in terms of relative position are then concatenated, to form a unique descriptor for the main amino acid sequence. In order to handle proteins constituted by more than one amino acid sequence, a descriptor is associated

with each sequence or chain. In the present implementation, a maximum of three chains may be accommodated by the descriptor.

The proposed descriptor does not depend on the number of residues since the probability density functions are by definition normalized. Also, it takes into account the local nature of the subsequences. These descriptors constitute the input of our multilayer neural network.

3 Molecular Surfaces Description

As stated earlier, the molecular surfaces are complex three-dimensional shapes characterized by an intricate geometrical structure. The molecular surface represents the part of the protein that is exposed and accessible by the surrounding. It is accessible to a solvent (such as water) and to other proteins; consequently it is the interface for interaction. Thus, the molecular surface is fundamental in understanding biological processes since the later are mostly based on protein-protein interaction. The molecular surface is computed from the position of the constituent atoms with a programme called a molecular solver. Various approaches are proposed in the literature; see, for instance, the work of [9].

The molecular surface is anything but a suitable output for a neural network. Indeed, the complexities of the surfaces vary immensely from one protein to the next as well as the number of vertices and triangles associated with the representation of the surface (assuming a triangular tessellation). Additionally, the molecular surface, as a whole, may have an arbitrary spatial orientation, or pose, in space. It is impossible to train a neural network for every possible spatial orientation of a given protein. Besides, it may even prove difficult to train the neural network for a discrete sample of orientations. Consequently, it is not possible to normalize the shape in a compact and efficient way in order to accommodate a neural network output.

For these reasons, we propose to characterize each protein by a shape descriptor (or index) which is invariant under translation and rotation (pose) [10]. The index has a normalized structure which makes it suitable for neural networks. The index is calculated as follow. Firstly, if required, the molecular surface is tessellated with triangular simplices. The barycentre and the symmetrical tensor of inertia of the molecular surface are calculated. Then, the Eigen decomposition of the tensor is evaluated. The Eigen vectors constitute a rotation invariant frame for the molecular surface: that is, the Eigen vectors are invariant under a rotation or a translation of the original molecular surface. The axes of the reference frame are identified by their corresponding Eigen value. That is, the first axis is the one with the highest Eigen value, the second axis is the one with the second highest Eigen values, etc. Note that the reference frame may become unstable under reflection of the axis if there is an Eigen value ambiguity, i.e. if two or more Eigen values are very similar, but such an ambiguity may be efficiently handled through a procedure described in [11].

Once the reference frame is determined, the probability density functions associated with of the radial distribution of the triangular simplices, the angular distribution of the triangular simplices relative to the Eigen vector corresponding to the largest Eigen value and the angular distribution of the triangular simplices relative to the Eigen vectors corresponding to the second largest Eigen value are evaluated. The distribution relative to the third axis is redundant: it may be obtained from the other two with the cosine law. The radial distribution is the distribution of the norm of the vectors starting at the barycentre of the molecular surface and ending at the barycentre of each triangular simplex. The angular distribution is the distribution of the angles in between the previous vectors and a given reference axis of the Eigen frame. The index is then formed by concatenating together the three distributions. Such an index is translation and rotation invariant which means that an alignment of the molecular surfaces is superfluous. Furthermore, its structure is normalized and does not depend on the molecular surface complexity or on its particular tessellation.

4 Molecular Surfaces Prediction

Recall that the aim of our neural network is to predict the shape index from an amino acid sequence descriptor, as introduced in Sections 2 and 3, respectively. In this paper, we restrict ourselves to the multilayer feed forward network. In our implementation, each neuron of each layer is fully connected to the neurons of the next layer and a sigmoid activation function is attributed to each neuron and a weight is associated with each connection. In the present paper, we consider neural networks with one and two hidden layers. The multilayer neural network has been selected because of its efficiency and its versatility.

The first layer has a fixed number of neurons which correspond to the various probability distribution functions associated with the correlation evaluated with the closest neighbourhood, the second closest neighbourhood, etc. as described in Section 2. In order to form the input layer, the probability distribution functions, which are by definition normalized, are simply concatenated with each other. The second and the third layers correspond to the hidden layers. We consider both neural networks with one and two hidden layer and study their relative efficiency in predicting the molecular surface. The last layer, which corresponds to the output layer, is associated with the shape index. As stated in Section 3, the shape index is formed of three probability density functions corresponding to the radial distribution and the two angular distributions of the surfaces elements in the pose invariant reference frame (the Eigen frame of the tensor of inertia).

The training of the neural network is paramount to determine the weights in between the neurons. Despite the fact that numerous algorithms have been introduced in order to compute the weights, the basic framework remains unaltered. That is, the weights are optimized in order to minimize the discrepancy in between the shape indexes

associated with the amino acid descriptors and the shape indexes predicted by the neural network. In our case, the discrepancy is defined as the square of the Euclidean distance in between the real shape index and the predicted shape index. The optimization process involves the minimisation of an objective function which is equal to the sum, over all proteins in the training set, of the square Euclidean distance in between the real and the predicted shape indexes. We consider various methods for the optimization namely three methods based on the gradient of the objective function: the gradient descent algorithm, the gradient descent algorithm with momentum and the Fletcher-Reeve conjugate gradient algorithm (FR). The reader is referred to [12] for more details about these algorithms. All these algorithms are based on the gradient of the objective function. The first algorithm relies

solely on the gradient, while the second algorithm aims to improve the performance by introducing a momentum or memory, while the later minimizes the objective function along mutually orthonormal directions.

We also consider two methods based on the Hessian of the objective function, namely the Levenberg-Marquardt (LM) and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods. The LM algorithm is a method in which the objective function is approximated or modelled (trust region) by a multidimensional Taylor development up to the second order (gradient and Hessian). The BFGS method, on the other hand, is a quasi-Newton method which approximates the Hessian of the objective function with a procedure based on the secant equation [12]. This suite of algorithms thus provides us with a variety of optimization methods.

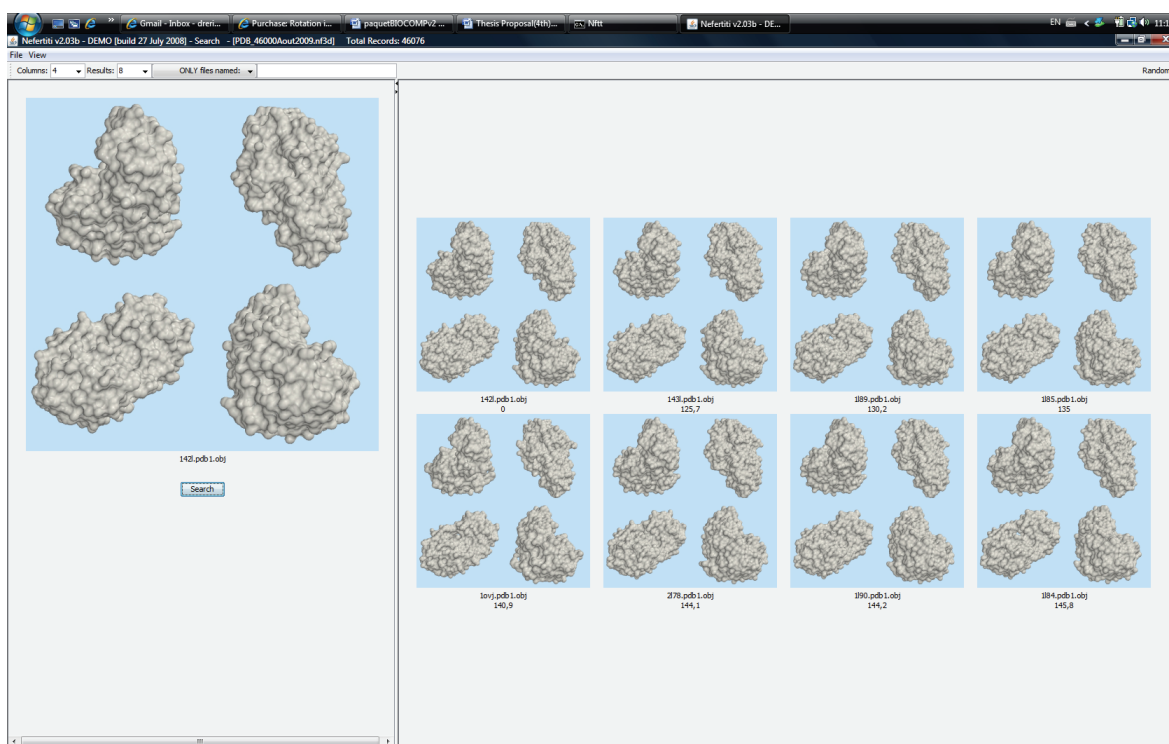


Figure 2. Molecular surface search engine. The unknown shape index obtained from the neural network is compared to the 45,000 shape indexes of the database and the most similar known molecular surfaces are retrieved.

The shape index, predicted by the neural network, provides an abstract description of the molecular surface associated with a given amino acid sequence. In its current version, our system does not allow for the direct reconstruction the molecular surface. Consequently, in order to retrieve the molecular surface, an additional step is required. Thus, instead of reconstructing the unknown molecular surface, we retrieve, from a database of known shape indexes, the known molecular surface which is the most similar to the unknown one. The unknown shape index is compared to the known shape indexes with the Euclidean metric.

In order to provide the most complete searching space, we have indexed, with our shape descriptors, the molecular surface of the proteins that may be found in the Protein Data Bank (PDB). A dedicated search engine, shown in Fig. 2, allows us to retrieve, for a given output of the neural network, the most similar known molecular surfaces out of the PDB. That is, we compare the proteins using a surface-based approach. In addition, the search engine provides us with an indication of the similarity between the known and the unknown structure. The retrieval time is very fast, i.e. the answer is returned in less than one second.

5 Experimental Results

Five multilayer neural networks are considered in our experiment. All networks have 200 neurons in their input layer and 120 neurons in their output layer. The networks differ by the structure of their hidden layer(s): the first, second, third and fourth neural network have one hidden layer with 18, 40, 55 and 70 neurons, respectively. The fifth neural network has two hidden layers composed of 18 and 19 neurons, respectively. As mentioned earlier, neural networks are trained with the gradient descent method, the gradient descent method with momentum, the Fletcher-Reeves conjugate gradient method, the BFGS method and the Levenberg-Marquadt method. The first three are based on the gradient of the objective function while the last two are based on an approximation of the Hessian of the objective function. The best results for training, in terms of mean squared error, are obtained with the gradient method with momentum when the neural network has one hidden layer and, with the Fletcher-Reeves conjugate gradient when the neural network has two hidden layers. The momentum, in the gradient method with momentum, reduces the training mean squared error by almost 50% compared to the plain gradient method.

The Fletcher-Reeves is by far the method with the fastest convergence (at least an order of magnitude) which most likely originates from the fact that the optimisation space is explored with mutually orthonormal directions. The BFGS method has by far the slowest convergence of all methods. The Levenberg-Marquadt performs poorly which means that the quadratic approximation of the objective function associated with this algorithm is not a priori justified.

In order to train our system and validate our results, we use the well-known 10-folds cross validation approach [13]. Here, we select randomly 90% of the members of each class in order to train the network and we validate the approach with the remaining 10%. The whole process is repeated ten times. The training set is constituted of 241 proteins divided in 19 families or classes: Complement control protein from Vaccinia virus, (Apo)ferritin from Mouse (*Mus musculus*), Immunoglobulin light chain kappa variable domain, VL-kappa from Mouse (*Mus musculus*), cluster 1.1, p53 tetramerization domain from Human (*Homo sapiens*), Penicillin-binding protein 5, C-terminal domain from *Escherichia coli*, Phosphoserine aminotransferase, PSAT from *Bacillus alcalophilus*, Chaperonin-10 (GroES) from *Escherichia coli*, Phage T4 lysozyme from Bacteriophage T4, Calmodulin from Cow (*Bos taurus*), Pertussis toxin S2/S3 subunits, C-terminal domain from *Bordetella pertussis*, Red fluorescent protein (fp583 or dsred(clontech)) from Coral (*Discosoma sp.*), Histidinol-phosphate aminotransferase HisC from *Escherichia coli*, Transcription initiation factor TFIIB, N-terminal domain from Human (*Homo sapiens*), P22 tailspike protein from *Salmonella* phage P22 and Catalase-peroxidase KatG from *Burkholderia pseudomallei*. Typical members of these families, identified by the PDB code, are 1g40, 1h96, 1rum, 1sak, 1z6f, 2bi3, 2c7c,

1421, 1xfw, 1bcp, 1ggx, 1gew, 1rly, 1tyv and 2b2s.

An unknown molecular surface is considered correctly predicted if both the known molecular surface retrieved from the unknown shape index output by the neural network and the unknown molecular surface belong to the same family. Among the families, eleven have a small number of proteins (2 to 8), two families have 28 and 32 members while the largest family is formed of 64 members. In general, the best results are obtained with the Fletcher-Reeves conjugate gradient method which means that, for that particular problem, the other approaches tend to be trapped in a local minimum. Increasing the number of neurons in the hidden layer tends to improve the results. The single hidden layer neural network outperforms the two hidden layers neural network if the number of neurons superior to forty. Following the 10-folds cross validation, the predictions of the system are 100% accurate for 8 families out of the 19 families of the training set (1ggx, 1rly, 1tyv, 1xfw, 1z6f, 2bi3, 2c7c, 1421), as well as 80% and more for another three families (1bcp, 1g40, 1qfg). The prediction accuracy for the low membership families is around 50% which may be explained by the fact that, due to random sampling, these families may not be always included in the training set during the 10-folds cross validation. One family with a relatively high membership (1fsn), 16, has relatively low prediction accuracy, 63 %. These results may be explained by the fact that, within that family, there is a relatively high variability in terms of amino acid sequence, while the variability is much lower from the molecular surface point of view. Consequently, it is more difficult to train the neural network as opposed to family in which similar amino acid sequences correspond to similar molecular surfaces.

6 Conclusions and Future Work

In this paper, we have shown that it is possible to predict the molecular surface of a protein from its amino acid sequence. We have introduced new descriptors for the amino acid sequence and the pose-invariant three-dimensional shape of the molecular surface. We have explored various approaches in order to optimize the objective function of the multilayer network. Finally, we have proposed an approach to retrieve, from an unknown shape index, the most similar known molecular surface out of a very large database of molecular surfaces. Our results against a database of 45,000 proteins, are promising, and showed that our molecular surface search engine allows one to retrieve the closest known molecular surface associated with an unknown, so-called query protein.

At this stage, our approach does not allow for the direct reconstruction of the molecular surface from the shape index. This is our main future research direction, and we are currently developing an algorithm to determine achieve this goal. We intend to replace the current shape index with one based on spherical harmonics [14] which, in addition to provide pose invariance, would allow for the direct

reconstruction of the molecular surface from the shape index by using the spherical harmonics decomposition.

7 References

- [1] P. C. Ng and S. Henikoff; "SIFT: predicting amino acid changes that affect protein function", *Nucl. Acids Res.*, 31 (13), 3812—3814, 2003.
- [2] A. Bairoch et al.; "The Universal Protein Resource (UniProt)", *Nucl. Acids Res.*, 33 (suppl 1), D154—D159, 2005.
- [3] S. Babaei, A. Geranmayeh and S. A. Seyyedsalehi; "Toward designing modular recurrent neural networks in learning protein secondary structure", *Expert Systems with Application*, 39, 6263—6274, 2012.
- [4] F. Bettella, D. Rasinski and E. W. Knapp; "Protein Secondary Structure Prediction with SPARROW", *J. Chem. Inf. Model.*, 52 (2), 545—556, 2012.
- [5] C. Branden, C. and J. Tooze; "Introduction to Protein Structure", Garland Publishing: New York, NY, 1999.
- [6] S. Ramachandran and N. V. Dokholyan,; "Homology Modeling: Generating Structural Models to Understand Protein Function and Mechanism", *Computational Modeling of Biological Systems, Biological and Medical Physics, Biomedical Engineering, Part 1*, 97—116, 2012.
- [7] H. M. Berman et al.; "The Protein Data Bank", *Nucl. Acids Res.* 28 (1), 235—242, 2000.
- [8] S. Henikoff and J. G. Henikoff; "Amino Acid Substitution Matrices from Protein Blocks", *PNAS*, 89 (22), 10915—10919, 1992.
- [9] W. Humphrey, A. Dalke and K. Schulten; "VMD - Visual Molecular Dynamics", *J. Molec. Graphics* 14, 33—38, 1996.
- [10] E. Paquet and H. L. Viktor; "Capri/MR: exploring protein databases from a structural and physicochemical point of view", *Proceedings of the VLDB Endowment (Very Large Database Endowment)*, 1 (1), 1504—1507, 2008.
- [11] R. Bro, E. Acar and T. G. Kolda; "Resolving the sign ambiguity in the singular value decomposition", *Journal of Chemometrics*, 22 (2), 135—140, 2008.
- [12] J. Nocedal and S. Wright; "Numerical Optimization", Springer, New York, NY, 2006.
- [13] R. Kohavi; "A study of cross-validation and bootstrap for accuracy estimation and model selection", *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2 (12), Morgan Kaufmann, San Mateo, CA, 1137—1143, 1995.
- [14] M. Kazhdan, T. Funkhouser and S. Rusinkiewicz; "Rotation invariant spherical harmonic representation of 3D shape descriptors", *SGP '03 Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 23-25 June, Aachen, Germany, 2003.

Fast Sequence Database Search Using Intra-Sequence Parallelized Smith-Waterman Algorithm

Chung-E Wang

Department of Computer Science
California State University, Sacramento
Sacramento, CA 95819-6021

wang@csus.edu

Abstract - *In this paper, we describe a fully pipelined VLSI architecture for sequence database search using Smith-Waterman algorithm. The architecture makes use of the principles of parallelism and pipelining to the greatest extent in order to take advantages of both intra-sequence and inter-sequence parallelization and to obtain high speed and throughput. First, we describe a parallel Smith-Waterman algorithm for general SIMD computers. The parallel algorithm has an execution time of $O(m+n)$, where m and n are lengths of the two biological sequences to be aligned. Next, we propose a VLSI implementation of the parallel algorithm. Finally, we incorporate a pipeline architecture in the proposed VLSI circuit and result in a pipeline processor that can do sequence database searches at the speed of $O(m+n+L)$, where L is the number of sequences in the database.*

Keywords: Sequence alignment, sequence database search, Smith-Waterman algorithm, parallel algorithm, VLSI circuit, pipelined architecture.

1. Introduction

A sequence alignment is a way of matching two biological sequences to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the two biological sequences. Sequence alignment is a fundamental operation of many bioinformatics applications such as genome assembly, sequence database search, multiple sequence alignment, and short read mapping.

The Smith-Waterman algorithm [1] is the most sensitive but slow algorithm for performing sequence alignment. Here sensitivity refers to the ability to find the optimal alignment. Smith-Water algorithm requires $O(m^2n)$ computational steps, where m and n are lengths of the two sequences to be aligned. Smith-Waterman algorithm was later improved by

Gotoh [2]. Gotoh's algorithm can find an optimal alignment of two biological sequences in $O(mn)$ computational steps. It was a great improvement for aligning two sequences. However, it's not fast enough for sequence database searches. A sequence database search is to compare a query sequence with a database of sequences, and identify database sequences that resemble the query sequence above a certain threshold.

Due to substantial improvements in multiprocessing systems and the rise of multi-core processors, parallel processing became a trend of accelerating Smith-Waterman's algorithm and sequence database searches. Many enhancements of Smith-Waterman algorithm based on the idea of parallel processing have been presented [3-18]. However, all these enhancements can only speed up Smith-Waterman algorithm by a constant factor. That is, all these enhancements still require an execution time of $O(mn)$ to align two biological sequences.

Besides parallel processing, heuristic methods are other commonly used approaches for speeding up sequence database searches. A heuristic method is a method which is able to produce an acceptable solution to a problem but for which there is no proof that it's an optimal solution. Heuristic methods are intended to gain computational performance, potentially at the cost of accuracy or precision. Popular alignment search tools such as FASTA [19], BLAST [20] and BLAT [21] are in this category. They did successfully gain some speed. However, the sensitivity is compromised. For sequence database searches, sensitivity refers to the ability to find all database sequences that resemble the query sequence above a threshold.

Without sacrificing any sensitivity, in this paper, we first describe a parallel Smith-Waterman algorithm for general SIMD (Single Instruction stream - Multiple Data stream) computers. This parallel algorithm requires an execution time of $O(m+n)$ to align two biological sequences. Second, we propose a VLSI (Very-Large-Scale Integration)

implementation of the parallel algorithm. Finally, we use the pipeline technique to overlap the execution times of alignment checking of database sequences. The resulting pipeline has a throughput rate of $O(1)$ execution time per database sequence. Consequently, the time complexity of the proposed pipeline processor is $O(m+n+L)$, where L is the number of sequences in the database.

2. The Smith-Waterman Algorithm

The Smith-Waterman algorithm is used to compute the optimal local-alignment score. Let $A = a_1 a_2 \dots a_m$ and $B = b_1 b_2 \dots b_n$ be the two sequences to be aligned. A weight $w(a_i, b_j)$ is defined for every pair of residues a_i and b_j . Usually $w(a_i, b_j) \leq 0$ if $a_i \neq b_j$, and $w(a_i, b_j) > 0$ if $a_i = b_j$. The penalties for starting a gap and continuing a gap are defined as g_{init} and g_{ext} respectively. The optimal local alignment score S can be computed by the following recursive relations:

$$E_{i,j} = \max \{ E_{i,j-1} - g_{ext}, H_{i,j-1} - g_{init} \} \quad (1)$$

$$F_{i,j} = \max \{ F_{i-1,j} - g_{ext}, H_{i-1,j} - g_{init} \} \quad (2)$$

$$H_{i,j} = \max \{ 0, E_{i,j}, F_{i,j}, H_{i-1,j-1} + w(a_i, b_j) \} \quad (3)$$

$$S = \max \{ H_{i,j} \}; \quad (4)$$

The values of $E_{i,j}$, $F_{i,j}$ and $H_{i,j}$ are 0 when $i < 1$ and $j < 1$.

Smith-Waterman algorithm is a dynamic programming algorithm. A dynamic programming algorithm is an algorithm that stores the results of certain calculations in a data structure, which are later used in subsequent calculations. Smith-Waterman algorithm uses a $m \times n$ matrix, called alignment matrix, to store and compute H , E , and F values column by column.

3. An Intra-sequence Parallelized Smith-Waterman Algorithm For SIMD Computers

Intra-sequence parallelization means the parallelization is within a single pair of sequences, in contrast to inter-sequence parallelization where the parallelization is carried out across multiple pairs of sequences.

Figure 1 shows the computational dependencies of Smith-Waterman alignment matrix. Most of the existing parallelized Smith-Waterman algorithms in the literature are originated from this dependency structure. Since cells on a bottom-left to top-right diagonal have the same sum of indices, we number those diagonals with their sums of indices. Noticeably cells of a diagonal don't depend on

cells of the same diagonal and thus can be computed simultaneously. Furthermore, cells of a diagonal only depend on cells of the previous two diagonals. So, the basic idea of the parallel algorithm is to fill the matrix diagonal by diagonal starting from the top-left corner. Moreover cells of a diagonal are filled simultaneously with multiple processors.

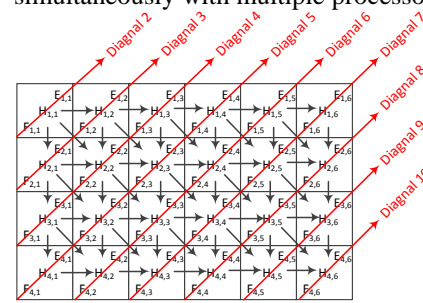


Fig. 1. Computational dependencies in the Smith-Waterman alignment matrix.

The pseudo code of the parallel algorithm is given in Figure 2. Note that in the pseudo code, we use the in-parallel statement to indicate things to be done by processors simultaneously. The in-parallel statement has the following syntax:

```
In parallel, all processor i, lo<=i<=hi do {
    ...
}
```

Only processors with processor numbers between lo and hi are activated. Also note that in the pseudo code, j_i is a variable for processor i only. In other words, different processors have different j 's. Furthermore, array $maxH$ is used for processors to keep track of the largest $H_{i,j}$. Since there are $m+n-1$ diagonals, the loop repeats $m+n-1$ times and thus the parallel algorithm has an execution time of $O(m+n)$.

```
In parallel, all processor i, 1<= i <=m do
    E[i][0] = F[i][0] = H[i][0] = maxH[i] = 0;
In parallel, all processor i, 1<= i <=n do
    E[0][i] = F[0][i] = H[0][i] = 0;
for (diag=2; diag<=m+n; ++diag) {
    if (diag<=m+1) first = diag-1;
    else first = m;
    if (diag-n<=1) last = 1;
    else last = diag-n;
    In parallel, all processor i, first<= i <= last do {
        j_i = diag - i;
        E[i][j_i] = max { E[i][j_i-1] - g_ext, H[i][j_i-1] - g_init };
        F[i][j_i] = max { F[i-1][j_i] - g_ext, H[i-1][j_i] - g_init };
        H[i][j_i] = max { 0, E[i][j_i], F[i][j_i], H[i-1][j_i-1] + w(a_i, b_j_i) };
        if (H[i][j_i] > maxH[i]) maxH[i] = H[i][j_i];
    }
}
S = max { maxH[1], maxH[2], ..., max H[m]};
```

Fig. 2. The pseudo code

4. A VLSI Implementation

First we design a small processing element according to the recursive relations (1), (2) and (3). Processing elements will be used to implement cells of Smith-Waterman's alignment matrix. Thus, each processing element will be numbered with the indices of its corresponding cell in Smith-Waterman's alignment matrix. As follows, processing element $PE_{i,j}$ will be used to compute values $E_{i,j}$, $F_{i,j}$ and $H_{i,j}$. As shown in Figure 3, processing element $PE_{i,j}$ consists of 3 registers $E_{i,j}$, $F_{i,j}$ and $H_{i,j}$ and several simple combinational circuits such as adders and comparators (for finding maximum of two or three values). Since $H_{i,j}$ depends on $E_{i,j}$ and $F_{i,j}$, processing element $PE_{i,j}$ has two computational states and thus requires two clock signals to complete its computation of values $E_{i,j}$, $F_{i,j}$ and $H_{i,j}$. Since the longest data path in the processing element consists of an adder and a comparator, the clock period, i.e. time between each clock signal, only needs to be set to the time delay caused by an adder and a comparator.

Each processing element $PE_{i,j}$ has 5 input ports and 3 output ports. Input ports A and B are for inputting residues a_i and b_j from the two sequences to be aligned. Input ports E_{in} , F_{in} and H_{in} are for inputting corresponding values from cells on which $PE_{i,j}$ depends. Output ports E_{out} , F_{out} and H_{out} are for exporting values to cells depending on $PE_{i,j}$.

Additionally, in order to keep track of the largest $h_{i,j}$ value, register $H_{i,j}$ of processing element $PE_{i,j}$ is connected to register $maxH_i$ as shown in Figures 3.

Next, we map our algorithm into a VLSI circuit. Figure 4 depicts our VLSI circuit at the register level for $m=4$ and $n=6$. Processing elements and registers $maxH_i$ are connected together according to recursive relations (1), (2), (3) and (4). As shown in Figure 4, processing elements are arranged into levels such that processing elements corresponding to cells of a diagonal are placed on the same level and thus will compute their values at the same time. For the clarity of the logic of the VLSI circuit, in Figure 4, levels are numbered with their corresponding diagonal numbers.

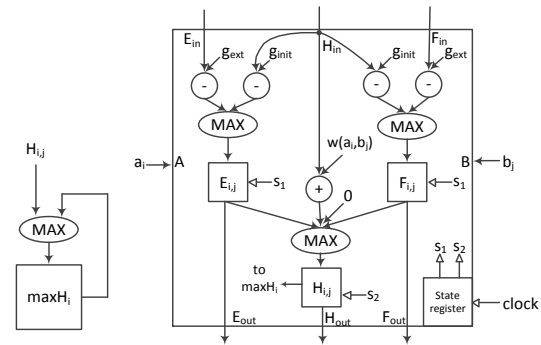


Fig. 3. Processing element $PE_{i,j}$ and register $maxH_i$

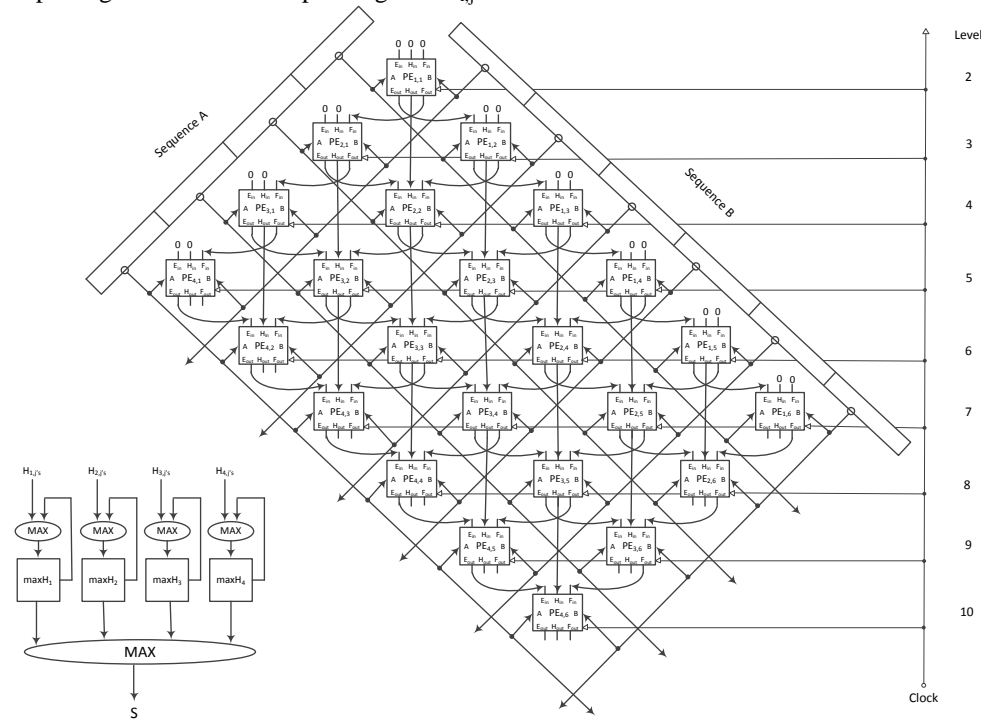


Fig. 4. The VLSI circuit, $m=4$, $n=6$.

5. A Pipeline Architecture for Sequence Database Searches

Since sequence database search is different from sequence alignment, first, we modify our processing element $PE_{i,j}$ as shown in Figure 5. Instead of keeping track of largest $h_{i,j}$, the new processing element $PE_{i,j}$ will generate a SELECT signal when its $H_{i,j}$ value is above a threshold.

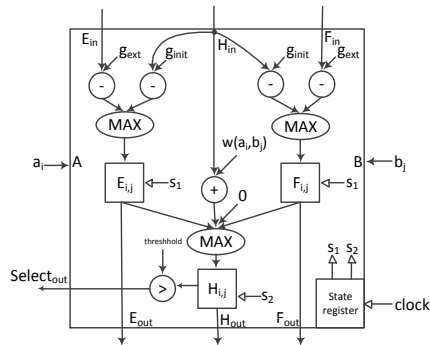


Fig. 5. Modified processing element $PE_{i,j}$.

To speedup sequence database searches, a pipelined architecture is incorporated in the VLSI circuit. Pipelining is a natural concept for increasing the throughput of a system when processing a stream of

data, even though pipelining cannot speed up the process of a single datum. The space-time diagram in Figure 6 reveals the advantages of the pipelined architecture in processing database sequences. The diagram shows the succession of the levels in the pipeline with respect to time. From the diagram one can observe how independent sequences are processed concurrently in the pipeline.

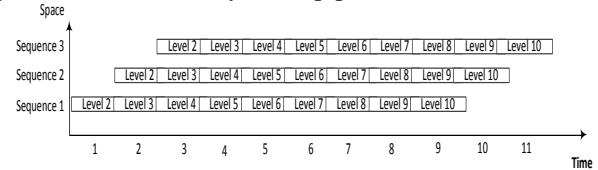


Fig. 6. Pipeline space-time diagram

Since there are $m+n-1$ levels in the VLSI circuit, it's very natural to organize the entire architecture as a linear pipeline with $m+n-1$ stages. To do so, as shown in Figure 7, we add $m+n-1$ registers to the VLSI circuit to hold database sequences one for each stage, i.e. level. Additionally, each database sequence register has a S flag which will be used to indicate whether the database sequence is selected or not. As shown in Figure 7, processing elements' select signals are connected to S flags of corresponding

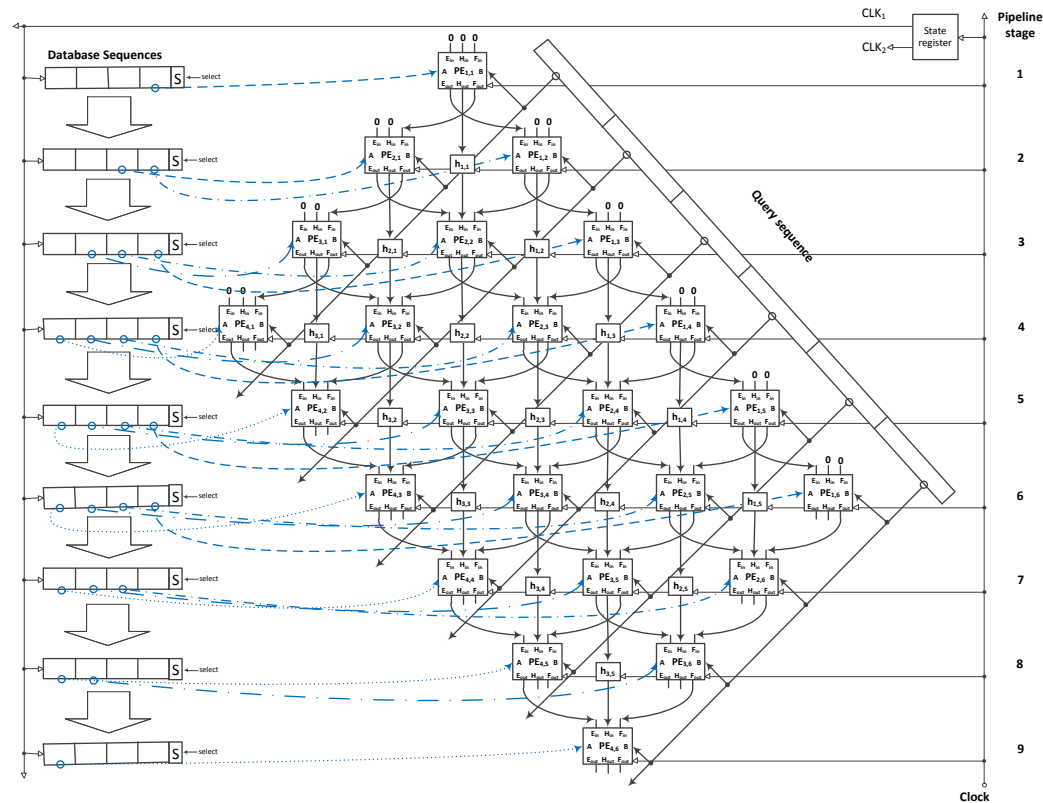


Fig. 7. The single-chip pipeline processor, $m=4, n=6$.

database sequence registers. Furthermore, according to recursive relation (3), $H_{i,j}$ depends on $H_{i-1,j-1}$ which is not in the immediate previous level of $H_{i,j}$. So, for the purposes of buffering and synchronization, we add $h_{i,j}$ buffer registers to hold $H_{i,j}$ values. Since a processing element $PE_{i,j}$ needs two clock signals to compute its values, our pipeline takes 2 clock signals to move from one pipeline stage to another.

Apparently as soon as the first database sequence completes its alignment checking, every one stage time there is a database sequence completes its alignment checking. Consequently, the pipeline processor has a throughput rate of one database sequence per two clock signals. In other words, the pipeline has a time complexity of $O(1)$ time per database sequence. Moreover, since the first sequence takes $O(m+n)$ time to completes its alignment checking, the total time complexity of the pipeline processor is $O(m+n+L)$, where L is the number of sequences in the database.

For most bioinformatics applications, m and n are in thousands and thus $m \times n$ is in millions. Since there are $m \times n$ processing elements in our VLSI circuit, the pipeline processor requires millions of combinational circuits and registers. As of today, a VLSI microchip can have billions of transistors. Since a register or a simple combinational circuit such as adder or comparator doesn't need thousands of transistors, billions of transistors should be enough for millions of our processing elements. As a result, it's possible to implement the pipeline processor in a single VLSI microchip.

For some applications such as genome assembly, the length of the query sequence may be more than thousands and thus requires multiple microchips to implement the pipeline processor. Figure 8 demonstrates the scalability of the pipeline processor. In circuit design, scalability refers to the ability to be expanded to cope with increased use. As shown in Figure 8, we decompose the pipeline circuit into three sub-circuits each of which can be implemented in a microchip. Type 1 chip is for the head of a pipeline. Type 3 chip is for the tail of a pipeline. Type 2 chip is for the middle part of a pipeline. To handle long query sequences, we simply use more type 2 chips in the middle to lengthen the pipeline.

6. Conclusion

In this paper, we have described a $O(m+n)$ time intra-sequence parallelized Smith-Water algorithm for general SIMD computers, where m and n are lengths of the two sequences to be aligned. We have shown a VLSI implementation of the parallel algorithm. We have also shown that by incorporating

a pipelined architecture into the VLSI circuit, we can speed-up sequence database searches without sacrificing the sensitivity. The resulting pipeline processor can do sequence database searches at the speed of $O(m+n+L)$, where L is the number of sequences in the database. Moreover, we have demonstrated the scalability of the pipeline processor.

7. References

- [1] TF Smith, MS Waterman, Identification of common molecular subsequences. *J Mol Biol.* 147, 195–197 (1981).
- [2] O. Gotoh, An improved algorithm for matching biological sequences. *J Mol Biol.* 162, 705–708 (1982).
- [3] M Borah, RS Bajwa, S Hannenhalli, MJ Irwin, A SIMD solution to the sequence comparison problem on the MGAP.
- [4] B Alpern, L Carter, KS Gatlin, Microparallelism and high performance protein matching. Proceedings of the 1995 ACM/IEEE Supercomputing Conference, San Diego, California, Dec 3-8, 1995.
- [5] R Hughey, Parallel hardware for sequence comparison and alignment. *Comp Appl Biosci.* 1996;12:473–479.
- [6] A Wozniak, Using video-oriented instructions to speed up sequence comparison. *Comput Appl Biosci.* 13, 145–150 (1997)
- [7] T Rognes, E Seeberg, Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics.* 16, 699–706 (2000).
- [8] ITS Li, W Shum, K Truong, 160-fold acceleration of the Smith-Waterman algorithm using a field programmable gate array (FPGA). *BMC Bioinformatics.* 8, 185 (2007).
- [9] M Farrar, Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics.* 23, 156–161 (2007).
- [10] Z. Nawaz, M. Shabbir, Z. Al-Ars, and K. Bertels, "Acceleration of smith-waterman using recursive variable expansion," in *DSD-2008*, pp. 915–922, September 2008.
- [11] A Szalkowski, C Ledergerber, P Krähenbühl, C Dessimoz, SWPS3 - fast multithreaded vectorized Smith-Waterman for IBM Cell/B.E. and x86/SSE2. *BMC Res Notes.* 1, 107 (2008).
- [12] A Wirawan, CK Kwok, NT Hieu, B Schmidt, CBESW: Sequence Alignment on the Playstation 3. *BMC Bioinformatics.* 9, 377 (2008).
- [13] W Rudnicki, A Jankowski, A Modzelewski, A Piotrowski, A Zadrożny, The new SIMD Implementation of the Smith-Waterman Algorithm on Cell Microprocessor. *Fund Infor.* 96, 181–194 (2009)
- [14] Y Liu, DL Maskell, B Schmidt, CUDASW++:

- optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units. *BMC Res Notes*. 2, 73 (2009).
- [15] Ł Ligowski, WR Rudnicki, An efficient implementation of Smith Waterman algorithm on GPU using CUDA, for massively parallel scanning of sequence databases. Eighth IEEE International. Workshop on High Performance Computational Biology, Rome, Italy, 2009.
- [16] Y Liu, B Schmidt, DL Maskell, CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA_enabled GPUs based on SIMT and virtualized SIMD abstractions. *BMC Res Notes*. 3, 93 (2010).
- [17] Ł Ligowski, WR Rudnicki, Y Liu, B Schmidt, Accurate Scanning of Sequence Databases with the Smith-Waterman Alg. *GPU Comput. Gems, Emerald Edition*. (Morgan Kaufmann, 2011), pp. 155–157.
- [18] R Torbjorn, Faster Smith-Waterman database searches with inter-sequence SIMD parallelization. *BMC Bioinformatics* 12:221 2011.
- [19] WR. Pearson, DJ. Lipman, Improved tools for biological sequence comparison. *PNAS*, 85 (8): 2444-8. (1988)
- [20] SF. Altschul, W. Gish, W. Miller, EW. Myers, DJ. Lipman, "Basic local alignment search tool". *J Mol Biol* 215 (3): 403–410. (October 1990)
- [21] WJ. Kent, "BLAT--the BLAST-like alignment tool". *Genome research* 12 (4): 656–664. (2002)

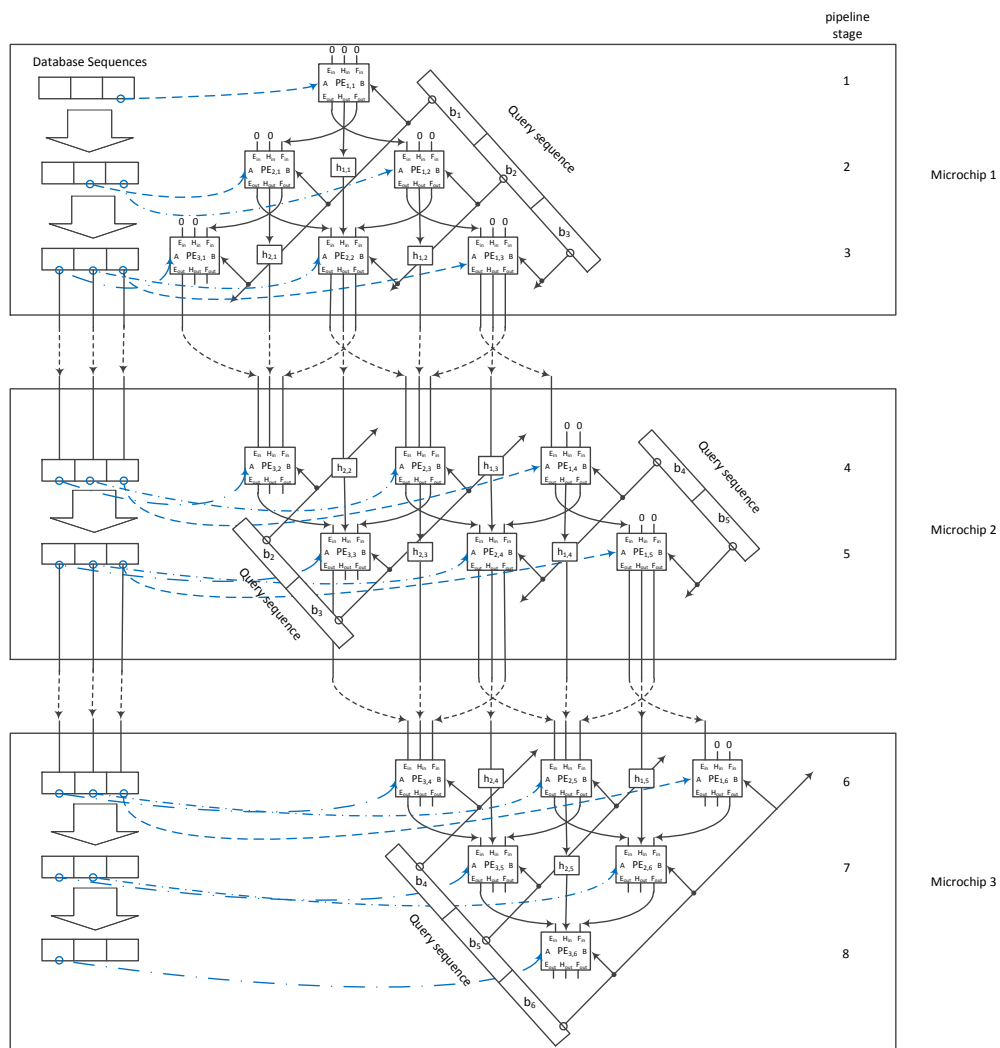


Fig. 8. A multiple-chip pipeline processor, $m=3$, $n=6$.

Bimodal Gene Prediction via Gap Maximisation

Abdullatif S. Al-Watban^{1,2}, Zheng Rong Yang¹

¹School of Biosciences, University of Exeter, EX4 4QD, UK

²Saudi Food and Drug Authority, Medical Devices Sector, Riyadh, KSA

Abstract- *Bimodal gene is one of the common phenomena frequently observed in gene expression data for certain types of studies including cancer studies and drug/therapy effect studies. There have been several algorithms proposed to predict bimodal genes with success. However, occasionally their performance is not very satisfied. We propose a new algorithm to detect bimodal genes. The new algorithm is based on the assumption that the bimodality is related with the gap between two consecutive expressions. We show that this new algorithm demonstrates better performance compared with several benchmark algorithms using both real and simulated data sets.*

Keywords: bimodal distribution, non-parametric analysis, differential genes, heterogeneity.

1 Introduction

Microarray experiments have benefitted the discovery of genetic differentiation pattern for interpreting the observed phenotypic differentiation for a decade [1]. The success is due to high-throughput and genome-wide examination. The discovery of differential genes in relation to phenotypic differentiation can be implemented using standard student t test if data satisfy the assumption. However biological diversity makes this difficult because a large number of genes appear to have bimodal or multi-modal distribution [2]. Fig 1 shows such a typical bimodal distribution of samples in the same category (such as cancer samples) of a gene.

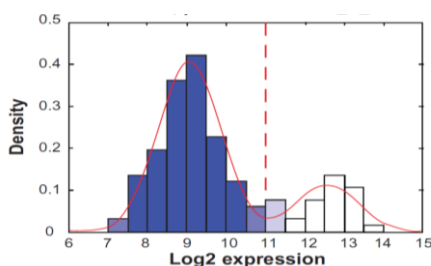


Figure 1: Histograms for ERBB2 gene. The gene has bimodal distribution with the dashed vertical line representing the classification threshold between the two modes [3].

Khalil et al have explained that cancer is a complex disease [4] because it has many subtypes. The existence of bimodal genes may be related to important subtypes of a disease. In medical science, bimodal genes can be the product of somatic mutations as the amplification of the receptor tyrosine kinase proto-oncogene "erbB2" during the development of cancer [5]. Another cause for the bimodality in cancers is germ cell mutations such as SNPs [6]. It has been noticed

that the majority of cancer data demonstrate this kind of heterogeneous pattern [7-9]. Genetic translocations are commonly occurred in cancer cell which is a result of the rearrangement of parts between non-homologous chromosomes [10]. However, these mutations play main role in cancer cell progression or, more generally, diseases development. Furthermore, the genomic lesions may affect some samples but not all leading to the occurrence of bimodality. An example of recurrent fusion was observed by Tomlins and others in prostate cancer datasets where they found ERG and ETV1 genes over expressed in some of the samples in multiple datasets [9]. A study has showed that oncogene HER2 is over-expressed in 15–20% of breast tumors compared with normal breast tissues [11]. In addition the bimodality appears in biological systems as noticed by Mason and his group [12]. It is observed that the expression levels for some genes showed a distinct bimodal distribution in human skeletal muscle tissue. Also bimodal distribution were studied in blood glucose samples [13, 14]. The bimodality can occur in humongous tissue as reported in these references [12, 15].

This heterogeneity demonstrated that the fully understanding to both genotype and phenotypes is the critical key for drug design [8]. The researchers have made a great effort to study the complexity of cancer disease aiming to understand the molecular characteristics [16, 17]. Cancer patients with similar tumour characteristics are likely not to response for the same treatment [18]. In breast cancer, for example, variant responses were found to drug such as Tamoxifen and Herceptin giving evidence of the heterogeneity in pathological factors such as estrogen receptor (ER) and HER2 status [19]. Large number of patients gained from using Tamoxifen for hormone receptor-positive but the same drug failed in subgroup of patients who carry specific variants in the cytochrome gene P450 2D6 (CYP2D6) [20, 21]. Trastuzumab, as a first drug approved by FDA for this purpose, has been a beneficial therapy, either alone or in combination with chemotherapy, in about 25% of patients with positive HRE2 cancer patients [22-25]. This raised an issue of an accurate grouping of HRE2-positive patients [21]. Gefitinib (Iressa) has been approved by FDA, which suppress the ATP binding function of EGFR, and has been of partially remission regression for 10-30% of patients with non-small cell lung cancer [26-30]. It has been noticed, that genetic alterations are associated with drug response as proven in their study [31].

Due to the often observed heterogeneity in gene expression data, the conventional t test and correlation analysis may not be able to well detect partial differentiation. The kurtosis analysis [32], the likelihood ratio test [33] and the bimodality index [34] have been proposed to examine the bimodality

among genes. PACK (Profile Analysis using Clustering and Kurtosis) [32] clusters samples first and then uses kurtosis to find relevant classifiers. It was reported that about 80%-20% bimodal genes were missed using PACK [34]. The likelihood ratio test (LRT) [33, 35] examines the likelihood of bimodal over unimodal [13, 14]. Ertel and Tozeren used the χ^2 test with six degree of freedoms. They set 0.001 as the significance level to predict bimodal genes. Bessarabova and colleagues developed a τ indicator for detecting bimodality [36]. They combined a statistical method based around t test like statistic for direct comparison of gene expression from different platforms to identify bimodal genes based on the relative difference average between each peak of gene expression value in breast cancer. The Bimodality Index [34] used a mixture of two homogeneous Gaussians to model bimodality and outweighed the high-expressed samples.

Have applied these algorithms to our data, we have found that they often show dissatisfied performance. Some often over-predict bimodal genes and some do not provide a statistical significance value for analysis. In this paper we present a novel algorithm further. The basic principle is to detect the maximum gaps between two clusters. This therefore avoids the parametric function to be used. We have evaluated this algorithm in comparison with several benchmark algorithms and demonstrate in this paper that this new algorithm provides another way to acquire insightful interpretation to bimodality among genes.

In the following sections, we discuss the implementation of hBI and evaluate its performance in comparison to some benchmark algorithms using real and simulated data.

2 Methods

Our algorithm is a revision of Bimodal Index - BI [34], which is defined as:

$$BI_i = \sqrt{\pi_i(1-\pi_i)}\delta_i \quad (1)$$

where π_i is the proportion of samples and δ_i is the distance between the two subgroups of the i^{th} genes. The use of this definition implies a homogeneous variance for two clusters of samples. A one-side t statistics of the i^{th} gene can be defined as

$$t_i = \frac{\mu_{H,i} - \mu_{L,i}}{\sqrt{\frac{\sigma_{L,i}^2}{n_{L,i}} + \frac{\sigma_{H,i}^2}{n_{H,i}}}} \quad (2)$$

where $\sigma_{L,i}^2$ is the variance of lowly expressed samples, $\sigma_{H,i}^2$ is the variance of highly expressed samples, $n_{L,i}$ is the number of lowly expressed samples, $n_{H,i}$ is the number of highly

expressed samples, $\mu_{L,i}$ is the mean of lowly expressed samples, and $\mu_{H,i}$ is the mean of highly expressed samples of the i^{th} gene. if $\sigma_{L,i}^2 = \sigma_{H,i}^2 = \sigma_i^2$, this one side t statistic becomes

$$t_i = \sqrt{\frac{n}{\sigma_i^2}}\delta_i\sqrt{\pi_{H,i}(1-\pi_{H,i})} \quad (3)$$

where $\pi_{H,i}$ is the proportion of highly expressed samples of the i^{th} gene. If the sample size is fixed for all genes,

$$t_i \propto \sigma_i^{-1}\delta_i\sqrt{\pi_{H,i}(1-\pi_{H,i})} \quad (4)$$

It can be seen that if homogeneous exists across subgroups and genes, BI is equivalent to one side t statistic. However this can hardly be true in real applications. We therefore revise BI employing heterogeneous variance. In the one side t statistic, we use percentile estimations to replace parametric estimation of means and variances shown below

$$t_i = \frac{q_{H,i}^{25} - q_{L,i}^{75}}{\sqrt{\frac{\sigma_{L,i}^2}{n_{L,i}} + \frac{\sigma_{H,i}^2}{n_{H,i}}}} \quad (5)$$

Here q_H^{25} is the 25th percentile of highly expressed samples, q_L^{75} is the 75th percentile of lowly expressed samples and the variances are calculated using

$$\sigma = \frac{IQR}{1.34896} \quad (6)$$

We assume that the separation between lowly expressed samples and highly expressed samples occurs at one of the largest gaps between consecutive sorted samples. Therefore we introduce the gap between lowly expressed samples and highly expressed samples to enhance the bimodality test. Our heterogeneous bimodal index (hBI) is defined below

$$hBI_i = \alpha(m_{H,i} - M_{L,i}) + (1-\alpha)t_i \quad (7)$$

where $m_{H,i}$ is the minimum of highly expressed samples, $M_{L,i}$ is the maximum of lowly expressed samples of the i^{th} gene and $\alpha > 0$ is a trade-off between the gap effect and t statistic. In this paper, $\alpha = 0.75$.

BI uses an arbitrary threshold to make decision based of the indexes, we employ the sequential Monte Carlo approach [37] (Besag and Clifford 1996) to deliver significance analysis. The procedure of our algorithm is shown below

Step 1. BI calculation for each gene

1.1. to sort expressions

- 1.2. to calculate the distance between every consecutive expressions and record them as a gap list
- 1.3. to sort the gap list
- 1.4. to calculate the revised BI for the top ten gaps and record them in a bimodality list
- 1.5. to maximise the bimodality list

Step 2. Apply BC algorithm to obtain p values

To evaluate our algorithm in comparison with likelihood test, Kurtosis test and BI test, we calculate sensitivity (Sen), specificity (Spe), total accuracy (Auc) and use receiver operative characteristic (ROC) [38] analysis. The sensitivity is the ratio of correctly predicted bimodal genes. The specificity is the ratio of correctly predicted non-bimodal genes. The total accuracy is the ratio of corrected identified unimodal and bimodal genes. Specially, we calculate the area under ROC curve (AUC) for comparison.

3 Results and discussions

3.1 Simulated Data

For all five scenarios, 950 genes were designed as unimodal and 50 genes were designed as bimodal. Each gene has 40 replicates. Thirty replicates were designed of low expressions. Ten replicates were designed of high expressions. Each simulation was repeated for ten times.

Scenario 1 - Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed samples of a bimodal gene were drawn from a normal distribution of mean ten and standard deviation one. Highly expressed samples of a bimodal gene were drawn from a normal distribution of mean 12 with variable standard deviation drawn from a uniform distribution between one and five. *Table 1* shows the comparison based on the mean values among ten simulations for four algorithms using specificity, sensitivity and AUC. It can be seen that hBI and Kurtosis have similar performance and hBI slightly outperforms Kurtosis analysis. Likelihood test shows the worst performance with the sensitivity as 0.06 although its specificity is 1.

Table 1: The averaged measurements for scenario 1

	LR	K	BI	hBI
Spe	1	0.983	0.975	0.992
Sen	0.062	0.858	0.532	0.84
Auc	0.995	0.964	0.852	0.992

Scenario 2 - Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed samples of a bimodal gene were drawn from a normal distribution of mean ten and standard deviation

one. Highly expressed replicates of a bimodal gene follow a uniform distribution in the interval between zero and five in addition to maximum of low expressions. The averaged measurements are shown in *Table 2*. In this scenario kurtosis has shown the worst accuracy (36%) while the other relatively similar and higher, 99.9%. Also the result has shown that the likelihood test has very low sensitivity while BI and hBI perform equally well.

Table 2: The averaged measurements for scenario 2

	LR	K	BI	hBI
Spe	1	0.986	0.997	0.9963
Sen	0.058	0	0.954	0.924
Auc	0.999	0.36	0.999	0.9985

Scenario 3 - Samples of unimodal genes were drawn from a uniform distribution in the interval between ten and 12. Lowly expressed replicates of a bimodal gene were drawn from the same low expression distribution as bimodal genes and highly expressed replicates of a bimodal gene were drawn from a normal distribution with two units added to the maximum of the low expressions. *Table 3* shows the summary of the simulations for this scenario. This scenario has shown that Kurtosis failed again to have a sensible accuracy (14%). hBI shows the highest AUC (0.999) similar to LR (0.996) and BI (0.993). hBI outweighs LR and BI in term of sensitivity, the sensitivities of BI and LR are 0.81 and 0.77, respectively while hBI's sensitivity is 0.94.

Table 3: The averaged measurements for scenario 3

	LR	K	BI	hBI
Spe	0.997	0.9519	0.9898	0.996
Sen	0.774	0	0.812	0.94
Auc	0.996	0.1413	0.9933	0.999

Scenario 4 - Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed replicates of a bimodal gene were drawn from a mixture of a normal distribution of mean ten and a normal distribution of mean 12. The standard deviation of the former was designed as one and that of the latter was designed as three. Highly expressed replicates of a bimodal gene were drawn from the low expressions plus white noise with two units above the maximum low expression. *Table 4* shows the summary of ten simulations on random samples for this scenario. All perform very well in terms of AUC. This means there are some suitable statistical significance levels by which perfect separation between unimodal and bimodal genes can be found.

Table 4 The averaged measurements for scenario 4

	LR	K	BI	hBI
Spe	1	0.9861	0.9963	0.997
Sen	0.468	0	0.93	0.952
Auc	0.998	0.9572	0.9984	0.9988

Scenario 5 - Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. We organised lowly expressed replicates of a bimodal gene as a mixture of three normal distributions with mean values as ten, 11 and 12 as well as standard deviation values as three, two and one. Highly expressed replicates of a bimodal gene were drawn in the same way as scenario 4. Based on ten random simulations for this scenario, we have observed that although LR and BI show reasonably good values of AUC, their sensitivities are not acceptable. This shows that these two algorithms have the same problem encountered in scenario 4 that their p values tend to be large, which leads to the difficulty of using command significance levels to make decision. Kurtosis analysis does not work well because its AUC value drops to 0.66 not very far away from 0.5, a random classification. In this scenario hBI perform the best in all measurements while BI has 69% sensitivity.

Table 5: The averaged measurements for scenario 5

	LR	K	BI	hBI
Spe	1	0.9844	0.9845	0.9873
Sen	0	0	0.698	0.766
Auc	0.9267	0.6676	0.9593	0.9867

3.2 Real data

GSE11121 dataset: The data set was downloaded from GEO (Gene Expression Omnibus). It contains 200 lymph node-negative breast cancer patients who were not treated by systemic therapy after surgery. The data was derivation study to find prognostic motifs [39]. Gene expression profiling of patients was done using the Affymetrix HG-U133A microarray platform comprising 22283 probs. The raw expression deposited at the NCBI GEO data repository under the accession number GSE11121. We have transformed the expression using base two logarithm before analysis. We used three significance levels (0.001, 0.01 and 0.05) to predict bimodal genes. *Table 6* shows the predicted bimodal genes using these three significance levels. The likelihood test predicted from 0.3% to 2.3% bimodal genes, BI predicted from 0.01% to 5% bimodal genes and hBI predicted bimodal genes from 0.01% to 5% as well. However Kurtosis analysis ends up with too many predictions up to 54.7%, which is unreasonable. Even for the significance level 0.001, it still predicts 36.3% bimodal genes, which is far more than a realistic level.

Table 6: Number of predicted bimodal genes for three significance levels for data set GDS11121

Significance levels	LHR	K	BI	hBI
0.001	72	8087	22	23
0.01	182	10065	227	221
0.05	523	12193	1112	1112

Fig 2 (a) shows the overlap analysis between four algorithms based on the significance level 0.001 values using VennDiagram [40]. We have found that hBI is most similar to BI. The overlap percentage between these two algorithms is 31.8%, i.e. $100 * 7 / (7 + 14 + 1)$. The overlap degree between LHR and hBI is 20.2%. The overlap degree between LHR and BI is 5.6%. 91.3% of predicted bimodal genes of hBI are predicted by Kurtosis as well. This percentage drops to 69.6% between hBI and LHR as well as 30% between hBI and BI. Also the overlap percentage between BI and hBI is 27.7% for significance level 0.01 and the overlap degree is 34.3% between the hBI and LHR and 7.9% between BI and LHR - *Fig 2* (b). 90.9% of predicted bimodal genes of hBI is predicted by Kurtosis as well. This percentage drops to 46.6% between hBI and LHR as well as 24% between hBI and BI. For significance level 0.05, we found the overlap between hBI and BI is 36.2% and the overlap degree between hBI and LHR is 28.2% and 7.07% between BI and LHR - *Fig 2* (c). In addition, 83.3% (32.3%, 36.2%) of hBI's predictions are consistent with Kurtosis (LHR, BI).

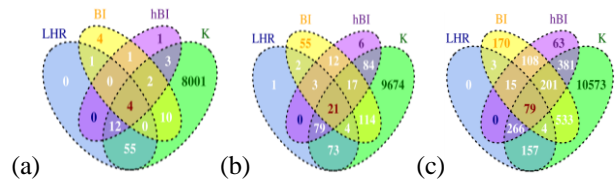
**Fig 2:** Venn diagram illustrates the overlapped between the methods for GSE11121 with the significance levels 0.001(a), 0.01(b) and 0.05(c).

Fig 3 shows top five bimodal genes predicted based on the significance level 0.001, where (a-d) predicted by all and (e) was predicted by hBI only. It can be seen that they show different types of distributions. Both GOXA1 - *Fig 3* (a) - and GATA3 - *Fig 3* (b) show a pattern that the high expressions form a tight cluster. However the low expressions demonstrate a more flat distribution or form more small clusters. TDRD12 - *Fig 3* (c) - and GRIA2 - *Fig 3* (d) have tight clusters formed by low expressions and their high expressions display flat distributions. SH3GL3 shows a different pattern from other four. It is composed of two more tightly formed clusters, one small and one large. The gap between two clusters is large. The analysis of these patterns proves one important concept that the use of restrict assumption of data distribution may not be sufficient for accurate prediction of bimodal genes in real applications, where distribution can vary in many different formats.

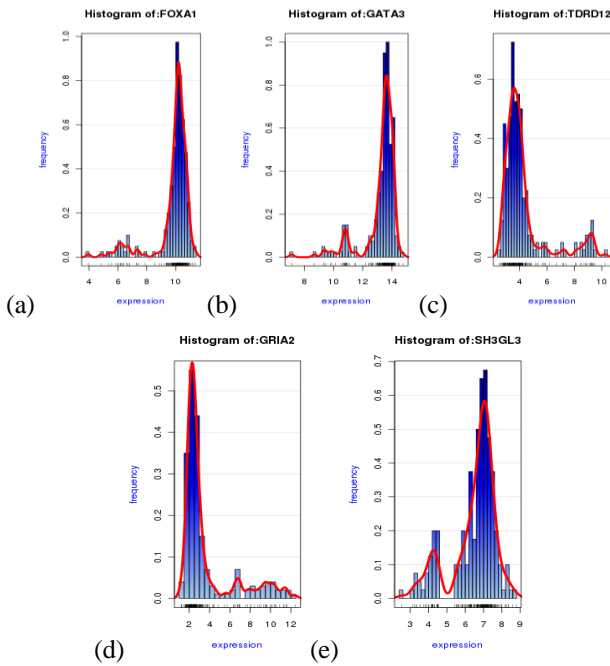


Fig 3: Density analysis of four bimodal genes; (a-d) predicted by all four algorithms at the significance level 0.001 and (e) only predicted by hBI at the same significance level. The horizontal axes represent \log_2 expressions and the vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

Table 7 shows the p values of four algorithms for the genes uniquely predicted by hBI at the significance level 0.01. The data shows that for those bimodal genes predicted by hBI, their ranks of other algorithms are far behind. For instance, the Kurtosis rank of C6orf64 is 18643 and the Kurtosis rank of GULP1 is 22101. Fig 4 shows four of them, which have gene symbols. They are indeed bimodal genes. However other three algorithms failed to predict them. For instance, PSPH was ranked by hBI at the 66th position ($p = 0.002$). Likelihood, Kurtosis and BI ranked it at the 2990th ($p = 0.2$), 13507th ($p = 0.1$), and the 1293th ($p = 0.05$) respectively. This gene is highly expressed in African Americans comparing to European Americans colorectal cancer patients [41]. Also PSPH is expressed at higher level in responding patients versus non-responding group, which support its importance as therapeutic target for non-small-cell lung cancer [42]. RBBP5 was ranked at the 113th position ($p = 0.005$), but was ranked at the 540th position ($p = 0.52$), the 17000th position ($p = 0.36$), and the 974th position ($p = 0.043$) by likelihood, Kurtosis and BI tests. RBBP5 was found to be active in only 40% of Pancreatic ductal adenocarcinomas (PDAs) [43].

Table 7: p values of bimodal genes predicted **ONLY** by hBI at significance level 0.01 for GDS11121

symbol	LH	K	BI	hBI
PSPH	0.27(2990)	0.1(13507)	0.058(1293)	0.002(66)
unknown	0.065(652)	0.03(11438)	0.039(864)	0.004(97)
RBBP9	0.052(540)	0.36(17000)	0.043(974)	0.005(113)
unknown	0.31(3584)	0.02(11311)	0.012(265)	0.008(195)
C6orf64	0.07(713)	0.54(18643)	0.018(416)	0.008(199)
GULP1	0.19(1879)	0.97(22101)	0.026(582)	0.009(226)

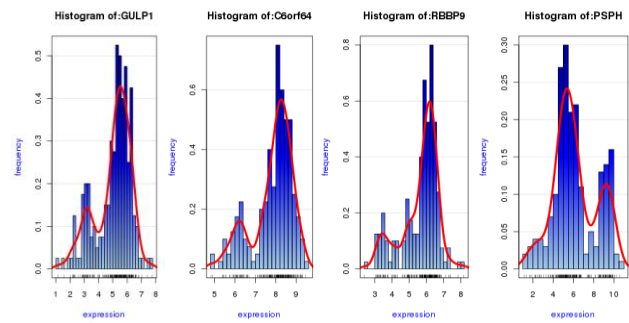


Fig 4: Density analysis of two bimodal genes only predicted by hBI at the significance level 0.01. The horizontal axes represent \log_2 expressions and vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

4 Conclusion

We have proposed a novel bimodal gene prediction algorithm via relaxing the constraints of BimodalIndex algorithm. First, the constraint of cross-cluster homogeneous variance has been removed. It is unrealistic to assume that two clusters of a bimodal gene should have the same variance. The examination of various data sets has clearly shown that one of two clusters, either being of lowly expressed samples or of highly expressed samples is very likely to demonstrate a comparatively flat distribution while the other shows a tight cluster. Second, we deliberately removed the constraint of homogeneous variance across genes because this constraint is certainly confusing. An obviously evidence is that the variance of unimodal genes and bimodal genes will not show homogeneous variance. In addition to these two revisions, we have also emphasised the impact of gaps between consecutive expressions of sorted samples on bimodal formulation. This is because we have observed in real data sets that often lowly expressed samples demonstrate a tight cluster and highly expressed samples show; *i*) a comparatively large variance; and *ii*) distantly departing from the tight cluster of lowly expressed samples or vice versa. In this case the t statistic, although using percentiles to estimate mean values and standard deviations, is still not working well, i.e. the t statistic can be very likely to be small due to the large variance of the highly expressed samples. We therefore introduced a gap impact onto the prediction of bimodal genes. Doing so, we admit that we have introduced a hyper-parameter. In order to remove this hyper-parameter, our future work will employ the Bayesian learning framework to overcome this difficulty. Nevertheless, we have documented our simulations, which all show that our new algorithm is better than the benchmark algorithms in simulated data sets. In the application to real data sets, we show that our new algorithm is partially consistent with benchmark algorithms and does provide some new insights to the analysis of bimodal genes. Importantly, most of the predicted bimodal genes by our new algorithm do show typical bimodality. Particularly, not a small percentage of our unique predictions is unfortunately not favoured by benchmark algorithms. We therefore look forward to some even advanced approach, such as meta-analysis of prediction

to deliver even robust predictions of bimodal genes. Finally, it worth to note that significance analysis is critical to real biological/medical application, we therefore have enhanced the BimodalIndex for using the Besag's sequential Monte Carlo approach to deliver significance analysis.

5 References

- [1] J. L. DeRisi, Iyer, Vishwanath.R., Brown, Patrick O., "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science*, vol. 278, pp. 680-686, 1997.
- [2] Y. Yang, Tashman, Adam., Lee, Jung., Yoon, Seungtai., Mao, Wenyang., Ahn, Kwangmi., Kim, Wonkuk., Mendell, Nancy., Gordon, Derek., Finch, Stephen., "Mixture modeling of microarray gene expression data," *BMC Proceedings*, vol. 1, p. S50, 2007.
- [3] A. Ertel, "Bimodal Gene Expression and Biomarker Discovery," *Cancer Informatics*, vol. 9, pp. 11-14, 2010.
- [4] I. G. Khalil, Hill, C., "Systems biology for cancer," *Current Opinion in Oncology*, vol. 17, pp. 44-48, 2005.
- [5] J. G. Hengstler, Lange, Jost., Kett, Alexandra., Dornhöfer, Nadja., Meinert, Rolf., Arand, Michael., Knapstein, Paul G., Becker, Roger., Oesch, Franz., Tanner, Berno., "Contribution of c-erbB-2 and Topoisomerase II α to Chemoresistance in Ovarian Cancer," *Cancer Research*, vol. 59, pp. 3206-3214, 1999.
- [6] V. N. Kristensen, Edvardsen, Hege., Tsalenko, Anya., Nordgard, Silje H., Sorlie, Therese., Sharan, Roded., Vailaya, Aditya., Ben-Dor, Amir., Lonning, Per Eystein., Lien, Sigbjorn., Omholt, Stig., Syvanen, Ann-Christine., Yakhini, Zohar., Borresen-Dale, Anne-Lise., "Genetic variation in putative regulatory loci controlling gene expression in breast cancer," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 7735-7740, 2006.
- [7] W. F. Anderson, Matsuno, Rayna., "Breast Cancer Heterogeneity: A Mixture of At Least Two Main Types?," *Journal of the National Cancer Institute*, vol. 98, pp. 948-951, 2006.
- [8] F. B. Bertucci, Daniel., "Reasons for breast cancer heterogeneity," *Journal of Biology*, vol. 7, p. 6, 2008.
- [9] S. A. Tomlins, Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A., Chinnaiyan, A.M., "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science*, vol. 310, pp. 644-8, 2005.
- [10] J. Hu, "Cancer outlier detection based on likelihood ratio test," *Bioinformatics*, vol. 24, pp. 2193-9, 2008.
- [11] D. Slamon, Clark, GM., Wong, SG., Levin, WJ., Ullrich, A., McGuire, WL., "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene," *Science*, vol. 235, pp. 177-182, 1987.
- [12] C. Mason, Hanson, Robert., Ossowski, Vicky., Bian, Li., Baier, Leslie., Krakoff, Jonathan., Bogardus, Clifton., "Bimodal distribution of RNA expression levels in human skeletal muscle tissue," *BMC Genomics*, vol. 12, p. 98, 2011.
- [13] T.-O. B. Lim, Rugayah. Morad, Zaki. Hamid, Maimunah A., "Bimodality in Blood Glucose Distribution: is it universal?," *Diabetes Care*, vol. 25, pp. 2212-2217, 2002.
- [14] J. M. Fan, Susanne.J. Zhou, Yue. Barrett-Connor, Elizabeth., "Bimodality of 2-h Plasma Glucose Distributions in Whites," *Diabetes Care*, vol. 28, pp. 1451-1456, 2005.
- [15] I. K. Dozmorov, Nicholas. Tang, Yuhong. Shields, Alan. Pathipvanich, Parima. Jarvis, James.N. Centola, Michael., "Hypervariable genes—experimental error or hidden dynamics," *Nucleic Acids Research*, vol. 32, p. e147, 2004.
- [16] C. Blenkiron, Goldstein, Leonard., Thorne, Natalie., Spiteri, Inmaculada., Chin, Suet-Feung., Dunning, Mark., Barbosa-Morais, Nuno., Teschendorff, Andrew., Green, Andrew., Ellis, Ian., Tavaré, Simon., Caldas, Carlos., Miska, Eric., "MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype," *Genome Biology*, vol. 8, p. R214, 2007.
- [17] S. Chin, Teschendorff, Andrew., Marioni, John., Wang, Yanzhong., Barbosa-Morais, Nuno., Thorne, Natalie., Costa, Jose., Pinder, Sarah., van de Wiel, Mark., Green, Andrew., Ellis, Ian., Porter, Peggy., Tavaré, Simon., Brenton, James., Ylstra, Bauke., Caldas, Carlos., "High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer," *Genome Biology*, vol. 8, p. R215, 2007.
- [18] M. Gort, Broekhuis, Manda., Otter, Renée., Klazinga, Niek., "Improvement of best practice in early breast cancer: actionable surgeon and hospital factors," *Breast Cancer Research and Treatment*, vol. 102, pp. 219-226, 2007.
- [19] R. E. Ellsworth, Hooke, Jeffrey.A., Shriver, Craig.D., Ellsworth, Darrell.L. , "Genomic Heterogeneity of Breast Tumor Pathogenesis," *Clinical Medicine Insights: Oncology* vol. 3, pp. 77-85, 2009.
- [20] L. D. Bradford, "CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants," *Pharmacogenomics*, vol. 3, pp. 229-243, 2002.
- [21] R. E. D. Ellsworth, David.J.; Shriver, Craig.D.; Ellsworth, Darrell.L. , "Breast Cancer in the Personal Genomics Era," *Current Genomics*, vol. 11, pp. 146-161, 2010.
- [22] D. B. A. Agus, Robert W.; Fox, William D.; Lewis, Gail D.; Higgins, Brian.; Pisacane, Paul I.; Lofgren,

- Julie A.; Tindell, Charles; Evans, Douglas P.; Maiese, Krista; Scher, Howard I.; Sliwkowski, Mark X., "Targeting ligand-activated ErbB2 signaling inhibits breast and prostate tumor growth," *Cancer Cell*, vol. 2, pp. 127-137, 2002.
- [23] M. Harris, "Monoclonal antibodies as therapeutic agents for cancer," *The Lancet Oncology*, vol. 5, pp. 292-302, 2004.
- [24] R. E. Nahta, Francisco., "HER2 therapy: Molecular mechanisms of trastuzumab resistance," *Breast Cancer Research*, vol. 8, p. 215, 2006A.
- [25] R. E. Nahta, Francisco.J., "Herceptin: mechanisms of action and resistance," *Cancer Letters*, vol. 232, pp. 123-138, 2006B.
- [26] B. A. R. Chabner, Thomas G., "Chemotherapy and the war on cancer," *Nat Rev Cancer*, vol. 5, pp. 65-72, 2005.
- [27] M. G. Muhsin, Joanne.; Kirkpatrick, Peter., "Gefitinib," *Nat Rev Cancer*, vol. 3, pp. 556-557, 2003.
- [28] P. A. E. Jänne, Jeffrey A.; Johnson, Bruce E., "Epidermal Growth Factor Receptor Mutations in Non-Small-Cell Lung Cancer: Implications for Treatment and Tumor Biology," *Journal of Clinical Oncology*, vol. 23, pp. 3227-3234, 2005.
- [29] M. G. Kris, Natale, Ronald B.,Herbst, Roy S., Lynch, Thomas J., Prager, Diane., Belani, Chandra P., Schiller, Joan H., Kelly, Karen., Spiridonidis, Harris., Sandler, Alan., Albain, Kathy S., Cella, David., Wolf, Michael.K., Averbuch, Steven.D., Ochs, Judith.J., Kay, Andrea.C., "Efficacy of Gefitinib, an Inhibitor of the Epidermal Growth Factor Receptor Tyrosine Kinase, in Symptomatic Patients With Non-Small Cell Lung Cancer," *JAMA: The Journal of the American Medical Association*, vol. 290, pp. 2149-2158, 2003.
- [30] H.-C. C. Wu, De-Kuan.; Huang, Chia-Ting., "Targeted Therapy for Cancer," *J. Cancer Mol.*, vol. 2, pp. 57-66, 2006.
- [31] J. G. J. Paez, Pasi A. Lee, Jeffrey C.; Tracy, Sean.; Greulich, Heidi.; Gabriel, Stacey.; Herman, Paula.; Kaye, Frederic J.; Lindeman, Neal.; Boggon, Titus J.; Naoki, Katsuhiko.; Sasaki, Hidefumi.; Fujii, Yoshitaka.; Eck, Michael J.; Sellers, William R.; Johnson, Bruce E.; Meyerson, Matthew., "EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy," *Science*, vol. 304, pp. 1497-1500, 2004.
- [32] A. E. Teschendorff, Naderi, A., Barbosa-Morais, N.L., Caldas, C., "PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer," *Bioinformatics*, vol. 22, pp. 2269-75, 2006.
- [33] A. Ertel, Tozeren, A., "Switch-like genes populate cell communication pathways and are enriched for extracellular proteins," *BMC Bioinformatics*, vol. 9, p. 3, 2008.
- [34] J. Wang, Wen, S., Symmans, W.F., Pusztai, L., Coombes, K.R., "The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data," *Cancer Inform*, vol. 7, pp. 199-216, 2009.
- [35] M. Gormley, Tozeren, A., "Expression profiles of switch-like genes accurately classify tissue and infectious disease phenotypes in model-based classification," *BMC Bioinformatics*, vol. 9, p. 486, 2008.
- [36] M. Bessarabova, Kirillov, E., Shi, W., Bugrim, A., Nikolsky, Y., Nikolskaya, T., "Bimodal gene expression patterns in breast cancer," *BMC Genomics*, vol. 11, p. S8, 2010.
- [37] J. BESAG and P. CLIFFORD, "Sequential Monte Carlo p-values," *Biometrika*, vol. 78, pp. 301-304, 1991.
- [38] C. E. Metz, "Basic principles of ROC analysis. ," *Seminars in Nuclear Medicine*, vol. 8, pp. 283-288, 1978.
- [39] M. Schmidt, Böhm, Daniel., von Törne, Christian., Steiner, Eric., Puhl, Alexander., Pilch, Henryk., Lehr, Hans-Anton., Hengstler, Jan G., Kölbl, Heinz.,Gehrmann, Mathias., "The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer," *Cancer Research*, vol. 68, pp. 5405-5413, 2008.
- [40] H. Chen, Boutros, Paul, "VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R," *BMC Bioinformatics*, vol. 12, p. 35, 2011.
- [41] B. Jovov, Araujo-Perez, Felix., Sigel, Carlie S., Stratford, Jeran K., McCoy, Amber N., Yeh, Jen Jen., Keku, Temitope., "Differential Gene Expression between African American and European American Colorectal Cancer Patients," *PLoS ONE*, vol. 7, p. e30168, 2012.
- [42] E.-H. Tan, Ramlau, R., Pluzanska, A., Kuo, H.-P., Reck, M., Milanowski, J., Au, J. S.-K., Felip, E., Yang, P.-C., Damyantov, D., Orlov, S., Akimov, M., Delmar, P.,Essioux, L., Hillenbach, C., Klughammer, B., McLoughlin, P. Baselga, J., "A multicentre phase II gene expression profiling study of putative relationships between tumour biomarkers and clinical response with erlotinib in non-small-cell lung cancer," *Annals of Oncology*, vol. 21, pp. 217-222, 2010.
- [43] D. J. Shields, Niessen, Sherry., Murphy, Eric A., Mielgo, Ainhoa., Desgrosellier, Jay S., Lau, Steven K. M., Barnes, Leo A., Lesperance, Jacqueline., Bouvet, Michael., Tarin, David., Cravatt, Benjamin F., Cheresch, David A., "RBBP9: A tumor-associated serine hydrolase activity required for pancreatic neoplasia," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 2189-2194, 2010.

Optimizing the Analysis of Clustered Data

Robert A. Warner, MD

Tigard Research Institute

Tigard, Oregon, USA

Abstract - *Diagnostic data from 2025 subjects were used to simulate a time series and the abilities of several electrocardiographic (ECG) diagnostic parameters to detect prior inferior myocardial infarction (MI) were assessed. The index parameter was the Q wave duration in standard ECG lead aVF. The remaining parameters were techniques used to smooth the Q wave duration data and consisted of moving ten-sample (10-s) means, medians, modes, maxima, minima and threshold counts. When the data were clustered with respect to diagnostic categories, all the smoothing parameters except the moving 10-s maxima had performances highly significantly superior to the Q wave duration parameter. In clustered data, the use of various techniques for smoothing the data substantially improves the detection of events of interest.*

Keywords: clustered data, diagnostic performance

1 Introduction

At scales greater than those of quantum phenomena, events and their attendant manifestations tend to occur in identifiable temporal clusters. The following is an operational definition of a clustered event:

1. *The event of interest is extended in time.*
2. *The event tends to change the measurable values of at least one parameter in a directionally consistent way.*
3. *The duration of the changes in the parameter(s) exceeds the sampling interval of the data that reflect these changes. If this were not the case, the equivalent of aliasing would occur.*

In time series of data, visual examination of line graphs of the values of relevant parameters can suggest temporal clustering of abnormalities of interest. For example, in cardiac monitoring, individual episodes of myocardial ischemia

experienced by patients are typically interspersed with longer periods without ischemia. Consequently, electrocardiographic (ECG) data obtained during the continuous monitoring of patients with coronary artery disease often show intermittent clusters of beats with ischemic ST segment displacement. Similarly, in seismography, foreshock and aftershock "swarms" often occur near the times of major earthquakes. These periods represent clusters of seismic activity that exhibit abnormally high amplitude and frequency.

Besides simply displaying the raw data used to identify events of interest, one can also use various techniques for smoothing the graphed data to more clearly visualize the onsets and offsets of the instances of the events. One technique for smoothing is to calculate and graph the moving averages (means) of the raw data for an appropriate number of samples.[1-3] Although such techniques for smoothing alter the visual appearance of graphed data, questions remain concerning their quantitative effects on the accuracy of detection of the events of interest. In other words, can techniques for smoothing data help assess the likelihood that an apparent occurrence of an event of interest is genuine or are they merely tools for improving the appearances of graphical displays of data? The answer to this question is highly relevant to the diagnostic accuracy of the tests used to identify these events.

Therefore, the first purpose of the present study is to determine statistically whether the use of moving averages (means) improves the diagnostic performances of criteria used to detect clustered events of interest. The second purpose of the study is to evaluate additional possible methods of numerically smoothing data. These include measures of central tendency other than the mean, i.e. the moving medians and modes of the data. Other possible strategies would be to determine the moving maxima and minima of sequential samples of data. Also, for each sequential sample of data, one can determine the number of data points in the sample that equal or exceed an appropriate threshold value. This

parameter can be referred to as the moving data threshold count in a given number of samples.

2 Materials and Methods

2.1 Description of Data

In the present study, I simulated a time series of data using ECG measurements obtained from a group of patients who had been evaluated for possible coronary artery disease. In the matrix used for the simulation, each column of data contains the automated measurements of various ECG parameters and each row contains the specific values of these measurements exhibited by each patient. Thus, in this matrix of ECG data, sequential changes in the values of each parameter that are observed from the top to the bottom of each column represent the differences from each patient to the next, rather than from one time interval to the next. Thus, changes in the parameters' values over a series of different patients are surrogates for changes in those values in the same patient over a period of time. In this simulated time series, the relevant diagnosis of each patient was independently corroborated by either cardiac catheterization (67%) or by systematic clinical evaluation using screening criteria exclusive of the ECG (33%).

2.2 Selection of Patients

I analyzed ECG data obtained from 2025 patients, each of whom had undergone coronary angiography and left ventriculography ($n = 1361$) or systematic clinical evaluation by two or more cardiologists ($n = 664$). Of the total, 366 had cath-proven prior inferior MI (myocardial infarction) and 497 patients had no significant coronary artery disease by catheterization. The other diagnostic categories in the database are prior posterior MI ($n = 66$), prior anterior MI ($n = 275$), combined prior inferior and anterior MI ($n = 157$) and screening normal ($n = 664$). The database is sequentially divided into a learning set and a test set of approximately equal size. In the learning set and in the test set, respectively, the ECG data were sequentially clustered with respect to each of the above six angiographic and clinical screening categories.

2.3 Descriptions of the Diagnostic Parameters

I evaluated the abilities of a variety of digital ECG parameters to detect prior inferior MI. The index parameter is the duration of Q waves in ms. in standard ECG Lead aVF and the other parameters are derivatives of Lead aVF Q wave duration. The duration of Q waves in ECG Lead aVF is a widely used parameter for diagnosing prior inferior MI.[4-5] One of these derivative parameters is the ten-sample central moving average (10-s CMA) of the Lead aVF Q wave duration data. Each value of the 10-s CMA is the mean of the five Q wave duration measurements immediately preceding it and the five Q wave duration measurements immediately following it. This contrasts with the type of moving average used in technical charts of stock market prices. In a chart that shows the changes in the price of a stock, a ten-sample moving average represents the mean of the closing prices of the preceding ten trading days. This would therefore be the ten-sample preceding moving average. To determine the effects of sample size, I also evaluated the eight, six, four and two-sample CMAs. These are calculated by averaging the four, three, two and one samples, respectively, before and after the row where the CMA is entered.

I also evaluated the diagnostic performances of the following derivative parameters besides the 10-s CMA:

1. The ten-sample central moving median
2. The ten-sample central moving mode
3. The ten-sample central moving maximum
4. The ten-sample central moving minimum
5. The ten-sample moving threshold count. The threshold value used was 30 ms., since a Q wave duration >30 ms. in Lead aVF is a widely used ECG diagnostic criterion for prior inferior MI.[4-5]

2.4 Analysis of the Data

I used receiver operating characteristic (ROC) curves to determine the diagnostic sensitivities of each of the above parameters at 100% specificity. I tested the statistical significances of any differences in diagnostic performance between the Q wave duration in Lead aVF vs. each of the above derivatives of this measurement using chi square analysis. To avoid a Type I error associated with multiple comparisons, an alpha <0.01 was selected a priori to indicate statistical significance.

In addition, to compare the effects of the various smoothing techniques on clustered vs. non-clustered data, I repeated the analyses with the data no longer grouped with respect to the six diagnostic categories, but instead sorted in ascending order of the subjects' arbitrarily assigned identification numbers.

Finally, I evaluated the effects on diagnostic performance of using various sample sizes other than ten.

3 Results

Figures 1A and 1B show the clustering of the ECG data by diagnostic category. Figure 1A shows the Lead aVF Q wave duration data and the Figure 1B shows the data for the 10-s CMAs. The left half of each figure shows the data for the learning set and the right half of each panel shows the data for the test set. In both the learning and the test sets, the values of both parameters are greater in the inferior MI (Diagnostic Categories 2 and 4) and the posterior MI (Diagnostic Category 5) groups. Clustering of the data in the posterior MI group occurs because posterior MI is commonly associated with inferior MI.[6-7] In each panel, the vertical arrows show the clusters of abnormal values of the diagnostic parameters associated with inferior MI. Comparing Figures 1A and 1B demonstrates the visual effects of smoothing the diagnostic data using the 10-s CMA. Despite the fact that Figures 1A and 1B incorporate nearly the same number of data points (2025 and 2020, respectively), the appearance of the graph in Figure 1A is much more granular than that of Figure 1B.

Figure 1A

Raw aVF QD Data

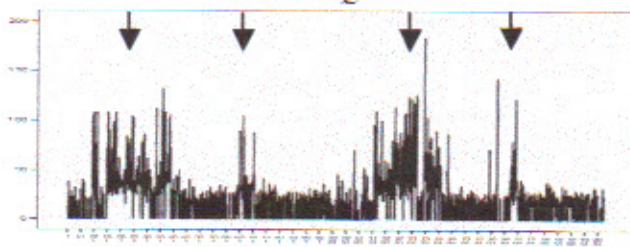


Figure 1B

10-s CMA Data

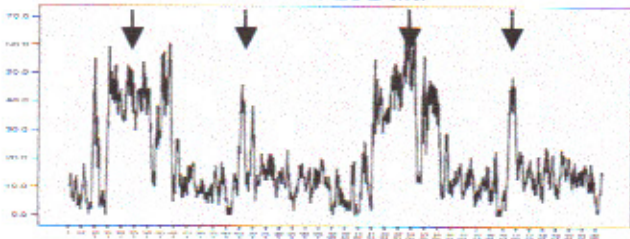


Table 1A shows the sensitivities at 100% specificity for detecting prior inferior MI exhibited by the Q wave duration in Lead aVF vs. six different ten-sample derivatives of that parameter. The results shown in Table 1A were obtained when the data were clustered with respect to the independently established diagnostic categories. In all cases, the differences in sensitivity between aVF Q wave duration and the ten-sample derivatives are highly statistically significant. For the ten-sample mean, median, mode, minimum and threshold count ≥ 30 ms., the sensitivities at 100% specificity are greater than that of the aVF Q wave duration. Conversely, the sensitivity of the 10-sample maximum is less than that of the aVF Q wave duration.

Table 1A

Diagnostic Performances in Clustered Data

Parameter	Threshold Value (ms.)	Sens. @ 100% Spec.	P Value*
AVF QD	73	11	
10-s CMA	28	90	4×10^{-101}
10-s Median	31	86	5×10^{-92}
10-s Mode	31	72	3×10^{-63}
10-s Maximum	143	0	8×10^{-11}
10-s Minimum	7	58	8×10^{-41}
10-s Ct. ≥ 30	8.5	54	1×10^{-35}

*Compared to the parameter aVF QD

Table 1B

Diagnostic Performances in Non-Clustered Data

Parameter	Threshold Value (ms.)	Sens. @ 100% Spec.	P Value*
AVF QD	73	11	
10-s CMA	41	3	3×10^{-5}
10-s Median	37	5	3×10^{-3}
10-s Mode	51	0	8×10^{-11}
10-s Maximum	163	1	2×10^{-8}
10-s Minimum	6	3	3×10^{-5}
10-s Ct. ≥ 30	8	3	3×10^{-5}

*Compared to the parameter aVF QD

Furthermore, the diagnostic threshold required to achieve 100% diagnostic specificity by aVF Q wave duration is over twice as great as the thresholds required by the ten-sample mean, median, mode, minimum and count ≥ 30 ms parameters. In contrast,

the diagnostic threshold required to achieve 100% by the ten-sample maximum parameter was nearly twice as great as that required by aVF Q duration. The highest sensitivity at 100% specificity is 90% and is exhibited by the 10-s CMA for this parameter at a diagnostic threshold of 28 ms. At 100% specificity, the ten-sample's median's sensitivity is numerically lower at 86%, but the difference between the sensitivities of the 10-s CMA and the 10-sample median is not statistically significant (chi square = 2.5, P = NS).

Further analysis of the ROC curves of the diagnostic parameters reveals that to duplicate the 90% sensitivity of the 10-s CMA, the requisite threshold value for aVF Q wave duration is 17 ms. and is associated with a diagnostic specificity of only 65%. In addition, applying the 73 ms. threshold of the aVF Q wave duration to the 10-s CMA parameter yields 100% specificity, but 0% sensitivity for prior inferior MI. Applying the 28 ms. threshold of the 10-s CMA parameter to the aVF Q wave duration parameter yields only 72% sensitivity and 94% specificity. The diagnostic superiority of using the 28 ms. threshold for the 10-s CMA parameter compared to using the same threshold for the aVF Q wave duration parameter is highly statistically significant (chi square = 59.2, P = 1×10^{-14}).

Table 1B also shows the sensitivities at 100% specificity for detecting prior inferior MI in the exhibited by the Q wave duration in Lead aVF vs. six different ten-sample derivatives of that parameter. However, in contrast to the results shown in Table 1A, the performance data in Table 1B were obtained when the data were not clustered with respect to diagnostic categories. Instead, the performances listed in Table 1B were obtained when the ECG data were sorted in ascending order of the subjects' identification numbers that have nothing to do with their diagnoses. For the non-clustered data, all the ten-sample derivatives have sensitivities at 100% specificity that are highly statistically significantly lower than the Lead aVF Q wave duration parameter. Thus, using the derivative parameters of the 10-sample means, medians, modes, minima and threshold counts improves diagnostic performance over the raw Lead aVF Q wave duration data only if the data are clustered with respect to the diagnostic categories. Conversely, if the data are not clustered with respect to the diagnostic categories, the above derivative parameters have significantly poorer diagnostic performances than the raw Lead aVF Q wave duration data.

Figure 2A shows histograms of the values of the parameter Lead aVF Q wave duration in the catheterization normal (lower panel) vs. the prior inferior MI (upper panel) groups. Figure 2B shows histograms of the values of the parameter 10-s CMA in the catheterization normal (lower panel) vs. the prior inferior MI (upper panel) groups.

Figure 2A

Histogram of Raw AVF QD Data

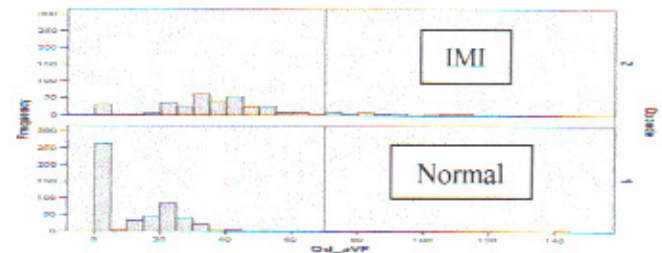
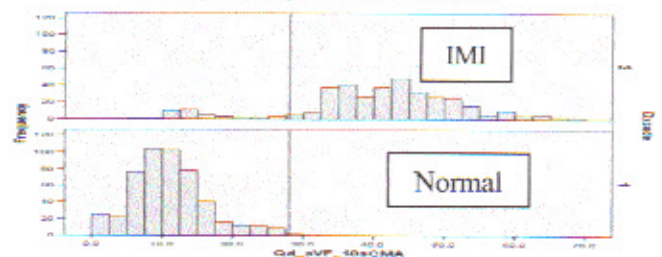


Figure 2B

Histograms of 10-s CMA Data



In both Figures 2A and 2B, the data are clustered with respect to the diagnostic categories. The vertical line near the middle of each panel indicates the diagnostic threshold value that yields 100% specificity for the detection of prior inferior MI. Comparing Figures 2A and 2B reveals that the 10-s CMA parameter provides much more complete separation of the normal vs. the inferior MI subjects than does the Lead aVF Q wave duration. Also, the lower panel of Figure 2A shows that distribution of Lead aVF Q wave durations in the normal subjects is highly skewed to the right. In contrast, the lower panel of Figure 2B shows that the distribution of values of the 10-s CMA in the normal subjects is gaussian.

Table 2 shows the diagnostic sensitivities at 100% specificity of the ten, eight, six, four and two-sample CMAs for the data clustered by diagnostic category. The diagnostic performances of the eight- and ten-sample CMAs are similar. However the sensitivities of the six-, four- and two-sample CMAs are significantly inferior to the sensitivity of the 10-s CMA.

Table 2
Effects of Sample Size on Diagnostic Performance

Parameter	Thresh old Value (ms.)	Scns. @ 100% Spec.	P Value*
10-Sample CMA	28	90	
8-Sample CMA	30	88	NS
6-Sample CMA	35	72	9×10^{-10}
4-Sample CMA	45	38	2×10^{-48}
2-Sample CMA	63	13	6×10^{-96}

*Compared to 10-Sample CMA

4 Conclusions

The present study shows that several techniques that can be used to smooth data for the visual depiction of clustered events significantly improve the diagnostic performances of diagnostic criteria used to identify those events. The greatest numerical improvement in diagnostic performance for prior inferior MI over the raw data (Q wave duration in ECG Lead aVF) was the 10-s CMA. This parameter is numerically superior to the other methods of data smoothing that were tested, and is statistically significantly superior to all the other methods except the 10-s central median. The diagnostic superiority of the 10-s CMA to the raw Lead aVF Q wave durations for the clustered diagnostic data is also supported by comparing Figure 1A and Figure 1B. In the clusters of diagnostic data in the inferior and posterior MI diagnostic categories, the peaks of the clusters of the abnormal data are proportionately further above baseline in the 10-s CMA data than in the Lead aVF Q wave duration data.

Figures 2A and 2B also show the superior ability of the 10-s CMA to the aVF Q wave duration data for distinguishing normal subjects from those with prior inferior MI. The histograms of the 10-s CMA data demonstrate much better separation of the normal from the prior inferior MI groups than do the Lead aVF Q wave duration data. In addition, the lower panel of Figure 2A shows that the distribution of values of Lead aVF Q wave duration in the normal subjects is heavily skewed to the right. This is because many subjects without cardiac disease have no Q waves in this ECG lead.[4-8] However, the lower panel of Figure 2B shows that the distribution of values of the parameter 10-s CMA in the normal

subjects is gaussian. Therefore, an additional advantage of using the Lead aVF 10-s CMA parameter rather than the Lead aVF Q wave duration parameter is that the former permits the use of more robust parametric statistics in the analysis of diagnostic data.

In clustered data, the analysis of moving samples incorporates the data from multiple proximate instances of the same abnormality in each cluster. In other words, in data relevant to clustered events of interest, the likelihood that each data point is a true positive or a true negative increases if it is surrounded by other data with similar values. The phenomenon of being surrounded by other data obviously confers a diagnostic advantage only if the data are clustered with respect to the same diagnosis. The data used in the present study represent a simulated, rather than a true, time series. However, whether each datum represents a different patient instead of a different point in time, the data are still clustered with respect to a diagnosis of interest.

The present study has emphasized the use of ten samples for calculating the central moving averages. However, as shown by the data in Table 2, using eight samples for calculating the central moving averages produced a statistically similar diagnostic result. Table 2 also shows that as the number of samples used to calculate the central moving average decreases, the diagnostic threshold required to achieve 100% specificity increases. The data in Table 2 show that this increase in the threshold values required for 100% diagnostic specificity is non-linear.

Comparing Figure 1B to Figure 1A confirms that plotting the 10-s CMA data produces a smoother graph than plotting the raw Lead aVF Q wave data. The greater smoothness of the graph in Figure 1B is evident despite the fact that it incorporates nearly as many data points as the graph in Figure 1A (2020 vs. 2025). In other words, the greater smoothness of the graph achieved by plotting the 10-s CMA data is achieved by consolidating, rather than by discarding diagnostically relevant data. Although both Figures 1A and 1B provide visual evidence that the data are clustered, it remains diagnostically important to analyze the data numerically. This is because artifacts are often clustered as well. For example, in the medical monitoring of patients with possible cardiac disease, changes in the ECG can occur not only because of episodes of myocardial ischemia, but also if the patient moves in bed or if one or more of the ECGs leads becomes loose. Therefore, it is important to

determine the diagnostic thresholds of the clusters associated with myocardial ischemia so that they can be distinguished from artifacts with qualitatively similar appearances. Also, in seismography, it is important to know the thresholds needed to detect foreshock swarms of seismic activity and distinguish them from vibrations produced by intermittent highway or railway traffic.

The findings of the present study are important because many events of interest are not instantaneous, but instead are extended in time, i.e. occur in clusters. The data show that using parameters such as the 10-s CMA to analyze the quantifiable manifestations of these events not only produces smoother visual depictions of the events, but also significantly improves the accuracy with which we can detect them.

5 References

- [1] Chatfield, C., 2004, The analysis of time series, an introduction, sixth edition: New York, Chapman & Hall/CRC.
- [2] Karl, J.H., 1989, An introduction to digital signal processing, Academic Press, Inc., San Diego, California 92101.
- [3] Lyons, Richard G. Understanding digital signal processing. Upper Saddle River: Prentice Hall PTR, 2001. ISBN 0-201-63467-8.
- [4] Warner RA, Wagner GS, Ideker, RE The ability of the QRS complex to determine the location and size of myocardial infarcts in Acute Coronary Care edited by Califf and Wagner, Boston, Martinus Nijhoff, 1984.
- [5] Warner RA, Hill N, Sheehe P, Mookherjee S, Fruchan CT. Improved criteria for the diagnosis of inferior myocardial infarction. *Circulation* 66:422-428, 1982.
- [6] Warner R. New developments in quantitative electrocardiography. *Proceedings of the Engineering Foundation*. 10:207-207, 1986.
- [7] Hill NE, Warner RA, Mookherjee S, Smulyan H. Comparison of optimal scalar electrocardiographic, orthogonal electrocardiographic and vectorcardiographic criteria for diagnosing inferior and anterior myocardial infarction. *Am. J. Cardiol.* 54:274-276, 1984.29
- [8] Warner RA, Battaglia, J, Hill NE, Mookherjee S, Smulyan H. Importance of the terminal portion of the QRS in the electrocardiographic diagnosis of inferior myocardial infarction. *Am. J. Cardiol.* 55:896-899, 1985.32.

Transition Initiation Sites (TIS) Recognition in DNA Sequence using Machine Learning

Muhammad Hossain and Kanaan Faisal

Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
{mdimtiazh, Kanaan}@kfupm.edu.sa

Abstract— Transition Initiation Sites (TIS) prediction is a challenging problem in computational biology. In the literature TIS is predicted using various machine learning techniques such as Neural Network (NN), Support Vector Machine, etc. We have applied Principal Component Analysis (PCA) to remove highly correlated features which improves the performance in terms of time and accuracy. In this paper we have used Group Model of Data Handling (GMDH) based algorithm Abductive Network (AN) to predict TIS and got accuracy of 93%.

Keywords- Bioinformatics, Transition Initiation Sites (TIS), mRNA sequence, Machine Learning, Neural Network, Abductive Network, GMDH.

I. INTRODUCTION

Proteins are synthesized from mRNAs by a process called translation. The region

at which the process initiates is called the Translation Initiation Site (TIS). The coding sequence is ranked by non-coding regions which are the 5' and 3' UnTranslated Region (UTR) respectively. The translation initiation site prediction problem is to correctly identify the TIS in a mRNA or cDNA sequence. This forms an important step in genomic analysis to determine protein coding from nucleotide sequences. In his research we have predicted TIS in human mRNA sequence.

In eukaryotes, the scanning model postulates that the ribosome attaches first to the 5' end of the mRNA and scans along the 5' to 3' direction until it encounters the first AUG. The problem of predicting the TIS is compounded in real-life sequence analysis by the difficulty of obtaining full-length and error-free mRNA sequences.

Machine learning techniques have been used successfully in TIS prediction using the mRNA or cDNA sequence.

In this research the feature dimension reduction is performed using PCA to select the most significant features and finally AN and Multi Layer Perceptron is used for TIS prediction.

The rest of the paper is organized as follows. Section 2 deals with recent literatures. Section 3 describes the proposed recognition system. Section 4 shows experimental results. Finally Section 5 mentioned conclusion and future work.

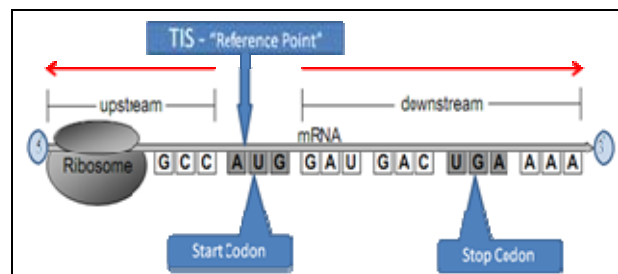


Figure 1. TIS Terminology

II. LITERATURE REVIEW

Pedersen and Nielsen [1] found that almost 40% of the mRNAs extracted from GenBank contain upstream AUGs. This accords with the scanning hypothesis that the ribosome operates in a linear fashion on the sequence to recognize the start site. They have trained an artificial neural network (ANN) on a 203 nucleotide window centered on the AUG. They obtained results of 78% accuracy on start AUGs and 87% accuracy on non-start AUGs on their vertebrate dataset, giving an overall accuracy of 85%. This system is available on the Internet as the NetStart 1.0 prediction server.

Zien et al. [2] obtain improved results on the same vertebrate dataset from Pedersen and Nielsen by using support vector machines (SVM). The same 203 nucleotide window is used as the underlying features to be learnt. They show how to obtain improvements by appropriate engineering of the kernel function - using a locality-improved kernel with a small window on each position, a codon-improved kernel using codon structure in the downstream sequence and a Salzberg kernel using conditional positional probabilities. With the nucleotide-based kernels [3], they obtain an accuracy of 69.9% and 94.1% on start and non-start AUGs respectively, giving an overall accuracy of 88.1%. The Salzberg kernel gives an overall accuracy of 88.6%.

Hatzigeorgiou [4] reports a highly accurate TIS prediction program, DIANA-TIS, using ANN trained on human sequences. Their dataset contains full-length cDNA sequences which has been altered for errors. An overall accuracy of 94% is obtained using an integrated method which combines a consensus ANN with a coding ANN together with the ribosome scanning model.

Zeng et al. [5] obtained 94% overall accuracy on the dataset used in [1, 2, 4] by using simple feature generation and selection on a variety of standard machine learning methods. In the work of Zien et al. [2] and Hatzigeorgiou [4], improved TIS prediction is obtained by a more complex method. Zeng et al showed that the use of simple feature generation followed by correlation-based feature selection allows a variety of standard machine learning methods such as ANN, decision trees, SVM, Naive Bayes to obtain accurate TIS prediction. Feature selection results in only a very small number of features, at most 13, to get good results. The results from the simple TIS prediction are directly comparable with Zien et al. [2] and Pedersen and Nielsen [1]. The highest overall accuracy obtained is 89.4% which is better than previous results on this dataset. Incorporating distance as a feature improves this result. Finally with the use of a scanning model, they have obtained an overall accuracy of 94.4% which compares very favorably to Hatzigeorgiou [4].

III. DATASET

The dataset used is the vertebrate dataset created by Pedersen and Nielsen [1]. This dataset was also used by [2, 4, 5]. So our results can be compared directly with the two previous works. The original dataset of Pedersen and Nielsen [1] consists of a selected set of vertebrate genomic sequences extracted from GenBank [6]. It consists of sequences from *Bos taurus* (cow), *Gallus gal-lus* (chicken), *Homo sapiens* (man), *Mus musculus* (mouse), *Oryctolagus cuniculus* (rabbit), *Rattus norvegicus* (rat), *Sus scrofa* (pig), and *Xenopus laevis* (African clawed frog). It has been shown by Pedersen and Nielsen that these vertebrates have similar start codon contexts [1]. These sequences are processed by removing possible introns and joining the exons. This is analogous to the splicing of mRNA sequences. From these sequences, only those with an annotated translation initiation site, and with at least 10 upstream nucleotides as well as 150 downstream nucleotides are selected. The sequences are altered to remove those belonging to same gene families, homologous genes from Different organisms, and redundant sequences, so as to avoid over-optimistic performance resulting from biased data [7]. This resulting dataset consists of 3312 sequences. Since the dataset is processed DNA, the TIS site is ATG. In total, there are 13503 ATG sites. Of the possible ATG start sites, 3312 (24.5%) are the true start ATGs while the other 10191 (75.5%) are non-start ATGs. The dataset is available in

<http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html>

An example entry from this dataset is given below in Figure 2 and Figure 1 shows the basic TIS terminologies.

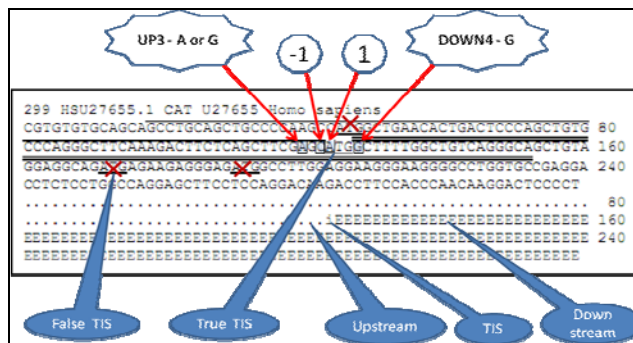


Figure 2. Human mRNA example

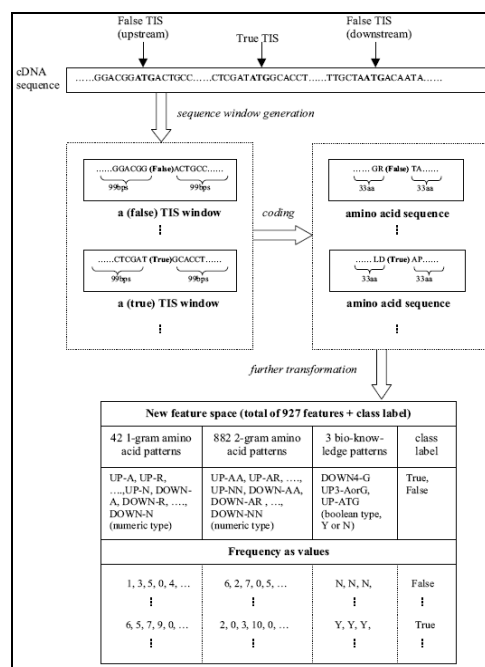


Figure 3. Feature Extraction

A. Feature Extraction

Frequency of k-gram amino acid. (k = 1,2,3.. Amino acid patterns) –

- Count the frequency of amino acid X in upstream and downstream. (20 amino acids + 1 stop symbols = 21 x 2).
- Count the frequency of amino acid of XY in upstream and downstream. (21 x 21 x 2 = 882).
- 3 biological knowledge: “DOWN4-G”, “UP3-AorG” and “UP-ATG”.

B. Feature Vector

After extracting feature we have a feather vector 13,310 X 927. We have used PCA to reduce the feature dimension from

927 to 70 which make the prediction faster by the machine learning techniques.

C. Feature Vector

After extracting feature we have a feather vector 13,310 X 927. We have used PCA to reduce the feature dimension from 927 to 70 which make the prediction faster by the machine learning techniques.

IV. MACHINE LEARNING TECHNIQUES

ANN, SVM, HMM, FUZZY LOGIC, Bayesian Network, Group Model of Data Handling Based Abductive Network (AN), etc are the well known techniques in machine learning. We have used ANN and AN to predict TIS.

A. Abductive Network (AN)

Abductive Networks approach based on the Group Method of Data Handling (GMDH) algorithm as an alternative learning tool. The GMDH approach to classification offers the advantage of simplified and more automated model synthesis. Abductory Inductive Mechanism (AIM) is a Machine Learning tool that automatically discover network solutions to complex decision, prediction, control and classification problems. The tool generate a mathematical models from relationships it finds in the training data. It does so by trying out all potential relationships of linear, multiple and polynomial on various combination of input variables. It iteratively build a network of numerical functional elements based on prediction performance using Predicted Square Error (PSE).

$$PSE = FSE + CPM (2K/N)\sigma^2$$

- FSE Fitted Square Error.
- CPM Complexity Penalty Multiplier.
- K # of Coefficients.
- N # of Inputs.
- σ Estimation of predicted error.

The unique property of automatic selection of only the most relevant input features by abductive network models gives useful insight into the contribution of the various features in the dataset.

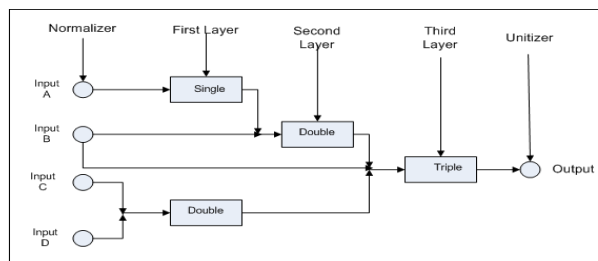


Figure 4. AN Functional Elements

AN Functional Elements

- **Normaliser:** Transforms the original input into a normalized variable having a mean of zero and a variance of unity.
- **Unitizer:** Restores the result to the original problem space
- **Node:** The node has input(1, 2 or 3) and the polynomial equation is limited to the third degree, that is:

$$y = z_0 + z_1x + z_2x^2 + z_3x^3$$

V. NEURAL NETWORK

In this paper Multi Layer Perceptron (MLP) classifier is used which is one of the popular Artificial neural networks (ANN) consist of simple processing elements and a high degree of interconnection. The elements are organized into an initial input layer, intermediate "hidden" layers, and a final output layer (Figure 7). In MLP information proceeds from the input layer to the output layer through hidden layer(s). It uses back propagation algorithm makes to learn the weights within the elements and construct arbitrarily complex nonlinear decision boundaries to separate multiple classes.

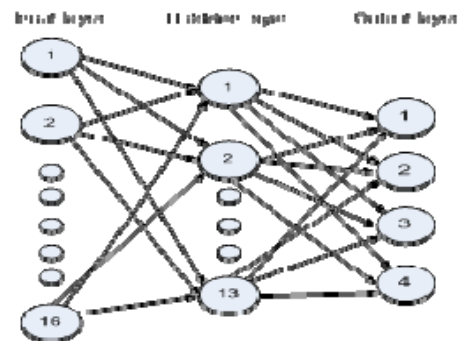


Figure 5. Multi Layer Perceptron (MLP)

Actually MLP operates in two distinct phases. The first is the recall phase in which the training pattern is presented to the input layer of the network and a corresponding output is recalled at the output layer. The second is the learning phase in which the network adjusts its synaptic weights in order to minimize the error between the recalled pattern and the correct pattern given by a teacher (supervised learning). The neural network is only as good as the data set with which it is trained upon. When selecting training data, the designer should consider:

- Whether all important features are covered
- What are the important/necessary features

VI. PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is used to reduce highly correlated features. PCA was first introduced by Pearson in 1901 and become a standard tool in modern data analysis.

PCA is actually a technique to find the directions in which a cloud of data points is stretched most. PCA perform linear transformation by choosing a new coordinate system in such a way that greatest variance by any projection of the data set comes to lie on the first axis (the first principal component). PCA can be used for reducing dimensionality by eliminating the later principal components.

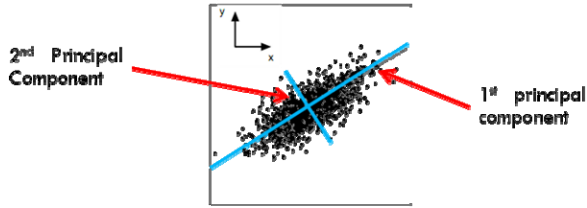


Figure 6. Figure 1: PCA

The objective of PCA is to perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible. PCA performs a linear mapping of the data to a lower dimensional space in such a way, that the variance of the data in the low-dimensional representation is maximized. At first, the correlation matrix of the data is constructed and the eigenvectors on this matrix are computed so that the eigenvectors that correspond to the largest eigenvalues (the principal components) can be used to reconstruct a large fraction of the variance of the original data. Moreover, the first few eigenvectors can often be interpreted in terms of the large-scale physical behavior of the system. The original space (with dimension of the number of points) has been reduced (with data loss, but hopefully retaining the most important variance) to the space spanned by a few eigenvectors.

VII. FRAMEWORK OF TIS PREDICTION

At first the TIS features are reduced using the Principal Component Analysis (PCA) by removing highly correlated features. 70% of the digits used for training the AN and MLP to build the model and 30% were used for testing. In the following sections the steps are described in details.

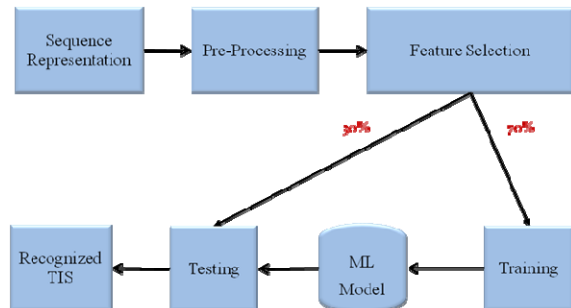


Figure 7. Figure 7: Framework for TIS Prediction

VIII. EXPERIMENTAL RESULTS

In this paper, the performance of two standard machine learning classifiers on the selected features is evaluated. We have used the Abductive Network model and Neural Network. We have used Abductive Network as a classifier. And later on we have used the features that are chosen by the AN as an input for Neural Network. Each ATG is labeled whether or not it's a true TIS site. Thus, each ATG in a sequence from the set of training sequences contributes one training instance. Training and testing is performed with a random sampling method. 70% of the dataset is taken as training and 30% is kept for testing.

The results testing are evaluated using standard performance measures. To describe the performance, the results from testing a classifier can be arranged in the following matrix:

TABLE 1 SLANDERED PERFORMANCE MEASURES

	Classified as Yes	Classified as No
Actual Yes Class	No. of True Positive	No. of False Negative
Actual No Class	No. of Fales Positive	No. of True Negative

We have the following measures:

$$\text{True Positive Rate (also called Sensitivity)} = 100 \times \frac{TP}{TP + FN}$$

$$\text{True Negative Rate} = 100 \times \frac{TN}{TN + FP}$$

$$\text{Specificity} = 100 \times \frac{TP}{TP + FP}$$

$$\text{Overall Accuracy} = 100 \times \frac{TP + TN}{TP + TN + FN + FP}$$

$$\text{Adjusted Accuracy} = \frac{TPRate + TNRate}{2}$$

Because the dataset consists of significantly more negative than positive examples, we have also used Adjusted Accuracy as a performance measure which gives a fairer comparison than overall accuracy for skewed datasets such as the one here where the number of non-start ATGs is disproportionately larger than the number of start ATGs.

For example, if 80% of the ATGs are non-start, then a trivial predictor which simply classifiers every ATG as non-start would already obtain an overall accuracy of 80%. Adjusted accuracy, on the other hand, is less skewed giving

50% accuracy. As the results in the literature do not give sufficient data to compare on the basis of adjusted accuracy, we continue to use overall accuracy in the comparisons with existing work.

TABLE 2 RESULTS

	Classified as No	Classified as Yes
Actual No Class	True Negative 2887	False Positive 177
Actual Yes Class	Fales Negative 211	True Positive 737

True Positive Rate (Sensitivity) = 77.74%
 True Negative Rate = 94.22%
 False Positive Rate (Specificity) = 5.77%
 Adjusted Accuracy = 85.98%
 Overall Accuracy = 90.33%

A. ROC CURVE ANALYSIS for AN

TABLE 3 RESULTS FOR AN

AUC	S.E.	95%	C.I.	Comment
0.95925	0.00455	0.95033	0.96816	Excellent test
Standardized AUC	100.9517		1-tail p-value	0.000000
The area is statistically greater than 0.5				

Cut-off point for best Sensitivity and Specificity (blu circle in plot) = 0.2536

In the ROC plot, the cut-off point is the closest to [0,1] point or, if you want, the closest to the green line

Table at cut-off point

TABLE 4 CUT-OFF VALUES

cut-off point	
867	423
81	2641

Prevalence: 23.6%

Sensitivity (probability that test is positive on unhealthy subject): 91.5%

95% confidence interval: 89.7% - 93.2%
 False positive proportion: 8.5%

Specificity (probability that test is negative on healthy subject): 86.2%
 95% confidence interval: 85.0% - 87.4%
 False negative proportion: 13.8%
 Youden's Index (a perfect test would have a Youden index of +1): 0.7765

Precision or Predictivity of positive test (probability that a subject is unhealthy when test is positive): 67.2%
 95% confidence interval: 64.6% - 69.8%
 Positive Likelihood Ratio: 6.6
 Moderate increase in possibility of disease presence

Predictivity of negative test (probability that a subject is healthy when test is negative): 97.0%
 95% confidence interval: 96.4% - 97.7%
 Negative Likelihood Ratio: 0.1
 Large (often conclusive) increase in possibility of disease absence

F-measure: 77.5%
 Accuracy or Potency: 87.4%
 Mis-classification Rate: 12.6%

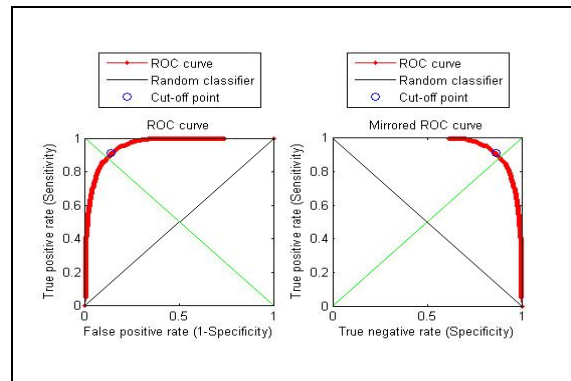


Figure 8. Figuroc for AN

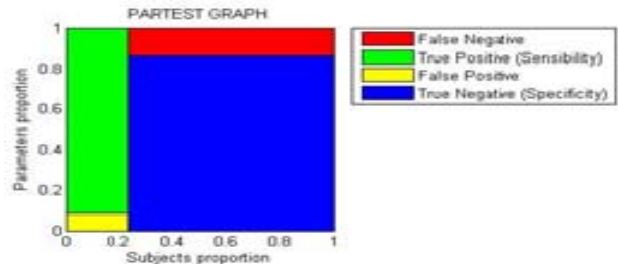


Figure 9. PARTEST GRAPH for AN

B. ROC CURVE ANALYSIS for ANN

TABLE 5 RESULTS OF ANN

AUC	S.E.	95%	C.I.	Comment
0.97858	0.00333	0.97205	0.98510	Excellent test
Standardized AUC	143.7515		1-tail p-value	0.000000
The area is statistically greater than 0.5				

Cut-off point for best Sensitivity and Specificity (blue circle in plot)= 0.3439

In the ROC plot, the cut-off point is the closest to [0,1] point or, if you want, the closest to the green line

Table at cut-off point

TABLE 6 CUT-OFF VALUES

cut-off point	
877	211
71	2853

Prevalence: 23.6%

Sensitivity (probability that test is positive on unhealthy subject): 92.5%

95% confidence interval: 90.8% - 94.2%

False positive proportion: 7.5%

Specificity (probability that test is negative on healthy subject): 93.1%

95% confidence interval: 92.2% - 94.0%

False negative proportion: 6.9%

Youden's Index (a perfect test would have a Youden index of +1): 0.8562

Precision or Predictivity of positive test (probability that a subject is unhealthy when test is positive): 80.6%

95% confidence interval: 78.3% - 83.0%

Positive Likelihood Ratio: 13.4

Large (often conclusive) increase in possibility of disease presence

Predictivity of negative test (probability that a subject is healthy when test is negative): 97.6%

95% confidence interval: 97.0% - 98.1%

Negative Likelihood Ratio: 0.1

Large (often conclusive) increase in possibility of disease absence

F-measure: 86.1%

Accuracy or Potency: 93.0%

Mis-classification Rate: 7.0%

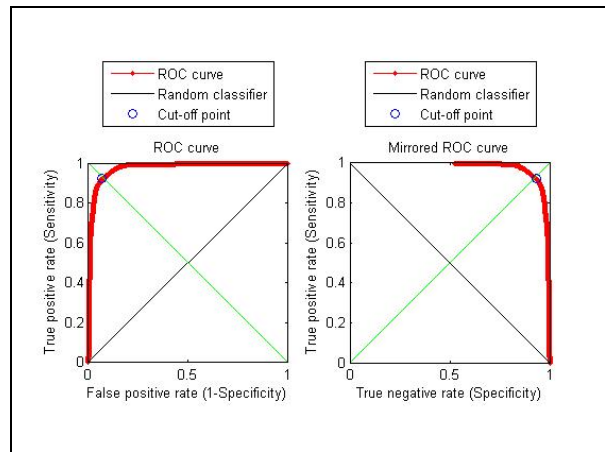


Figure 10. ROC for ANN

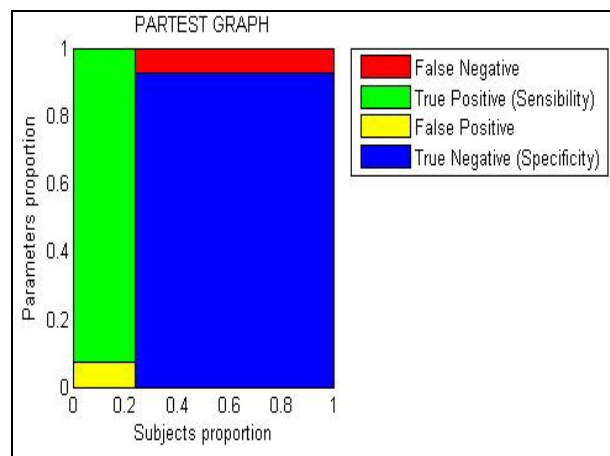


Figure 11. PARTEST GRAPH of ANN

Figure 12.

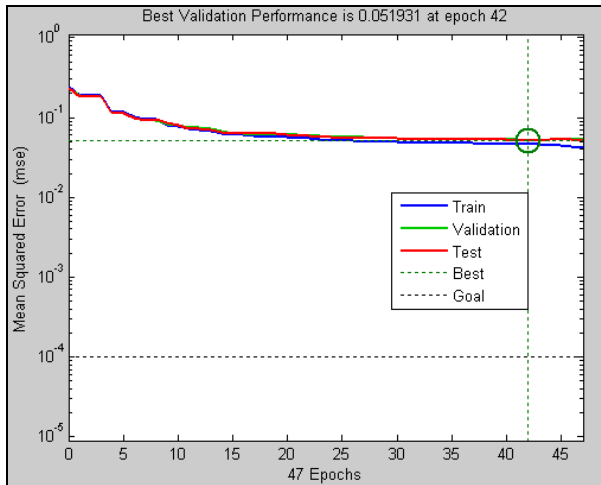


Figure 13. Learning Curve of ANN

Figure 14.

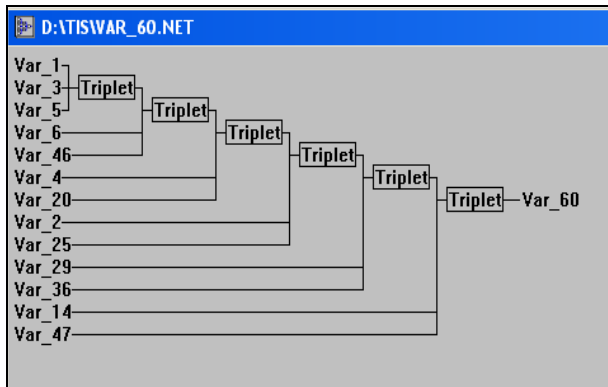


Figure 15. Trained Network of AN

D:\TIS\VAR_69.EVL

Evaluation of Network 'D:\TIS\VAR_69.NET'
Using Data from Table 'TIS68_EVAL'

Summary Statistics for output 'Var_69'

	evaluation	training
number of observations	4012	9363
average absolute error	0.14655	
absolute error standard deviation	0.21789	
average squared error	0.068939	0.076815
normalized mean squared error	0.38166	
squared error standard deviation	0.15794	
maximum absolute error	1.0000	
database output minimum	0.00000	0.00000
database output maximum	1.0000	1.0000
database output mean	0.23629	0.25248
database output standard deviation	0.42486	0.43446
network output mean	0.24941	
network output standard deviation	0.33109	
R-squared	0.61802	
root of predicted squared error		0.27994
predicted squared error		0.078367

Figure 16. Abductive Network of TIS Prediction Performance

IX. CONCLUSIONS

The experiments showed that the performance of Neural Network is better than Abductive Network. The overall accuracy of Neural Network is about 93.8% while the overall accuracy of Abductive Network is about 90.3%.

X. ACKNOWLEDGMENT

The Authors would like to acknowledge the support of King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

XI. REFERENCES

[1] Pedersen, A.G. and Nielsen, H., Neural network prediction of translation initiation sites in eukaryotes: Perspectives for EST and genome analysis, Proc. 5th International Conference on Intelligent Systems for Molecular Biology, pp.226-233, 1997.

[2] Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lemmen, C., Smola, A., Lengauer, T., and Muller,

K.-R., Engineering support vector machine kernels that recognize translation initiation sites,

Bioinformatics, 16: pp.799-807, 2000.

[3] Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lemmen, C., Smola, A., Lengauer, T., and Muller,

K.-R., Engineering support vector machine kernels that recognize translation initiation sites, Proc. German Conference on Bioinformatics '99, pp.37-43, 1999.

[4] Hatzigeorgiou, A.G., Translation initiation start prediction in human cDNAs with high accuracy, Bioinformatics, 18: pp.343-350, 2002.

[5] F. Zeng et al. "Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites", Gen. Inf., vol. 13, pp.192-200, 2002,

[6] Benson, D., Boguski, M., Lipman, D., and Ostell, J., Genbank, Nucleic Acids Res., 25:1 {6, 1997.

[7] Hobohm, U., Scharf, M., Schneider, R., and Sander, C., Selection of representative data sets, Prot. Sci., 1:409{417, 1992.

[8] <http://www.cbs.dtu.dk/services/NetStart>

[9] Ricardo Gutierrez-Ozuna, Lecture: Dimensionality reduction (PCA), "Introduction to Pattern Recognition", Wright State University,

http://courses.cs.tamu.edu/tgutier/cs790_w02/15.pdf.

[10] Wikipedia Article, "Dimension Reduction",

http://en.wikipedia.org/wiki/Dimension_reduction.

[11] C. Ma et al. "Feature Mining Integration for Improving the Prediction Accuracy of Translation Initiation Sites in Eukariotic mRNAs", IEEE, 2006.

[12] Y. Saeys et al. "Translation Initiation Site Prediction on a Genomic Scale: Beauty of Simplicity", Bioinformatics, 2007, vol. 23, pp. i418-i423.

[13] S. Tikole and R. Sankararamakrishnan, "Prediction of Translation Initiation Sites in Human mRNA sequences with AUG Start Codon in Weak Kozak Context: A Neural Network Approach", 2008, BBRC, pp. 1166-1168.

[14] G. Li and T. Leong, "Feature Selection for the Prediction of Translation Initiation Sites", 2005, Gen. Prot. Bioinfo., vol. 3, no. 2.

[15] J. Wegrzyn, et al. "Bioinformatic Analyses of mammalian 5'UTR sequence properties of mRNA predicts alternative translation initiation sites", 2008, BMC.

[16] F. Zeng et al. "Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites", 2002, Gen. Inf., vol. 13, pp. 192-200.

[17] G. Tzani and I. Vlahavas, "Prediction of Translation Initiation Sites Using Classifier Selection".

[18] R. Akbani and S. Kwek, "Adapting Support Vector Machines to Predict Translation Initiation Sites in the Human Genome", 2005, IEEE.

SESSION

PROTEIN CLASSIFICATION AND STRUCTURE PREDICTION, AND COMPUTATIONAL STRUCTURAL BIOLOGY

Chair(s)

TBA

A New Hybrid De Novo Sequencing Method For Protein Identification

Penghao Wang^{1*}, Albert Zomaya², Susan Wilson^{1,3}

1. Prince of Wales Clinical School, University of New South Wales, Kensington NSW 2052, Australia

2. School of Information Technologies, University of Sydney, Camperdown NSW 2006, Australia

3. Mathematical Sciences Institute, Australian National University, Canberra ACT 0020, Australia

*. Corresponding author

Email: penghao.wang@unsw.edu.au; albert.zomaya@sydney.edu.au; sue.wilson@anu.edu.au

Abstract—Tandem mass spectrometry is a powerful tool for studying proteins. However, an open problem for proteomics research is how to accurately identify proteins from the experimental mass spectra. De novo sequencing based protein identification is the only feasible approach for finding new proteins and studying protein post-translational modifications. In this paper, we describe our novel hybrid de novo sequencing based protein identification method. It differs from existing methods which rely on finding one maximum path from a spectrum graph. Instead, to identify peptides, our method applies a novel Bayesian network and dynamic programming hybrid algorithm to explore the sub-optimal space. Thus our method can better accommodate various interferences and artefacts present in the mass spectra. Evaluated on a large number of spectra, our method outperforms the most popular de novo sequencing methods and can significantly improve the accuracy of de novo sequencing based protein identification.

Keywords-Protein identification, de novo sequencing, Bayesian network, dynamic programming, proteomics.

I. INTRODUCTION

In recent years, tandem mass spectrometry (MS/MS) has become the leading technology for proteomics research [1, 2]. In a single mass spectrometry (MS) experiment, thousands of proteins from multiple complex biological samples can be identified and their expressions accurately measured at nano-mol level, thus providing a high throughput and high sensitivity approach for proteomics research. In a typical MS experiment, samples are first mixed and treated with proteolytic enzymes (*e.g.*, trypsin) to break the proteins down into shorter peptides. The peptides are then separated using High Performance Liquid Chromatography (HPLC) and injected into the mass spectrometer, where the peptides are fragmented into peptide fragments, ionised, and finally captured by the mass spectrometer. One experiment may generate thousands of MS/MS spectra, each of which theoretically corresponds to one of the proteins in the sample. However, mass spectra are usually tempered with noise and various artefacts. Thus the identification of proteins from mass spectra is a very challenging and error-prone process. Recent advances in mass spectrometry instruments and new fragmentation technologies provide unprecedented resolving power and mass accuracy in acquired spectra, which present a new opportunity to potentially identify 100% of the proteins and many more protein modifications than before [2 - 4].

However, with existing identification methods, only 50% of the proteins can be successfully identified and the protein post-translational modifications (PTM) are virtually unidentifiable [5 - 8]. Therefore, it has become a serious bottleneck for proteomics research and there is a critical need for more accurate protein identification methods that can fully utilise the resolving power of new instruments and identify more proteins and protein modifications.

Existing identification methods may be roughly classified into two categories: the database search approach and the de novo sequencing approach. The database search approach has been widely used due to its accuracy and reliability. Database search methods identify proteins by generating theoretical spectra *in silico* from a given protein database and comparing the experimental spectra with the theoretical spectra to find the best match. The main difference between database search methods lies in the type of scoring functions utilised to rank-order the most probable protein matches. One popular scoring method is exemplified by the SEQUEST algorithm [9], which applies a signal processing technique known as cross correlation to mathematically determine the overlap between the theoretical spectra and the experimental spectra to find the best match. Another important scoring method is to employ a probability model to estimate the likelihood of a match between the experimental spectrum and the theoretical spectrum being a random event. A number of methods have been proposed using such an approach, including X!Tandem [10] which uses a hyper-geometric model, OMSSA [11] which applies a Poisson model, and MASCOT [12]. It is very desirable that the probability-based database search methods provide direct measurement of the statistical confidence of an identified protein.

Despite the sophistication of database search methods, they have several limitations. Firstly, they are only effective if the proteins of interest are already known and the database used in the identification process contains the correct protein sequences. Unfortunately, for many scenarios this is difficult since many studies involve unknown proteins or proteins that have not been completely annotated [13]. Secondly, the database search methods have limited capability in detecting protein modifications. If the proteins in the samples are heavily modified, it usually leads to incorrect identifications for database search methods [14, 15]. Thirdly, specifying the enzyme used in the proteolytic digestion can also exclude the correct peptides from the search space and lead to

misidentifications [16]. The de novo sequencing approach on the other hand is able to address these issues because it identifies proteins by extracting protein sequence information directly from experimental spectra and does not require any protein database. De novo sequencing methods are the only feasible means for applications such as finding novel proteins, detecting amino acid mutations, studying the proteome at the same time as the genome, and so on. However, the main obstacle for the de novo sequencing approach is that it usually requires relatively higher quality spectra. The recent development of mass spectrometry instruments enables the measurement of high dimensional mass spectra and provides unprecedented mass accuracy, and this has removed the main obstacle for the de novo sequencing approach.

Two different de novo sequencing methods have been developed. The first method, such as Sherenga [17] and Lutfisk [18], projects the problem into graph theory and applies algorithms for finding maximum path lengths in a network topology to achieve protein identification. The second method applies probability models in inferring the proteins from the spectra, for example NovoHMM [19] and PepNovo [20]. However, the main idea of these two methods is the same: to find the longest possible peptide sequence that best suits the observed experimental spectrum. Because many peaks in the spectra corresponding to real peptide fragment ions cannot be detected in the presence of protein modifications, and ion degradation generates many intensive peaks that cannot be explained, the optimal path may not always be the correct peptide identification. Therefore, we propose a new Bayesian network and dynamic programming hybrid de novo sequencing method to infer the most likely peptide sequences by exploring the sub-optimal space. The method firstly applies a Bayesian network probability model to infer a number of most probable peptide sequences given the spectra, and then utilises a dynamic programming algorithm to find the most likely sequence. Evaluated on a large number of tandem mass spectra, our method is able to outperform the most popular de novo sequencing algorithms.

II. METHOD

A. Terminology

A peptide P which has n amino acids can be formalised as: $P = p_1 p_2 \dots p_n$. The total mass of the peptide therefore can be formalised as:

$$M = \sum_{i=1}^n m_i + 18, \quad (1)$$

where m_i is the residue amino acid mass, and 18 is the mass of H_2O . When peptides are subjected to fragmentation, a typical event is a single cleavage along the peptide's backbone. For an n amino acids peptide, there will be n possible cleavage positions, including the case that no cleavage happens. As a result, a peptide may result in a series of different ions based on the cleavage position. The N-terminal fragments (also called prefix fragments) can be denoted as: $p_1, p_2 \dots p_i$, and the C-terminal fragments (suffix

fragments) are then denoted as: $p_{i+1}, \dots p_n$. These peptide fragments will generate corresponding fragment ions with positive charges after ionisation, and a tandem mass spectrum is the collection of all detected signals of generated peptide fragment ions. N-terminal ions are called a-, b-, and c-ions, while the C-terminal ions are called x-, y-, and z-ions. If a cleavage happens at the i^{th} peptide bond, it will produce a_i, b_i, c_i ions and $x_{n-i}, y_{n-i}, z_{n-i}$ ions. An illustration of possible peptide fragmentation positions and corresponding notations for the fragment ions is given in Figure 1. The peptide fragment ions may also have neutral losses, where chemical groups such as water or ammonia (NH_3) are separated from the fragment ions.

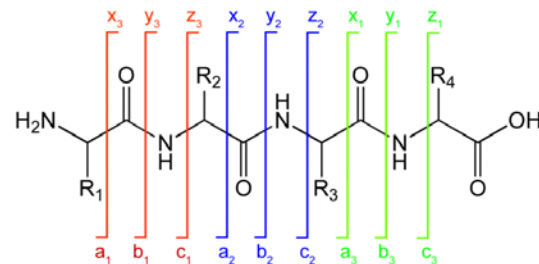


Figure 1. An illustration of a 4 amino acids peptide fragmentation pattern and notation for the fragment ions.

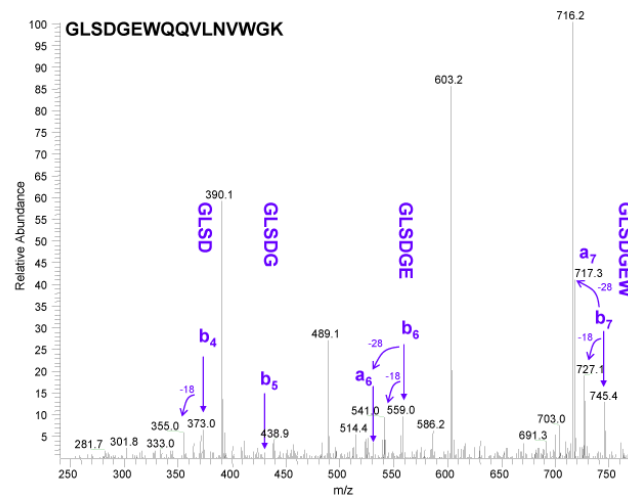


Figure 2. An example of identifying a peptide from a tandem mass spectrum using a de novo sequencing approach. The peptide precursor is singly charged and the spectrum is generated from an ion-trap mass spectrometer.

The mass spectrum of one peptide is a list of pairs of mass to charge ratio (m/z) and an associated intensity (m_1, i_1), (m_2, i_2), \dots (m_j, i_j) known as peaks, coupled with a parent (also called precursor) peptide mass M . The de novo sequencing problem is to infer the sequence of the peptide that gives rise to these peaks. Ideally each peak corresponds to one fragment ion, and the peptide sequence can be inferred from the mass difference between two adjacent peaks. An example is given in Figure 2. This is a very difficult task in reality, because spectra are very noisy and of

complex nature. In addition, different fragment ions are not detected at the same probability and many fragment ions are hardly distinguishable from the background spectrum noise. For example, the signals of b- and y-ions may be up to 5 times stronger than those of a- and x-ions; 1/5 of the b- and y-ions may suffer from neural losses; z-ions usually have very low intensities and so on [9].

B. Step 1: Spectrum Preprocessing

Our method has three major steps: (1) spectrum preprocessing, (2) Bayesian network-based identification, and (3) inferring the most likely sequence. The first step is to preprocess the spectra peaks and normalise the peak intensities prior to the main de novo sequencing algorithm. Our method adopts the peak preprocessing procedure of PepNovo [20]. The method firstly determines the baseline intensity as the average intensity of the weakest 1/3 of the peaks in the spectrum. The method divides each peak's intensity to the baseline intensity so that a normalised intensity is obtained. The normalised peak intensities are discretised into 4 levels: no signal, low signal, medium signal, and strong signal. The method then removes the low signal peaks by sliding a window of width h across the spectrum and removing all the peaks except the top k peaks. For our method, we use $h = 15$ Da and $k = 3$. The method also constrains the total number of selected peaks to be no more than 100.

Because different regions of the spectrum have different characteristics and distributions of the peaks, our method organises the peaks into 5 regions based on their m/z positions and adds this information to the Bayesian network-based model. Therefore, the correlation between the peptide fragmentation and the observed peak intensities can be better captured. For example, peaks are usually more intensive in the middle region of the spectrum because peptides are less likely to be cleaved at the positions near the two termini.

C. Step 2: Bayesian Network Identification

The second step is to infer a number of most probable peptide sequences using a Bayesian network probability model. This step involves 4 procedures.

Procedure 1: The method constructs a spectrum graph as introduced in [17]. A spectrum graph is a directed acyclic graph, whose vertices correspond to putative ions of the peptide fragmentation. Two vertices are connected by a directed edge from the vertex with a lower mass to the one with a higher mass if the mass difference between these two vertices approximates the residue mass of an amino acid or other mass offsets like ion neural losses (see Table 1 for the complete list of all considered mass offsets). Given a preprocessed mass spectrum S , we build the entire spectrum graph and connect all the edges given the peaks of S . Since the most intensive peaks in the spectrum tend to be b- and y-ions, our spectrum graph has vertices for both interpretations: given a peak at mass m_i , we create a vertex at mass $m_i - 1$ interpreting the peak as a b-ion and also a vertex at mass $M - m_i + 1$ interpreting the peak as a y-ion, where M is the sum of residue amino acid masses. A vertex for an empty peptide of mass zero and a vertex for intact peptide of

mass $M - 18$ are also added to the graph. If vertices are too close to each other (mass difference < 0.5 Da), these peaks are likely to be isotopic peaks of the same ion and are therefore merged. DiMaggio and Floudas [16] gave visualisation of a spectrum graph (also see Figure 3).

TABLE I. THE LIST OF ALL THE FRAGMENTATIONS THAT ARE MODELLED; M IS THE SUM OF THE AMINO ACID RESIDUE MASSES.

Ion Type	Notation	
	Mass offset	Terminus
b^+	$M + 1$	C-Terminus
$b^+ - H_2O$	$M - 17$	C-Terminus
$b^+ - NH_3$	$M - 16$	C-Terminus
$b^+ - 2H_2O$	$M - 35$	C-Terminus
$b^+ - NH_3 - H_2O$	$M - 34$	C-Terminus
b^{2+}	$(M + 2)/2$	C-Terminus
a^+	$M - 27$	C-Terminus
$a^+ - H_2O$	$M - 45$	C-Terminus
$a^+ - NH_3$	$M - 44$	C-Terminus
y^+	$M + 19$	N-Terminus
$y^+ - H_2O$	$M + 1$	N-Terminus
$y^+ - NH_3$	$M + 2$	N-Terminus
$y^+ - 2H_2O$	$M - 17$	N-Terminus
$y^+ - NH_3 - H_2O$	$M - 16$	N-Terminus
y^{2+}	$(M + 20)/2$	N-Terminus

Procedure 2: Our method uses a Bayesian network model to calculate the probability of observing each vertex of the constructed spectrum graph. We adopted the fragmentation model proposed in [20] which incorporates several ion degradations (given in Table 1) and 3 additional factors into the model. These 3 factors are: (1) the relationship among different types of fragment ions; (2) the correlation between peptide cleavage position and the fragmentation efficiency; and (3) the influence of the last amino acid that is adjacent to the peptide terminus. Factor 1 models the strong correlation among a-, b- and y-ions. For instance, if a b-ion is detected, it is very common that its corresponding y-ion can be detected with high intensities, and its associated a-ion is usually detected. Although all ions have correlations, only a-, b- and y-ions regularly have strong signals therefore our method focuses on these ions. Factor 2 models that ions have different probabilities of being observed depending on the cleavage positions. For example, a-ions tend to be observed more often near the N-terminus, while b- and y-ions show much higher intensities in the middle region of the spectrum, and so on. Factor 3 models the N-terminal and C-terminal amino acids' chemical effects on the peptide cleavage as reported in the literature [21, 22]. The rest of the vertices model the probabilities of observing ion degradations and ions carrying multiple charges. The whole Bayesian network

is given in Figure 4. Except for the top 3 vertices, which represent the 3 additional factors, each vertex of the network contains a conditional probability table given the values of its parent vertices. For instance, if we use the second red path in Figure 4, vertex y^+ holds the probability table $P(y^+ = t_i | b^+ = t_j, \text{region}(i) = R_k, \text{NT}(i - 1 \text{ or } i + 1) = \{\text{any AA}\}, \text{CT}(i - 1 \text{ or } i + 1) = \{\text{any AA}\})$, where t_i is the intensity of the y^+ ion, t_j is the corresponding intensity of b^+ ion, R_k is the cleavage region of the spectrum, and NT and CT are the effects of the adjacent N-terminal and C-terminal amino acids respectively.

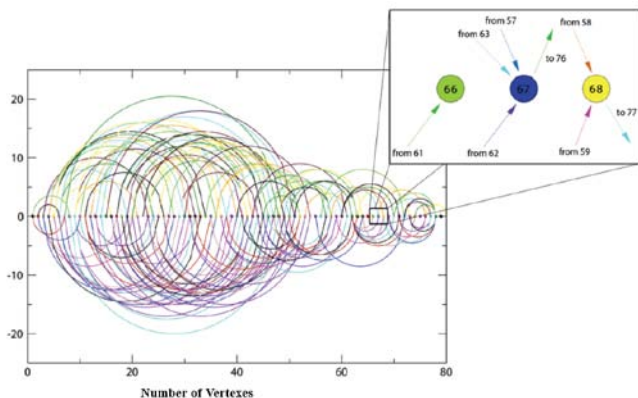


Figure 3. Visualisation of one spectrum graph, where each vertex represents one possible peptide cleavage position, and one directed edge is added if the mass difference between 2 vertices approximates the mass of an amino acid or an ion neutral loss.

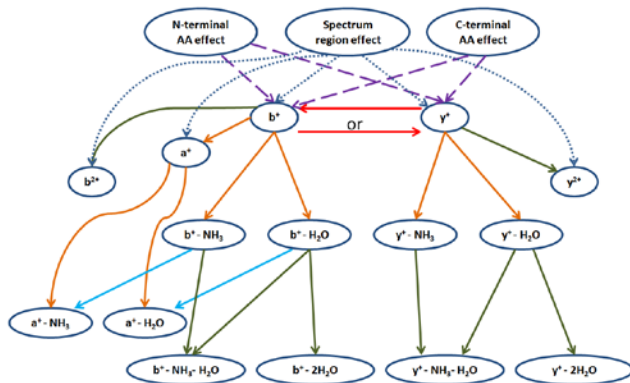


Figure 4. The Bayesian network model for our method. One of the two red paths will be randomly selected. The probabilities of the fragment ions are indicated by the colour of the solid paths: red > yellow > light blue > green. The dash paths are the additional 3 factors.

Our model differentiates itself from the one proposed in [20] in that it extends the way the probability tables are generated by incorporating singly charged, doubly charged, and triply charged tandem mass spectra from a mixture of mass spectrometry instruments. The Seattle dataset [23], which contains spectra from both ion-trap and quadrupole time of flight (TOF) mass spectrometers, was used for estimating the probability tables. More details are given in Section III. In this way, the model becomes more robust and can be applied to a much wider range of experiments.

Procedure 3: Each vertex of the constructed spectrum graph is scored using the described Bayesian network. This is achieved by comparing one hypothesis that the peak is a real fragment ion to the other hypothesis that the match is random. It is calculated by the likelihood ratio given in Equation (2):

$$O_i(m_j, S) = \log \frac{P_{\text{real}}(t | m_j, S, R, NT, CT)}{P_{\text{random}}(t | m_j, S)}, \quad (2)$$

where O_i represents the score for vertex i , m_j is the mass of the peak, S is the mass spectrum, t is the complete set of all peak intensities of S , R is the peak region, NT represents the N-terminal amino acid's chemical effect, and CT represents the C-terminal amino acid's chemical effect. Assume V is the set of the vertices in the probability network except the top 3 vertices, then $V = \{b^+, y^+, b^+ - \text{H}_2\text{O}, y^{2+}, \dots\}$. For each vertex v of V , $w(v)$ denotes v 's parents' assigned intensities given the network topology. $P_{\text{real}}(t_v = i | w(v) = \{t_1, t_2, \dots\})$ is the probability of detecting intensity i at fragment ion v given the intensities detected at its parents. Because all the conditional probability tables of the network have been obtained through training the Seattle dataset and vertex v is to be independent of the other vertices given that the values of its parents are known, the probability of observing ion fragment intensities t given that the possible cleavage occurred at mass m_j in spectrum S can be calculated by Equation (3):

$$P_{\text{real}}(t | m_j, S) = \prod_{v \in V} P_{\text{real}}(t_v | w(v), m_j, S, R, NT, CT). \quad (3)$$

One advantage of the model is that P_{real} can distinguish the likely combinations of ions and ion degradations from unlikely combinations, since the conditional probability tables are learnt from real data. For example, the probability of observing a y^+ ion and its neutral loss $y^+ - \text{NH}_3$ is higher than the probability of observing a $y^+ - \text{NH}_3$ ion without detecting the y^+ ion itself.

Under the hypothesis that the mass matches are random events, each peak is therefore considered to be independent. The probability of $P_{\text{random}}(t | m_j, S)$ can be easily calculated as the product of the probability of observing individual peaks at their mass positions. Once we have both P_{real} and P_{random} , the score for each vertex can be calculated.

Procedure 4: Given the spectrum graph and the score for each vertex, the method then finds several highest scoring asymmetric paths as the most probable peptide sequences. It is important to preserve the asymmetry because each peak from the spectrum contributes to two vertices in the constructed spectrum graph since we model both b- and y-ions for each peak. Dynamic programming is able to solve this problem and finds the highest scoring maximum path that goes through every pair of vertices corresponding to the same peak at most once. However, it has been shown that the maximum path may not be the best solution [24, 25]. There are two reasons: (1) a certain number of vertices on the optimal paths may be false positives because many high intensive peaks in the spectrum are signals from various

interferences, including protein modifications, unexpected peptide internal fragments, contaminations, *etc*; and (2) several vertices representing the real peptide fragment ions may not have the highest score so will not be included in the optimal path. It is common that real fragment ions have low intensive signals or even cannot be detected at all. Therefore, we utilise the algorithm proposed by Lu and Chen [25] to obtain a set of most probable peptide sequences by exploring the sub-optimal solutions from the spectrum graph. The algorithm firstly transforms the spectrum graph into a matrix and uses an iterative depth-first search algorithm to find the optimal path. Sub-optimal solutions are obtained by backtracking: at a certain iteration if a path showing close enough score to the optimal path, a sub-optimal path is then spawned and continued. Details of this algorithm can be found in [25].

D. Step 3: Inferring The Most Likely Sequence

The third step is to infer the most likely peptide sequence given the optimal sequence and a set of sub-optimal sequences. This set of peptide sequences has two main characteristics: (1) the majority of these sequences will have identical or highly similar segments of sequences; and (2) certain regions or sites may have ambiguities and show conflicting sequences. An example is given in Figure 5. The highly similar segments of sequences correspond to the high intensity fragment ions that are very likely to be correctly identified, while the ambiguous segments are where the peaks do not match fragment ions well or the intensities of the ions are hardly distinguishable from baseline noise. In addition, these sub-optimal solutions may have different numbers of amino acids.

Figure 5. A set of sub-optimal peptide sequences generated in Step 3. The red regions are the highly likely regions; the yellow region is the borderline region; the green regions are ambiguous regions.

Given these characteristics, the most likely peptide sequence can be extracted by adapting a dynamic programming-based algorithm similar to ClustalW [26] which has been used in multiple sequence alignment. In our case, the introduced “gaps” between the sub-optimal peptide sequences correspond to the ambiguous sections of the tandem mass spectrum. Our algorithm employs a progressive design and has 4 procedures in total.

Procedure 1: The pairwise distances of the sub-optimal peptide sequences are calculated using the Smith-Waterman dynamic programming algorithm [27]. An n by n distance matrix is then constructed from the pairwise distances, where n is the number of sub-optimal peptide sequences.

Procedure 2: A relationship for the sub-optimal peptide sequences is obtained given the distance matrix. The relationship is represented as a binary tree topology, and is constructed by applying the Neighbour Joining algorithm [28]. This algorithm is guaranteed to find the relationship topology that has the minimum overall distance.

Procedure 3: The peptide sequences are progressively aligned following the branching order of the constructed binary tree representing the relationship. The alignment proceeds from the tips of the relationship tree toward the root. In this way, the closest peptide sequences are aligned first, while the order of the most distant peptide sequences to be aligned is delayed.

Procedure 4: The final peptide sequence is obtained by identifying the highly likely segments of peptide sequences. Our method considers the regions highly likely if 85% or more of the peptide sequences agree on them. The most frequently appearing sequences will be used for these segments. The segments that are agreed by more than 55% (and less than 85%) of the sequences will be classified as borderline segments. Each amino acid in borderline segments will be determined based on its frequency across all the sub-optimal sequences. For example, if the frequencies for Glycine, Serine and Valine are 68%, 23% and 9% respectively at one site, then the algorithm will select one of these amino acids using the same probabilities as their frequencies. On the other hand, the “gaps” are interpreted as ambiguous sequence segments, which are denoted as undetermined “X” in the final identified peptide sequence.

III. RESULTS

A. Evaluation Strategy

As mentioned, we used the Seattle dataset [23] to learn the Bayesian network conditional probability tables. The Seattle dataset is a collection of reference mass spectra of 18 commercial purified proteins generated by several mass spectrometers. We selected singly charged, doubly charged, and triply charged spectra to learn the conditional probability tables. We ignored all the quadruply charged spectra because they are less common and usually of poor quality. We also excluded all the spectra generated by the MALDI TOF mass spectrometers, because spectra acquired from these machines have low resolution.

We compare the performance of our method with the most popular PepNovo and NovoHMM de novo sequencing methods by the criterion of identification accuracy. The identification accuracy is defined as the ratio of the number of correct amino acids to the number of identified amino acids. We use one large publicly available dataset to evaluate these 3 methods. The dataset is a collection of MS and MS/MS spectra of a mixture of 9 commercial purified proteins, generated by the Thermo Electron LTQ quadrupole linear ion-trap mass spectrometer. There are 3 technical replicas for this dataset, and in total the dataset contains 58,081 tandem mass spectra.

B. Evaluation Results

Our evaluation results are presented in Figure 6. Trypsin digestion was specified for running all 3 methods. Two amino acid pairs (Q and K), (I and L) are considered identical, since they have identical monoisotopic masses. Identification of either of these amino acids is considered correct. For example, if the peptide sequence is QFIER, the identifications such as QFLER and KFIER are all considered to be correct. PepNovo and NovoHMM were executed at default parameters. For our method, we used error tolerance of 0.1 Da and the maximum number of sub-optimal solutions that are explored to generate the final result was set to 20, which seemed to produce the best results.

PepNovo and NovoHMM seem to have similar overall performance in terms of identification accuracy. However, NovoHMM tends to have slightly higher accuracy in identifying short length peptide sequences. As shown in Figure 6, NovoHMM outperforms PepNovo at sequence lengths from 3 to 6 amino acids; while PepNovo starts to display better accuracy than NovoHMM for sequence length of 7 and onward. This may be due to NovoHMM's Hidden Markov model beginning to overfit when the spectra are more complicated. In any case, the performance difference between these two methods is quite small.

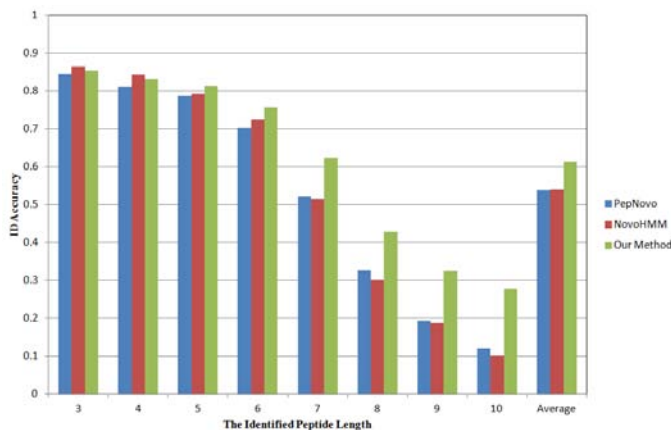


Figure 6. The comparison of identification accuracy. The x-axis is the identified peptide length in number of amino acids, the y-axis is the accuracy. The blue bar is PepNovo, the red is NovoHMM and the green is our method. The last 3 bars at the right end of the graph are the average accuracy across all peptide lengths.

Our method, compared to PepNovo and NovoHMM, has significantly better performance. It can be clearly seen from Figure 6 that our method on average achieved around 10% higher accuracy than PepNovo and NovoHMM. It is very promising that our method has much better accuracy in identifying peptide sequences of more than 7 amino acids. This is important because the majority of the tryptic peptides have 7-13 amino acids. Our evaluation results also indicate that our method has increasingly higher accuracy for longer peptides. Figure 6 shows that the accuracy improvement of our method at length 5 is minor, then it keeps increasing, and becomes almost doubled at peptide lengths of 9 and 10. This is probably because our method is not constrained to the

maximum path and takes advantage of sub-optimal solutions. When peptides have more amino acids, the number of observed fragment ions may grow very quickly. Therefore, the likelihood that the optimal path is the correct peptide sequence becomes smaller and smaller. The results demonstrate that the exploration of sub-optimal space can significantly improve the identification accuracy.

IV. DISCUSSION AND FUTURE WORK

De novo sequencing based protein identification methods are commonly considered by the community as inferior to database search methods. This might be true for older MS instruments but is not the case anymore. Database search methods may be the first choice for low resolution spectra generated by older instruments; however database search methods render useless the resolving power of the new instruments. The identification coverage of database search methods simply cannot be significantly improved by using high resolution spectra. This is due to their reliance on protein databases, which are seldom complete. The de novo sequencing approach on the other hand is able to make better use of the high resolution spectra from new instruments and does not suffer from the issues of the database search approach. From our experiments, de novo sequencing methods are able to outperform typical database search methods on high resolution Orbitrap spectra data (results not shown). Therefore, the applicability of the de novo sequencing approach should be reconsidered and more research effort should be devoted to the development of new de novo sequencing methods.

Due to the complicated nature of mass spectra, not only the optimal solution but also the sub-optimal solutions should be utilised in order to improve the identification accuracy. Several de novo sequencing methods have been developed, all of which apply sophisticated algorithms. However, the central dogma of these methods remains the same: to find the maximum path in a spectrum graph under a specific model. Unfortunately, the optimal solution may not always be the correct identification. There are several explanations. Firstly, a large portion of highly intensive peaks in the spectra are not the expected signals from peptide fragment ions. This may be due to various reasons, such as peptide internal fragmentation, peptide post-translational modifications, contamination, chemical reactions, isotopic interferences, machine error, and many others. Secondly, many fragment ions are difficult to detect and usually have low intensities, for example c- and z-ions are barely distinguishable from noise. It is possible that even the dominant b- and y-ions are partially missing from the spectra. In any case, the fragmentation patterns still are not fully understood today. Therefore, the sub-optimal solutions are of great interest. The performance of our method clearly demonstrates that in a large number of cases the correct peptide sequences are not the optimal solutions, but can be obtained by exploring the top ranking sub-optimal solutions. This creates a new research direction and it would be very desirable to develop more efficient algorithms for exploring the sub-optimal space for accurate peptide identification.

The de novo sequencing approach has great potential for identifying protein modifications. One major advantage of our method is its ability to find the regions where the spectrum is difficult to explain. Many identified ambiguous regions turn out to be the locations where modifications tend to occur, especially phosphorylation. This is very interesting since phosphorylation is one of the most important protein modifications. It has been shown to activate or deactivate many protein enzymes and play key roles in cellular processes. This also indicates that protein modification is one important factor that greatly influences the accuracy of the de novo sequencing based identification. Although the identification of protein modifications is not the central concern of de novo sequencing, it remains the most effective approach because it infers the actual peptide sequences directly from the spectra rather than matching a database. If the de novo sequencing method has an efficient protein modification model, multiple protein modifications can be identified accurately by exploring the sub-optimal space. Our method may be easily extended for this purpose by incorporating further consideration of protein modifications into the Bayesian network, and this would be an interesting direction for future research.

ACKNOWLEDGMENT

This research was funded by Australian National Health and Medical Research Council (NHMRC) grant 525453. We adapted source codes from PepNovo for constructing the spectrum graph and probability model. We adapted source codes from ClustalW for generating the final peptide sequence.

REFERENCES

- [1] Zhang Q., Faca V., and Hanash S., "Mining the plasma proteome for disease applications across seven logs of protein abundance", *J. Proteome Res.*, vol. 10, pp. 46-50, 2011.
- [2] Rinner O., et al. "Identification of cross-linked peptides from large sequence databases", *Nat. Methods*, vol. 5, pp. 315-318, 2008.
- [3] Tran J.C., et al. "Mapping intact protein isoforms in discovery mode using top-down proteomics", *Nature*, vol. 480, pp. 254-258, 2011.
- [4] Durbin K.R., et al. "Intact mass detection interpretation, and visualization to automate top-down proteomics on a large scale", *Proteomics*, vol. 10, pp. 3589-3597, 2010.
- [5] Spirin V., et al. "Assigning spectrum-specific P-values to protein identifications by mass spectrometry", *Bioinformatics*, vol. 27, pp. 112801134, 2011.
- [6] Deutsh E.W., Lam H., and Aebersold R., "Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics", *Physiol. Genomics*, vol. 14, pp. 18-25, 2008.
- [7] Polaco, B.J., et al. "Discovering mercury protein modifications in whole proteomes using natural isotope distributions observed in liquid chromatography-tandem mass spectrometry", *Mol. Cell Proteomics*, vol. 10, Epub 2011 Apr 30, 2011.
- [8] Nesvizhskii A.I. and Aebersold R., "Analysis and validation of proteomic data generated by tandem mass spectrometry", *Nat. Method*, vol. 4, pp. 787-797, 2007.
- [9] Eng J.K., McCormack, A.L., and Yates J.R. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database", *J. Am. Soc. Mass Spectrom.*, vol. 5, pp. 976-989, 1994.
- [10] Craig R., Beavis R.C. "TANDEM: matching proteins with tandem mass spectra", *Bioinformatics*, vol. 20, pp. 1466-1467, 2004.
- [11] Geer L.Y., et al. "Open mass spectrometry search algorithm", *J. Proteome Res.*, vol. 3, pp. 958-964, 2004.
- [12] Perkins D.N., Pappin D.J.C., Creasy D.M., and Cottrell J.S. "Probability-based protein identification by searching sequence databases using mass spectrometry data", *Electrophoresis*, vol. 20, pp. 3551-3567, 1999.
- [13] Lu B., and Chen T. "Algorithms for de novo sequencing using tandem mass spectrometry", *Biosilico*, vol 2, pp. 2, 2004.
- [14] Searle B.C., et al. "Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm", *J. Proteome Res.*, vol. 4, pp. 546-554, 2005.
- [15] Bandeira N., Tsur D., Frank A., and Pevzner P.A. "Protein identification by spectral networks analysis", *Proc. Natl. Acad. Sci.*, vol. 10, pp. 6140-6145, 2007.
- [16] DiMaggio P.A., and Floudas, C.A. "De novo peptide identification via tandem mass spectrometry and integer linear optimisation", *Anal. Chem.* vol. 79, pp. 1433-1446, 2007.
- [17] Dancik, V., et al. "De novo peptide sequencing via tandem mass spectrometry", *J. Comput. Biol.*, vol. 6, pp. 327-342, 1999.
- [18] Taylor J. A., and Johnson, R.S. "Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry", *Anal. Chem.*, vol. 73, pp. 2594-2604, 2001.
- [19] Fischer B., et al. "NovoHMM: a hidden Markov model for de novo peptide sequencing", *Anal. Chem.*, vol. 77, pp. 7265-7273, 2005.
- [20] Frank A. and Pevzner P. "PepNovo: de novo peptide sequencing via probabilistic network modelling", *Anal. Chem.*, vol. 77, pp. 964:973, 2005.
- [21] Tabb D.L., Smith L.L., Breci L.A., Wysocki V.H., Lin D., and Yates J.R. "GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model", *Anal. Chem.*, vol. 75, pp. 1155-1163, 2003.
- [22] Breci L.A., Tabb D.L., Yates J.R., and Wysocki V.H., "Cleavage N-terminal to Proline: analysis of a database of peptide tandem mass spectra", *Anal. Chem.*, vol. 75, pp. 1963-1971, 2003.
- [23] Limek J., et al. "The standard protein mix database : a diverse dataset to assist in the production of improved peptide and protein identification software tools", *J. Proteome Res.*, vol. 7, pp. 96-103, 2008.
- [24] Chen T., Kao M.Y. "A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry", *J. Comput. Biol.*, vol. 8, pp. 325-337, 2001.
- [25] Lu B. and Chen T.J. "A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry", *J. Comput. Biol.*, vol. 10, pp. 1-12, 2003.
- [26] Thompson J.D., Higgins D.G., and Gibson T.J., "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", *Nucleic Acids Res.*, vol. 22, pp. 4673-4680, 1994.
- [27] Smith T.F. and Waterman M.S., "Identification of common molecular subsequences", *J. Mol. Biol.*, vol. 25, pp. 195-197, 1981.
- [28] Saitou N. and Nei M., "The neighbor-joining method: a new method for reconstructing phylogenetic trees", *Mol. Biol. Evol.*, vol. 4, pp. 406-425, 1987.
- [29] Thompson J.D., Higgins D.G., and Gibson T.J., "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", *Nucleic Acids Res.*, vol. 22, pp. 4673-4680, 1994.

Protein Localization by Integrating Multiple Protein Correlation Networks

A. M. Mondal^{1,2} and J. Hu^{2*}

¹Mathematics and Computer Science, Claflin University, Orangeburg, SC, USA

²Computer Science and Engineering, University of South Carolina, Columbia, SC, USA

Abstract - We explored how integration of different protein-protein correlation (PPC) networks improves the performance of a network based classifier, NetLoc, in predicting protein subcellular localization. We investigated different integration approaches such as integration with or without changing the scope of the base network and evaluated NetLoc performance using the resulting network. Results showed that integration of different PPC networks improves NetLoc performance significantly depending on the base networks for integration and the integration approaches. This significant improvement is due to the increase in connectivity in the resulting network and contribution of positive signals imported with co-localized interactions from other networks.

Keywords: protein localization; protein-protein correlation network; network integration; diffusion kernel; PPI network.

1 Introduction

Literature shows that integrating multiple evidences can greatly improve the prediction accuracy [1-3] of classifiers for predicting protein localization. Wolf-PSort [1] achieved competitive results by combining features from PSORT, iPSort, amino acid content, and sequence length. Drawid and Gerstein [2] proposed a naïve Bayesian classifier to integrate features including motifs, sequence properties, and whole-genome gene expression features. Recently, Scott et al. [3] proposed a two-level Bayesian network approach to integrate information from InterPro motifs, targeting signals, and protein interacting partner relationships.

In our previous work [4], we proposed a network based approach for protein localization prediction. We showed that different protein-protein correlation networks such as physical protein-protein interaction (PPPI), genetic PPI (GPPI), mixed PPI (MPPI), and co-expressed PPI (COEXP) carry different levels of localization information and the performance of the proposed algorithm, NetLoc [4], depends on the topological characteristics such as connectivity and percentages of co-localized PPIs in the network [5]. Figure 1 presents the distribution of PPIs among 4 different networks: PPPI(P), GPPI(G), MPPI(M),

and COEXP70(C). Most of the PPIs of each network are not shared by other networks. For example, in PPPI network 43363 out of 50997 PPIs are not shared by other three networks. Similarly, GPPI has 103631 PPIs and 95120 PPIs are not shared by other three networks. So, integration of different networks would change the topological characteristics of the resulting network and may improve the prediction performance.

In the present study, we developed a PPC network based integration framework for protein localization prediction. This method is inspired by the successful application of network integration methods in protein/gene function prediction [6]. Integration of different networks may or may not change the scope of the resulting network from the original networks depending on the integration approach. The scope of a network in the present context is concerned with either the number of proteins in the network or the number of annotated proteins in the network or the number of PPIs in the network. Our objective is to find a unified network, with maximum scope in terms of network proteins and annotated proteins for a species by combining all available networks, which could be used as the standard network for protein localization prediction for that species.

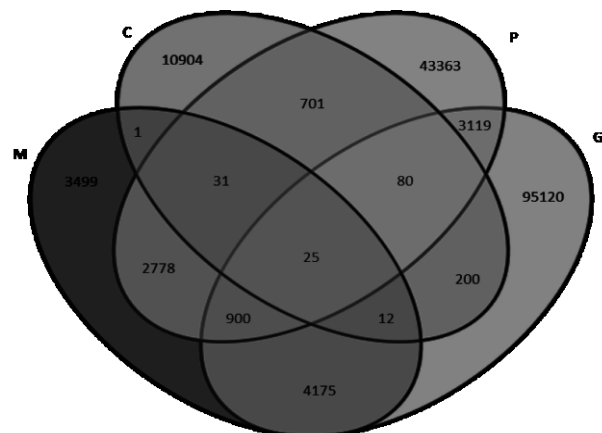


Figure 1. Distribution of PPIs in different networks.

2 Unified Network for a Species

Different kinds of PPI networks exist for a species and they provide different level of information for protein localization as mentioned earlier. One reasonable question to ask is how to come up with a unified network for network based classifiers such as NetLoc, which can be then used as the standard network for predicting protein subcellular localization for that species. Before enumerating the properties of the unified network, we define the following terms:

Co-Localized PPIs (coPPIs): PPIs for which both proteins are localized at the same location.

Non-co-localized PPIs (ncPPIs): PPIs for which two proteins are localized at two different locations.

Signal to Noise Ratio (SNR): Ratio of coPPIs to ncPPIs.

Density of coPPI (DCOP): Number of coPPI per annotated protein.

Based on the results presented in [5, 7], criteria for a unified network for protein localization are: i) the network should have high values of SNR and DCOP [7]; ii) the network should have large connected components [5]; iii) the network should have maximum possible scope with respect to the number of proteins and the number of annotated proteins i.e., network with most of the proteins in a genome; and iv) the network should have more coPPIs. The more is the coPPI the better is the network [7].

Answers to the following questions would help in finding the unified network for a species. Question-1: which type of PPIs carries more information about protein localization? Question-2: does removing some PPIs from any network improve the performance? Question-3: how does the integration approach affect the performance? Question-4: which approach should we use to integrate different networks?

3 Data and Methods

3.1 Datasets

We conducted experiments on data sets for *Saccharomyces cerevisiae* used by Mondal and Hu [4, 5, 7]. Two networks, physical PPI (PPPI) network and genetic PPI (GPPI) network, are obtained from BioGRID [8], mixed PPI (MPPI) network is from MIPS [9] and the co-expression (COEXP) network is from gene expression data of Stanford University [10]. PPPI contains only physical interactions whereas MPPI contains both physical and genetic interactions. MPPI has much less interactions since it has not been updated since 2006.

The localization data of Huh et al. [11] was used as the basis for annotation. The experiment was carried out using high-resolution localization (22 locations) for networks COEXP70, GPPI, MPPI and PPPI. Table 1 shows the summary of the four network datasets used in this study. In terms of the number of interactions, GPPI is the largest

network followed by PPPI, COEXP70 and MPPI. Considering the number of proteins, PPPI is the largest network followed by GPPI, MPPI and COEXP70. GPPI is the densest graph, meaning it has the highest values in terms of the average degree of nodes, followed by PPPI, COEXP70 and MPPI. The PPPI network has the largest number of proteins with annotated localization followed by GPPI, MPPI, and COEXP70.

TABLE 1. PPC Networks and Annotation

Property	COEXP70	GPPI	MPPI	PPPI
Number of PPIs	11954	103631	11421	50997
Number of Proteins	2004	5252	4319	5477
Average Degree of Nodes	11.92	39.46	5.28	18.62
Number of Annotated Proteins	1479	3732	3026	3803
Localization	1961	4947	4049	5039

3.2 Integration approaches

3.2.1 Integration without changing the scope of the base network

In this approach, a network is selected as the base network. Interactions from other networks that fit into the base network are imported to the base network. This integration does not change the scope of the base network in terms of network proteins and annotated proteins. The only changes are the number of PPIs or edges in the integrated network. This integration can be carried out in two different methods. In the first method, all types of PPIs from other networks that fit into the base network are imported and in the second method, only the coPPIs from other networks that fit into the base network are imported. In the second method we are avoiding importing noises or ncPPIs to maintain lower level of noise in the integrated network. For subsequent discussion, the scope of the integrated network in first method is called scope-1 and in second method it is called scope-2.

Table 2 summarizes the network structures before and after integration without changing the scope of the base network in terms of network proteins and annotated proteins. For example, for integration considering MPPI as the base network, the number of network proteins (4319) and annotated proteins (3026) in the resulting integrated network remains the same as the base network. Integration using scope-1 produces a network with 119965 PPIs and scope-2 produces a network with 49066 PPIs. It is clear that integrated network with scope-2 is more connected (more edges or PPIs) than the base network (49066 > 11421) and network with scope-1 is more connected than network with scope-2 (119965 > 49066) as expected.

TABLE 2. Networks upon integration without changing the scope of the base network

Networks	Proteins		PPIs		
	Network	Annotated	Base	Scope-1	Scope-2
COEXP70	2004	1479	11954	34688	20468
GPPI	5252	3732	103631	157423	124839
MPPI	4319	3026	11421	119965	49066
PPPI	5477	3803	50997	158983	83457

3.2.2 Integration with changing the scope of the base network

The resulting network upon union of two or more networks would have different scopes than the original networks in terms of network proteins and annotated proteins. In general, union of two or more networks would broaden the scope by increasing the numbers of both network proteins and annotated proteins. Two different methods are employed to integrate the networks in union approach. In the first method, the resulting network proteins are union of four original networks and the resulting annotated proteins are union of annotated proteins of four original networks. In the second method, a network is considered as the base and only the coPPIs from other networks are imported where coPPIs are determined based on the resulting annotated proteins found in the first method. In the second method we are avoiding importing noises or ncPPIs to maintain lower level of noise in the integrated network. For subsequent discussion, scope in the first union method is called scope-3 and that in the second is called scope-4.

TABLE 3. Networks upon integration with changing the scope of the base network.

Networks	Network Proteins			Annotated Proteins			PPIs		
	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4
COEXP70	2004	6079	4296	1479	3899	3771	11954	164908	62203
GPPI	5252	6079	5389	3732	3899	3869	103631	164908	126807
MPPI	4319	6079	5132	3026	3899	3839	11421	164908	62375
PPPI	5477	6079	5544	3803	3899	3870	50997	164908	84057

Table 3 summarizes the network structures before and after integration with changing the scope of the base network in terms of network proteins and annotated proteins. By definition, integrated networks in scope-3 have only one value for each network for each of the network attributes such as network proteins (= 6079), annotated proteins (= 3899), and number of PPIs (=164908). For completeness, the same value is shown for each of the base networks. Integration using both scope-3 and scope-4 increases the scope in terms of network proteins and annotated proteins but the increase is less in scope-4. For

example, for COEXP70, network proteins increase from 2004 to 4296 in scope-4 and 2004 to 6079 in scope-3. Similarly, annotated proteins increase from 1479 to 3771 in scope-4 and 1479 to 3899 in scope-3. Scope-4 produces integrated networks of different sizes ranging from 62203 PPIs for COEXP70 to 126807 PPIs for GPPI. In general, integrated networks are more connected (more PPIs or edges) than the base network.

3.3 Classification algorithm

We applied the diffusion kernel-based logistic regression (KLR) model [12] as used in [4, 5, 7] to predict protein subcellular localization. The KLR model based subcellular prediction problem can be formulated as in [12]. Given a protein-protein interaction network with N proteins X_1, \dots, X_N with n of them X_1, \dots, X_n with unknown subcellular locations, the task is to assign subcellular location labels to the n unknown proteins based on the location labels of known proteins and the protein-protein interaction network.

Let $X_{[-i]} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N)$,

$$M_0(i) = \sum_{j \neq i, x_j \text{ known}} K(i, j) I\{x_j = 0\}$$

$$\text{And } M_1(i) = \sum_{j \neq i, x_j \text{ known}} K(i, j) I\{x_j = 1\},$$

where $K(i, j)$ is the kernel function for calculating the similarity distances between two proteins in the network. $I(x_j = 0)$ is an indicator which indicates the interacting protein j does not have the location of interest and $I(x_j = 1)$ indicates that protein j does have the location of interest. Diffusion kernel K , to represent the interaction network, is defined using the following equation.

$$K = e^{\{\tau L\}}$$

Where

$$L(i, j) = \begin{cases} 1 & \text{if protein } i \text{ interacts with protein } j \\ -d_i & \text{if protein } i \text{ is the same as protein } j \\ 0 & \text{otherwise} \end{cases}$$

Where d_i is the number of interaction partners of protein i , τ is the diffusion constant, and $e^{\{L\}}$ represents the matrix exponential of the Laplacian matrix L . Then the KLR model is given by:

$$\log \frac{\Pr(X_i = 1 | X_{[-i]}, \theta)}{1 - \Pr(X_i = 1 | X_{[-i]}, \theta)} = \gamma + \delta M_0(i) + \eta M_1(i)$$

which means that the logit of $\Pr(X_i = 1 | X_{[-i]}, \theta)$, the probability of a protein targeting a location L is linear based on the summed distances of proteins targeting to L or other location. We then have:

$$\Pr(X_i = 1 | X_{[-i]}, \theta) = \frac{1}{1 + e^{-(\gamma + \delta M_0(i) + \eta M_1(i))}}$$

The parameters γ , δ , and η can be estimated using the maximum likelihood estimation (MLE) method. Note that here only the annotated proteins are used in the estimation procedure.

Fig. 2 presents the schematic overview of the network-based framework for protein localization prediction using the KLR model by integrating different PPC networks. First, an integrated network is obtained by combining different PPC networks using one of the four scopes. Then diffusion kernel type feature, which is a square matrix consisting of 1 (interaction) and 0 (no interaction), is developed for the integrated network.

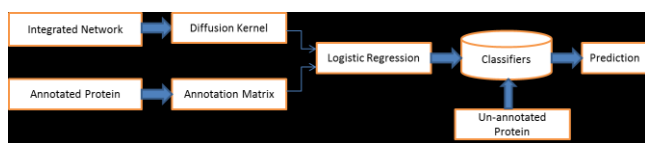


Figure 2. Protein localization prediction using the KLR model by integrating PPC networks.

Annotation matrix, which is an m by n matrix, consists of 1 (annotated) and 0 (not annotated), where m is the number of annotated proteins and n is the number of localizations, is developed from annotated proteins. KLR model is developed using kernel type features and annotation matrix using logistic regression. The KLR model produces confidences for each protein for all locations. Then a threshold on confidences is used to classify the proteins to be localized at a location or not.

4 Results and discussion

4.1 Quality of PPI

To answer question-1, we need to find the quality of each type of PPI network, which depends on how much information is carried out by that type in predicting protein localization. There are three fundamental types of PPIs used in the present study – physical PPI, genetic PPI, and co-expressed PPI. In order to determine the quality of different types of PPI, we need to fix the scope of networks with respect to i) number of network proteins (same number of same proteins), ii) number of annotated proteins (same number of same proteins) and iii) number of PPIs (same number of PPIs but different types). Table 4 shows the common proteins among three fundamental networks and the corresponding PPIs in different networks. It is clear that there are 1710 network proteins and 1390 annotated proteins which are common among three fundamental networks but they have different number of PPIs (COEXP70:9007, GPPI:12369, PPPI:10136). Now, NetLoc performance is determined by selecting a fixed number of PPIs (6000, 7000, 8000, 9000) randomly for each network. For each selection, 10 different sets of PPIs are selected

randomly and NetLoc performance i.e, AUC value, is evaluated using each set of PPIs. Then the mean and standard deviation of 10 AUC values are determined. Table 5 shows the statistics of performances for 10 experiments for each selection of PPIs. It is clear that for a specific number of PPIs, AUC for 10 experiments are very close for each network since the standard deviations are very small compared to mean values. This indicates that the performance results or AUC values produced by each selection are statistically significant.

TABLE 4. Numbers of common proteins and corresponding PPIs

Item	COEXP70	GPPI	PPPI
Original PPIs	11954	103631	50997
Original Network Proteins	2004	5252	5477
Original Annotated Proteins	1479	3732	3803
Common Network Proteins	1710	1710	1710
Common Annotated Proteins	1390	1390	1390
PPIs wrt common proteins	9007	12369	10136

TABLE 5. Statistics of 10 AUC values obtained using 10 different sets of edges for each selection

Selection Of PPIs	COEXP70		GPPI		PPPI	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
6000	0.7211	0.0072	0.6757	0.0089	0.7612	0.0085
7000	0.7295	0.0042	0.6852	0.0084	0.7696	0.0062
8000	0.7374	0.0054	0.6883	0.0077	0.7730	0.0037
9000	0.7460	0.0001	0.6960	0.0066	0.7844	0.0040

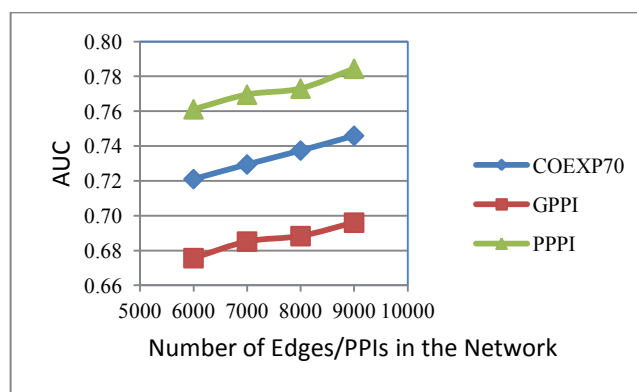


Figure 3. Contribution of PPI types in predicting protein localization.

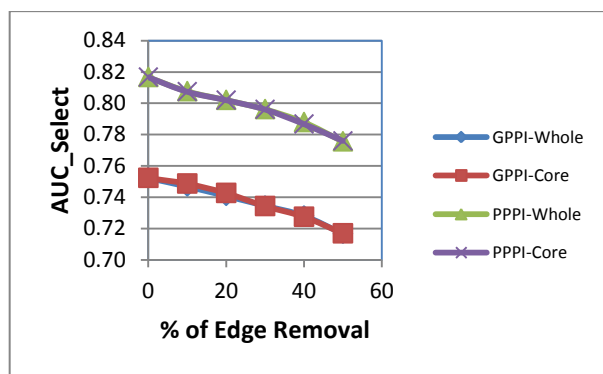
Figure 3 shows the trend of performance with different types of PPIs. It is clear that for a specific number of edges/PPIs, physical PPI produces the best performance, followed by Co-expressed PPI and then genetic PPI. For example, at edge equal to 7000, AUC values are 0.7696 for PPPI, 0.7295 for COEXP70, and 0.6852 for GPPI. This

trend increases with the increase of number of edges in the network. It can be concluded from this experiment that physical PPI has the highest contribution to predicting protein localization followed by co-expressed PPI and then by genetic PPI. So, Physical PPI network could be used as the basis for unified network.

4.2 Effect of removing some interactions

Both GPPI and PPPI are composed of only one connected component, table 4 of [4]. Any of these two networks could be a good candidate as the basis of a unified network. GPPI (network proteins = 5252, annotated proteins = 3732, PPIs = 103631) is the densest network or it has too many PPIs. On the other hand PPPI (network proteins = 5477, annotated proteins = 3803, PPIs = 50997) has less PPIs (about 50% of GPPI) and lower annotation coverage ($69.44\% < 71.06\%$). But PPPI produces better results than GPPI (AUC: $0.82 > 0.75$), figure 2 of [4]. This suggests that for a unified network, we may not need too many interactions. Then question arises, does removing some PPIs from GPPI network improve the performance (Question-2)? Removing edges from the whole network makes some of the proteins isolated from the network, specially, proteins with single-degree of interaction. These single-degree proteins are located at the edge of the network. In order to avoid producing isolated proteins, removal is also carried out from the core of the network. The core for the present study is composed of proteins with at least degree equal to 4.

Figure 4 shows the performance after removing edges from the whole network and from the core for both GPPI and PPPI networks. It is clear that removal of edges deteriorates the performance for both networks. But there is hardly any difference in performance in two different removal approaches. This suggests that for a unified network, we should not remove any edges or PPIs from any network.



Whole: represents removal from the whole network
Core: represents removal from the core of the network

Figure 4. Effect of edge removal.

4.3 Effect of without changing the scope of the base network

Table 6 summarizes the performance of integrated networks without changing the scope of the base network considering all locations (22 locations). It is clear that NetLoc performance significantly improves upon network integration. Using scope-1, performance improvement ranges from 3% for PPPI to 28% for COEXP70 and using scope-2, it ranges from 10% for PPPI to 36% for COEXP70. Two main reasons for improvement are- (i) each network becomes more connected (more edges) upon integration (Table 2) and (ii) increase in values for either SNR or DCOP or both. In our earlier study, we showed that NetLoc performance improves with the increase of SNR and DCOP [7]. In integration using scope-1, values of SNR for some integrated networks are slightly decreased from the corresponding base network but values for DCOP are significantly increased for each of the integrated networks compare to base networks, which in turn improve the performance of integrated networks. For a specific base network, values of DCOP for integrated networks using both scope-1 and scope-2 remain the same (14.16 for PPPI) but value of SNR in scope-2 (3.869 for PPPI) is significantly higher than that in scope-1 (0.987 for PPPI). As a result, scope-2 produces better results than scope-1 in general.

TABLE 6. NetLoc performance upon integration without changing the scope of the base network

Networks	SNR			DCOP			AUC_All			Improve	
	Base	Scope-1	Scope-2	Base	Scope-1	Scope-2	Base	Scope-1	Scope-2	Scope-1	Scope-2
COEXP70	1.451	1.147	4.388	2.84	8.60	8.60	0.6407	0.8229	0.8728	28%	36%
GPPI	0.806	0.959	1.352	8.38	14.06	14.06	0.7851	0.8813	0.9086	12%	16%
MPPI	0.996	0.961	11.709	1.16	13.60	13.60	0.7132	0.8692	0.9496	22%	33%
PPPI	1.537	0.987	3.869	5.63	14.16	14.16	0.8525	0.8787	0.9401	3%	10%

4.4 Effect of changing the scope of the base network

Table 7 summarizes the performance of integrated networks with changing scope of the base network considering all locations (22 locations). It is clear that NetLoc performance also significantly improves upon network integration with changing scope of the base network. Using scope-3, performance improvement ranges from 3% for PPPI network to 37% for COEXP70 and using scope-4, it ranges from 10% for PPPI to 49% for COEXP70. As explained earlier, the improvement in the performance is due to increase either in SNR or DCOP or in both. For example, for base network COEXP70, integration using scope-3 decreases SNR from 1.451 to 0.973 but increases DCOP significantly from 2.84 to 13.97,

which in turn improves the performance from 0.6407 to 0.8809. In integration using scope-4, a significant increase happened to both SNR (from 1.451 to 18.784) and DCOP (from 2.84 to 14.44), which results in huge improvement in performance from 0.6407 to 0.9562.

TABLE 7. NetLoc performance upon integration with changing the scope of the base network

Networks	SNR			DCOP			AUC_All			Improve	
	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4	Base	Scope-3	Scope-4	Scope-3	Scope-4
COEXP70	1.451	0.973	18.784	2.84	13.97	14.44	0.6407	0.8809	0.9562	37%	49%
GPPI	0.806	0.973	1.402	8.38	13.97	14.07	0.7851	0.8809	0.9041	12%	15%
MPPI	0.996	0.973	15.497	1.16	13.97	14.18	0.7132	0.8809	0.9565	24%	34%
PPPI	1.537	0.973	3.913	5.63	13.97	14.07	0.8525	0.8809	0.9351	3%	10%

4.5 Identifying unified network

Figure 5 presents the performance of integrated networks using four different scopes compare to base network. It is clear that integration improves performance in all methods of integration. Now the question is which integrated network should we select as the unified network or which approach should we use for integration (Question-4).

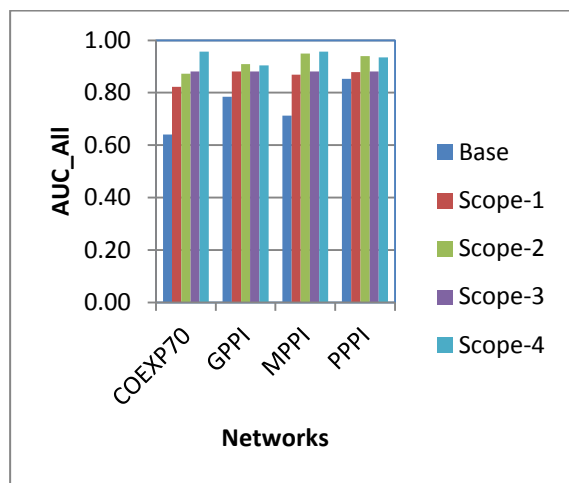


Figure 5. NetLoc performance upon integration with different scopes.

Unified Network based on Performance

Integration using Scope-2 produces better performance than scope-1 for all networks since scope-2 comes with better signals (relatively more co-localized PPIs) than scope-1. Similarly, scope-4 produces better performance than scope-3 for all networks. Considering performance, integrated networks using scope-2 and scope-4 are possible candidates for unified network. Out of 8 integrated

networks, integration using scope-4 with base network MPPI produces the best performance of AUC = 0.9565 (Figure 5). So, integrated network obtained from MPPI network using scope-4 can be considered as the unified network.

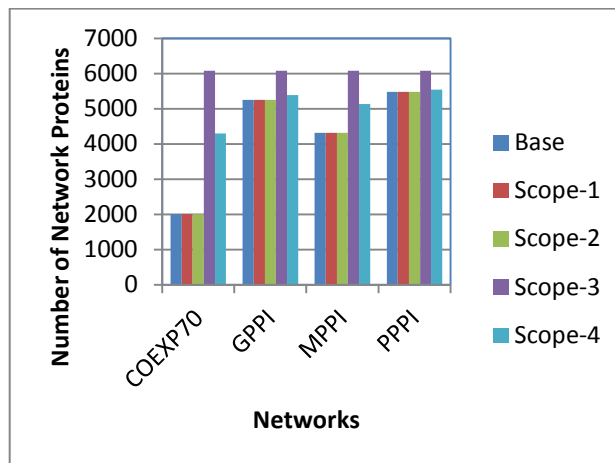


Figure 6. Network proteins upon integration with different scopes.

Unified Network based on Scope

Integration using scope-1 and scope-2 has the minimum scope, which is the same as base network, in terms of both network proteins (Figure 6) and annotated proteins (Figure 7) for each of the base networks. Integration using scope-3 has the maximum scope in terms of both network proteins (Figure 6) and annotated proteins (Figure 7), which are same for each of the base network. Integration using scope-4 has the intermediate scope in terms of both network proteins (Figure 6) and annotated proteins (Figure 7) for each of the base networks. So, considering scope, integrated network using scope-3 can be used as the unified network.

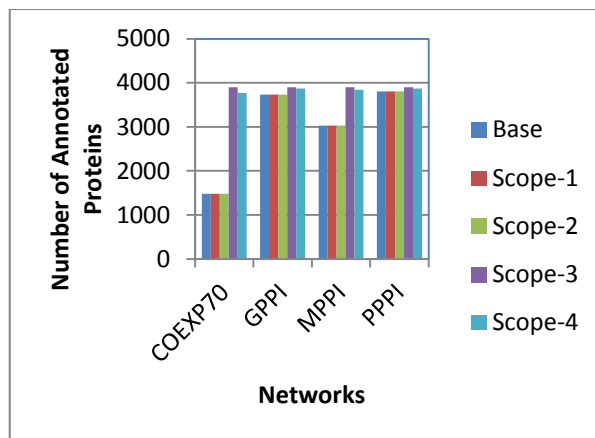


Figure 7. Annotated proteins upon integration with different scopes.

Balanced Unified Network

Unified networks based on performance and on scope represent networks based on two extremes. The first unified network produces a maximum performance of AUC = 0.9565 with a scope of 5132 network proteins and 3839 annotated proteins. The latter produces a performance of AUC = 0.8809 with the maximum scope of 6079 network proteins and 3899 annotated proteins. A balanced unified network is the one that provides a balance between performance and scope. Considering base PPPI, integration using scope-4 achieved a performance of AUC = 0.9351 with a scope of 5544 network proteins and 3870 annotated proteins. This network has both scope and performance in between the two unified networks based on two extremes. So, integrated network obtained from PPPI network using scope-4 can be considered as the balanced unified network for predicting protein localization. The overall performance (AUC = 0.9351) is improved by 10% over the individual best performance (AUC = 0.8525) with base network PPPI. This proves our hypothesis that the unified network should be based on high quality network which is physical PPI in the present study (Figure 3).

5 Conclusion

Different kinds of integration approaches such as integration with or without changing the scope of the base network are explored to observe the influence of integrating different PPC networks on the performance of a classifier, NetLoc, to predict protein localization. We use four different PPC networks for integration such as physical PPI, genetic PPI, mixed PPI, and co-expressed PPI. Our results showed that integration of different networks significantly improves NetLoc performance. The resulting network upon integration has higher number of co-localized PPI per annotated protein and/or higher signal to noise ratio, which in turn improves the NetLoc performance significantly. This study also showed that physical PPI has the highest contribution to predicting protein localization followed by co-expressed PPI and genetic PPI. Finally, we proposed a balanced unified network based on performance and scope of the integrated networks, and we found that the balanced unified network is based on a network with the best quality, which is physical PPI.

ACKNOWLEDGMENT

This work is partially supported by NSF Career Award DBI-0845381, HBCU-UP grant HRD-0713853, and Center for Excellence in Teaching of Claflin University.

References

[1] P. Horton, et al., "WoLF PSORT: protein localization predictor," *Nucleic Acids Research*, vol. 35, pp. W585-W587, 2007.

[2] A. Drawid and M. Gerstein, "A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome," *J Mol Biol*, vol. 301, pp. 1059-75, Aug 25 2000.

[3] M. S. Scott, et al., "Refining protein subcellular localization," *PLoS Comput Biol*, vol. 1, p. e66, Nov 2005.

[4] A. M. Mondal and J. Hu, "NetLoc: Network Based Protein Localization Prediction Using Protein-Protein Interaction and Co-expression Networks," in *IEEE International Conference on Bioinformatics & Biomedicine (BIBM2010)*, Hong Kong, 2010, pp. 142-148.

[5] A. M. Mondal and J. Hu, "Network Based Prediction of Protein Localization Using Diffusion Kernel," *International Journal of Data Mining and Bioinformatics* 2011.

[6] T. Hawkins, et al., "New paradigm in protein function prediction for large scale omics analysis" *Mol. Biosyst.*, vol. 4, pp. 223-231, 2008.

[7] A. M. Mondal, et al., "Network Based Subcellular Localization Prediction for Multi-Label Proteins," in *BIBM-International Workshop on Biomolecular Network Analysis (IWBNA)*, 2011.

[8] C. Stark, et al., "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res*, vol. 34, pp. D535-9, Jan 1 2006.

[9] U. Guldener, et al., "MPact: the MIPS protein interaction resource on yeast," *Nucleic Acids Res*, vol. 34, pp. D436-41, Jan 1 2006.

[10] P. T. Spellman, et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol Biol Cell*, vol. 9, pp. 3273-97, Dec 1998.

[11] W. K. Huh, et al., "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, pp. 686-91, Oct 16 2003.

[12] H. Lee, et al., "Diffusion kernel-based logistic regression models for protein function prediction," *OMICS*, vol. 10, pp. 40-55, Spring 2006.

Finding better partitions and conserved modules in Wnt signaling pathways

L. Nayak and R. K. De

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, West Bengal, India

Abstract—Human Wnt signaling pathway is involved in many crucial biological processes and its haywired behavior is found to be associated with various kinds of human cancers and other disorders. Modularized analysis will help in understanding *modus operandi* of this pathway. Here we partition the human Wnt signaling pathway into multiple partitions/modules by five algorithms inspired from different concepts. Greedy, Farhat's and Kernighan-Lin's algorithms are graph partitioning techniques. Newman's algorithm is dedicated towards finding communities in networks. Modularization algorithm detects functional modules in biological networks. A comparative study was done among partitions created by these algorithms by considering 'valid attribute' and 'functional enrichment' scores. Based on the functional enrichment score comparison, Modularization algorithm was found to create best partitions from the human Wnt signaling pathway. Later modules of 31 species-specific Wnt signaling pathways were studied and compared for detection of conserved modules.

Keywords: KEGG, Gene Ontology, Modularization algorithm, Functional enrichment score

1. Introduction

The justification for dividing a network into a number of modules lies in the fact that the complexity of each module is much less than that of the entire pathway. It provides an easier means of studying the network by part. The task is difficult, because the components of a pathway always unite their mettle towards a common function. Hence, separating them into different classes/clusters/partitions or the latest term 'modules' is difficult. The partitions obtained as a result of the separation process is expected to upgrade existing knowledge and to simplify a task. There exist several methods for creating partitions from networks, but only a few of them have been applied to biological networks like graph partitioning algorithms and community finding algorithms. Methods based on graph partitioning algorithms ([1], [2], [3]) are rigid, as they demand cut number and cut-size information. It is not possible always to provide this information.

Community finding algorithms may help in finding existing communities in undirected metabolic pathways [4], directed networks [5] as well as overlapping community structure in DIP core list of protein-protein interactions of

S. cerevisiae [6]. But, they have not been able to divide a network without existence of natural partition(s). Most of the biochemical networks come into this undividable category partially or fully. A newer flexible algorithm was required to overcome such kind of restrictions. The authors have devised an algorithm known as 'Modularization Algorithm' [7], in this regard. Modularization is a process by which one can split a network into smaller sub-networks called as modules. A module can be defined as a partition of the original network. It tends to be self-sufficient by maintaining minimal dependency on the rest part of the network. The algorithm is based on connectivity and topology of networks but does not require any cut-size or cut-number. It creates partitions from a network by using a complexity parameter 'c' [7]. It can also split a network without existence of any natural partition.

There are other existing partitioning approaches that can help towards designing an efficient modularization algorithm. A novel method to decompose biochemical networks is based on minimizing retroactivity among the created modules [8]. Retroactivity is the effect of the downstream elements on upstream elements. Another method claims to modularize biochemical networks based on classification of Petri net t-variants [9]. MODularized NETWORK learning (MONET) draws a whole network into overlapping modules and then tries to get the global picture by integrating the learned sub-networks [10]. Deterministic Modularization of Networks (dMoNet), a new agglomerative algorithm, finds even better modules in large-scale yeast and human protein interaction networks [11]. Bayesian networks and Probabilistic models are already used for identifying regulatory modules from gene expression data to identify functionally coherent modules and their correct regulators in *S. cerevisiae* [12]. Repeated random walk (RRW) based methods are used for discovering functional modules within large-scale protein interaction networks. They can find multi-functional proteins by allowing overlapping clusters [13].

Netsplitter [14] creates partitions progressively and the interactive visual matrix presentation allows considerable control over the process by the user, while incorporating special strategies to maintain the network integrity and minimize the information loss due to partitioning. Iterative Network Partition (iNP) identified modules in yeast protein complex network and breast cancer gene co-expression network [15]. Structural Clustering Algorithm for Networks

(SCAN) finds clusters or functional modules, hubs and outliers in complex biological networks [16]. Cartographic representation of networks can be used to find functional modules and uncover important new results in metabolic networks, such as the significant conservation of non-hub connector metabolites [17]. But none of them have been applied to signal transduction pathways. It will constitute an interesting work to combine these ideas to create a more robust partitioning algorithm and apply it to different kinds of pathways including that of signal transduction.

In this article, we have partitioned the human Wnt signaling pathway using various algorithms, *viz.*, Modularization algorithm [7], Newman's community finding algorithm [4], Greedy algorithm [1], Farhat's algorithm [2], and Kernighan-Lin's algorithm [3]. Their performances are compared based on 'valid attribute' and 'functional enrichment' scores in order to find the best partitioning algorithm. In addition, we have detected presence of conserved modules in 31 species-specific Wnt signaling pathways.

2. Materials and Methods

Here, we describe various partitioning algorithms. Then, we formulate a method for comparing these partitions by associating them to gene ontology terms. First of all, we describe different sources of pathway data.

2.1 Data

An exclusive list of all the signaling pathway databases is provided at <http://www.pathguide.org/>. Wnt signaling pathway data can be availed from some of these databases, *i.e.*, Reactome [18], BioCarta [19], PID [20], NetPath [21], STKE [22] and KEGG/PATHWAY [23] in various formats. No species-specific Wnt signaling pathway data is available other than hsa in PID and NetPath. Wnt data is available for hsa and mmu only in BioCarta. STKE has data for a few species (dme, dre, cel and hsa). In Reactome database Wnt signaling pathway information is available for 12 species. But, there is no option to download the molecular interactions of Wnt signaling pathway specific to each species. On the other hand, KEGG contains 48 species-specific Wnt signaling pathways (maximum number of species covered in any database at present). XML data files of the pathways along with their KGML and PNG diagrams are publicly accessible. We took 31 species-specific Wnt signaling pathways as raw data from this database (data taken in August 2009). These species-specific data are used for analysis in this work. Detailed information of these species is given in Table 1. The database uses a unique three letter code for each species along with their biological and common names (wherever applicable), *viz.*, 'hsa' for *H. sapiens* (human). These three letter codes are used extensively in this manuscript.

Table 1: Details of species taken from KEGG/PATHWAY database. For all these species, separate species-specific pathways are available in KEGG/PATHWAY database. The database uses a unique three letter code, *viz.*, 'hsa' for *H. sapiens* (human) for each species along with their biological and common names.

Sl. No.	Species Name	Common Name	KEGG code
01	<i>H. sapiens</i>	Human	hsa
02	<i>M. musculus</i>	Mouse	mmu
03	<i>R. norvegicus</i>	Rat	rno
04	<i>B. taurus</i>	Cow	bta
05	<i>C. familiaris</i>	Dog	cfa
06	<i>P. troglodytes</i>	Chimpanzee	ptr
07	<i>M. mulatta</i>	Rhesus Monkey	mcc
08	<i>M. domestica</i>	Opossum	mdo
09	<i>G. gallus</i>	Chicken	gga
10	<i>D. rerio</i>	Zebrafish	dre
11	<i>X. laevis</i>	African clawed frog	xla
12	<i>S. purpuratus</i>	Purple sea urchin	spu
13	<i>X. tropicalis</i>	Western clawed frog	xtr
14	<i>D. melanogaster</i>	Fruitfly	dme
15	<i>E. caballus</i>	Horse	ecb
16	<i>N. vectensis</i>	Sea anemone	nve
17	<i>A. mellifera</i>	Honey bee	ame
18	<i>D. pseudoobscura</i>	-	dpo
19	<i>T. castaneum</i>	Red flour beetle	tca
20	<i>A. aegypti</i>	Yellow fever mosquito	aag
21	<i>O. anatinus</i>	Platypus	oaa
22	<i>C. elegans</i>	Nematode	cel
23	<i>A. gambiae</i>	Mosquito	aga
24	<i>S. scrofa</i>	Pig	ssc
25	<i>B. floridae</i>	Florida lancelet	bfo
26	<i>C. intestinalis</i>	Sea squirt	cin
27	<i>D. ananassae</i>	-	dan
28	<i>B. malayi</i>	Filaria	bmy
29	<i>A. pisum</i>	Pea aphid	api
30	<i>T. adhaerens</i>	-	tad
31	<i>C. briggsae</i>	-	cbr

2.2 Algorithms

We have used the Biological Networks Gene Ontology tool (BINGO) [24] for comparing performance among Modularization [7], Newman's community finding [4], Greedy [1], Farhat's [2], and Kernighan-Lin's [3] algorithms. C and Matlab (Version 7.0.4) have been used for implementation of these algorithms.

2.3 Scoring Method

BINGO is an open source java tool to determine the Gene Ontology (GO) terms that are significantly over-represented in a set of genes. GO [25] is a public consortium of databases that provides a controlled vocabulary of terms aiming at a gene's or a cluster of genes' biological annotations. It consists of three hierarchically structured sets of vocabularies that describe gene products in terms of their associated 'Biological Process (BP)', 'Molecular Function (MF)' and 'Cellular Component (CC)' information; 'Go Full (GF)'

being the superset of these sets. BINGO runs as a plug-in to Cytoscape [26]. BINGO retrieves the relevant GO annotations and propagates them upward through the GO hierarchy, *i.e.*, any gene annotated to a certain GO category is also explicitly included in all parental categories. It tries to answer the basic question, “While sampling X genes (test set) out of N genes (reference set), what is the probability that x or more of these genes belong to a functional category C shared by n of the N genes in the reference set?” Hypergeometric test answers this question in the form of a P-value. P-values depict a created partition’s capability to lie in one category of biological function. If a particular partition created by a partitioning algorithm returns more number of valid GO terms with lower P-values than the others, the algorithm is believed as a better algorithm for creating partitions. Based on this belief, we have designed the ‘valid attribute score’.

Valid attribute-wise analysis takes into consideration the number of valid GO attributes that the algorithm in consideration gets as result from a query with respect to a background database. Here, we have considered GO attributes obtained with P-value of the order of 10^{-5} or smaller as valid. The threshold P-value was fixed in such a manner that valid attributes from majority of the partitions can be collected. Counting the number of valid attributes that a partition is found to be associated with, is a well established way of determining the biological validity of that partition. Many clustering algorithms follow it as a comparative measure to establish their superiority among the others [27]. Here, we have considered three background databases, namely ‘BP’, ‘CC’ and ‘GF’.

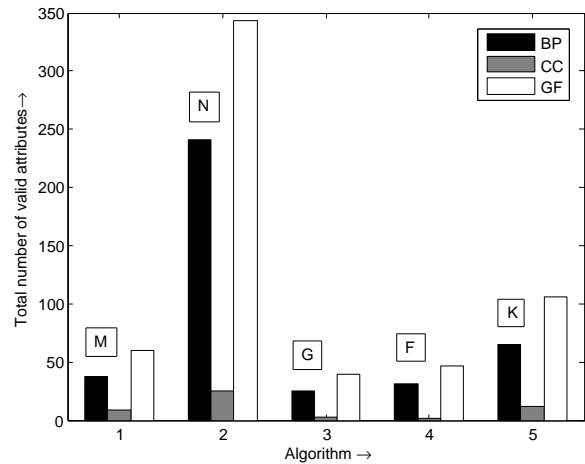
P-values give a good indication about the prominence of a certain functional category. But, no index of validity exists among the valid GO terms with lower P-values. Are they all equally valid or some of them are more valid than the others? Does such an index affect comparative results? By devising a validity index (‘functional enrichment score’), the authors have showcased the change in results. Functional enrichment score-wise analysis takes into account functional enrichment scores of a set of partitioning algorithms. The functional enrichment score S_A of an algorithm A is defined as the mean of enrichment scores of the p partitions it has created.

$$S_A = \frac{1}{p} \sum_{i=1}^p S_{P_i} \quad (1)$$

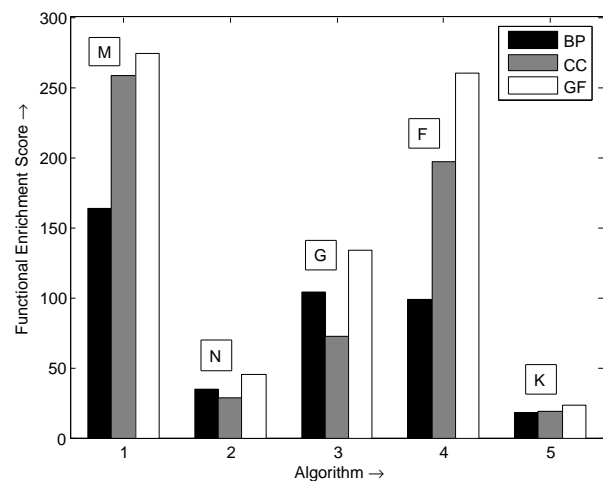
In turn the enrichment score S_{P_i} of a partition P_i is the average of the individual enrichment scores ($S_{T_{ij}}$) of associated individual attributes (T_{ij} s). Thus S_{P_i} is given by

$$S_{P_i} = \frac{1}{q} \sum_{j=1}^q S_{T_{ij}} \quad (2)$$

where q is the number of attributes. Enrichment score $S_{T_{ij}}$



(a)



(b)

Fig. 1: Performance comparison of various partitioning algorithms. (a) Comparison based on valid attribute score. Newman’s community finding algorithm is performing better. (b) Comparison based on functional enrichment score of valid attributes. Modularization algorithm is performing better. [BP- Biological Process, CC- Cellular Component, GF- GO Full, M- Modularization algorithm, N- Newman’s algorithm, G- Greedy algorithm, F- Farhat’s algorithm and K- Kernighan-Lin’s algorithm]

of an individual attribute T_{ij} is calculated by comparing the performance of algorithm A with the performance of

a background database in detecting over-expressed gene category(s) associated with the attribute. $S_{T_{ij}}$ depicts the efficiency of the partitioning algorithm in placing nodes having a common attribute in a partition with respect to a background database. Let x be the number of nodes associated with an attribute T_{ij} , which lies in a partition P_i , and $X (\geq x)$ be the number of nodes present in partition P_i . Then x/X is the ability of an algorithm for placing nodes in a partition that are associated with attribute T_{ij} . Let y be the number of nodes associated with an attribute T_{ij} in a background database, and $Y (\geq y)$ be the number of attributes in that database. Then y/Y is the ability of the background database to associate genes to attribute T_{ij} . Thus $S_{T_{ij}}$ can be defined as

$$S_{T_{ij}} = \frac{x}{X} / \frac{y}{Y} \quad (3)$$

In other words, we have taken the ratio of the performance of an algorithm with respect to the performance of a background database in assigning an attribute to a partition. Functional enrichment score is a measure to quantify the level of performance of an algorithm in creating biologically significant partitions. While comparing a few algorithms, higher the value of S_A , better is the algorithm for creating significant partitions. We have created three sets of enrichment scores, corresponding to three background databases, for each algorithm (Modularization, Newman's community finding, Greedy, Farhat's and Kernighan-Lin's) to get a better comparison.

3. Results and Discussions

Partitions of the human Wnt signaling pathway obtained by Modularization [7], Newman's community finding [4], Greedy [1], Farhat's [2], and Kernighan-Lin's [3] algorithms are described here. Human Wnt signaling pathway is a network of 60 nodes and 70 relations.

The best set of partitions created by each of the aforementioned algorithms was needed for the purpose of comparison. Hence, multiple sets of partitions were obtained by Modularization, Greedy and Farhat's algorithms where cut-number can be predesigned. Every individual set of partitions was evaluated by calculating their average functional enrichment score of associated valid attributes. The set of partitions having the highest functional enrichment score was deemed the best and used for comparison. The Modularization algorithm produced the best set of partitions (8 modules) for $c = 3$. Hence, one way of comparison was to create a set containing the same number of partitions from Greedy and Farhat's algorithm and then tally their average functional enrichment score of associated valid attributes. But, it would have been a biased way of comparison as some other set of partitions created by Greedy and Farhat's algorithm may yield a better functional enrichment score. So for Greedy and Farhat's algorithm, we have considered three immediate

lower and higher cut-numbers including the cut-number 8 for creating sets of partitions [range: 5-11]. The best set among them (11 partitions for Farhat's algorithm and 9 partitions for Greedy algorithm) was used then for comparison. Newman's community finding algorithm created the best set of partitions (8 partitions) for ΔQ value of $1.0470e^{-017}$. Two modules were generated by Kernighan-Lin's algorithm. The best sets of partitions created by all these algorithms are given in Table 2.

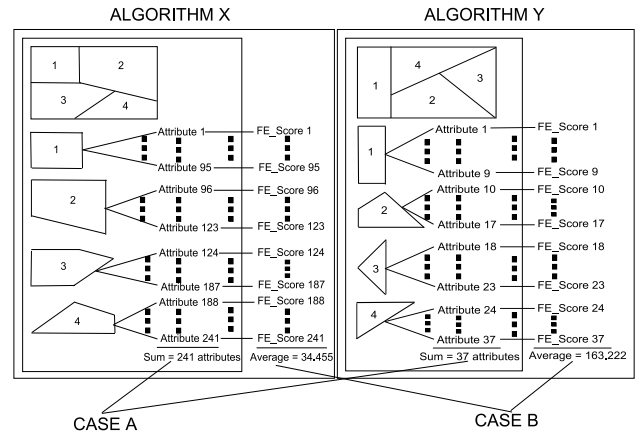


Fig. 2: **Methods of Algorithm Comparison.** (CASE A) Comparison based on valid attribute score shows that algorithm X is better than algorithm Y in creating partitions as the partitions are associated with more number of valid attributes. (CASE B) Comparison based on functional enrichment score of valid attributes shows that algorithm Y is better than algorithm X in creating partitions as the partitions are associated with some attributes, those have high association index (associated with more number of nodes in the partitions). Functional enrichment score is denoted as FE_Score. The later option is a better way in assigning biological significance to a partition, as the method of comparison can reflect the inner picture among the valid attributes rather than treating them as equals.

3.1 Performance Comparison of algorithms

The attribute-wise study takes into account the total number of valid attributes associated with the partitions obtained by an algorithm as a measure of their performance. Their overall performance is demonstrated in Figure 1(a). It shows that the Newman's community finding algorithm's partitions are returning maximum number of valid attributes (241, 25 and 343) with respect to all the three background databases namely 'BP', 'CC' and 'GF' followed by Kernighan-Lin's algorithm (65, 12 and 106). The next better performance is that of Modularization algorithm (37, 9 and 60) followed by Farhat's algorithm (31, 2 and 47) and Greedy algorithm (25, 3 and 39). Newman's algorithm is appeared to be the

Table 2: The best sets of partitions created by different partitioning algorithms. All the results are based on human Wnt signaling pathway data. Entries list the nodes in a partition. [Farhat's algorithm: 11 partitions; Greedy algorithm: 9 partitions; Modularization algorithm: $c = 3$, 8 modules; Newman's community finding algorithm: 8 partitions, $\Delta Q = 1.0470e^{-017}$; Kernighan-Lin's algorithm: 2 partitions, initial cut-size 8, final cut-size 4]

Partition No.	Farhat	Greedy	Modularization	Newman	Kernighan-Lin
01	PSEN1, CTNNB1, PRKACA, CTNNBIP1, CHD8, SIAH1	PSEN1, CTNNB1, PRKACA, GSK3B, AXIN1, CSNK1A1L, SIAH1, TP53	LEF1, SMAD4, NLK, SOX17, CTBP1, CREBBP, RUVBL1, MYC, JUN, FOSL1, CCND1, PPARC, MMP7, MAP3K7	MAPK8, RAC1, ROCK1, RHOA, DAAM1, DVL1, PRICKLE1, FZD10, VANGL2, WNT9A	DKK1, PORCN, LRP6, CER1, WIF1, FZD10, WNT16, FZD10, SFRP1, WNT9A, VANGL2, PRICKLE1, WNT5A, PRKCA, DVL1, RAC1, DAAM1, FZD10, MAPK8, RHOA, ROCK1, PLCB1, CHP, CAMK2A, NFAT5, SIAH1, TP53, NKD1, JUN, FOSL1
02	GSK3B, DVL1, AXIN1, FRAT1, FZD10	APC2, DVL1, TBLIX, FRAT1, FZD10, CXXC4	CTNNB1, PSEN1, CTNNBIP1, CHD8, PRKACA, CSNK1A1L, FBXW11, TBLIX	WIF1, CER1, PORCN, WNT16	CCND1, MMP7, MYC, PRKACA, PSEN1, TBLIX, APC2, CTNNB1, GSK3B, AXIN1, FBXW11, DVL1, CSNK1E, PPP2CA, CTNNBIP1, CHD8, FRAT1, CXXC4, SENP2, CSNK2A1, CSNK1A1L, RUVBL1, MAP3K7, PPARC, NLK, SMAD4, LEF1, SOX17, CTBP1, CREBBP
03	WNT16, SFRP1, LRP6, PORCN, CER1	SENP2, NKD1, DVL1, FZD10, RAC1, DAAM1	DVL1, CXXC4, SENP2, CSNK2A1, FRAT1, APC2, NKD1	DKK1	-
04	DKK1, PPARC, APC2, SMAD4, LEF1	VANGL2, PRICKLE1, WNT9A, FBXW11, CSNK2A1, PPP2CA	WNT16, PORCN, FZD10, SFRP1, CER1, WIF1, LRP6, DKK1	SIAH1, TP53	-
05	NLK, SOX17, CTBP1, CREBBP, RUVBL1	CSNK1E, RUVBL1, LEF1, SMAD4, NLK, SOX17	DVL1, FZD10, RAC1, DAAM1, VANGL2, PRICKLE1, WNT9A, MAPK8, RHOA, ROCK1	NFAT5, PRKCA, CHP, CAMK2A, PLCB1, FZD10, WNT5A	-
06	MAP3K7, MMP7, WIF1, CXXC4, SENP2	CTBP1, MMP7, PPARC, CCND1, FOSL1, JUN	AXIN1, CSNK1E, GSK3B, PPP2CA	MMP7, PPARC, CCND1, FOSL1, JUN, MYC, CTBP1, SOX17, SMAD4, CREBBP, RUVBL1, NLK, MAP3K7, LEF1	-
07	APC2, TBLIX, CSNK1A1L, CCND1, FOSL1	MYC, CREBBP, CHD8, CTNNBIP1, LRP6, DKK1	PLCB1, FZD10, CAMK2A, CHP, PRKCA, WNT5A, NFAT5	CHD8, CTNNBIP1, CSNK1A1L, FBXW11, TBLIX, AXIN1, PPP2CA, APC2, CTNNB1, PRKACA, PSEN1, GSK3B, CSNK1E	-
08	JUN, MYC, NKD1, DVL1, FZD10	SFRP1, WNT16, PORCN, CER1, WIF1, MAPK8	TP53, SIAH1	NKD1, FRAT1, SENP2, DVL1, CSNK2A1, CXXC4, LRP6, FZD10, SFRP1	-
09	WNT9A, PRICKLE1, VANGL2, DAAM1, RHOA	RHOA, ROCK1, MAP3K7, WNT5A, FZD10, PLCB1, PRKCA, CHP, NFAT5, CAMK2A	-	-	-
10	MAPK8, ROCK1, RAC1, FBXW11, CSNK2A1	-	-	-	-
11	PPP2CA, CSNK1E, WNT5A, FZD10, PLCB1, PRKCA, CHP, NFAT5, CAMK2A	-	-	-	-

best algorithm for creating partitions as they are found to be associated with the highest number of valid attributes.

At a deeper level we have found that small subsets of a large partition were always found to be associated with many attributes (Figure 2). A large partition ensures presence of many subsets in it, which are associated with GO attributes; some of them being unique. Thus the corresponding P-values will be lower and they will be considered as valid. But only validity of an attribute is not sufficient for defining goodness of a module. Ideally, a valid attribute must be given more preference if it is associated with more number of nodes present in a partition than another one associated with less number of nodes in the same partition. In other words, we needed to know the number of attributes that actually show some goodness (associated with more number of nodes) in justifying a partition. Hence, a functional enrichment score system was defined to give weightage to valid attributes according to their goodness of performance.

Functional enrichment score depicts the efficiency of a partitioning algorithm in placing nodes (having a com-

mon attribute) in a partition with respect to a background database. Higher the value of the score, better is the algorithm for creating partitions. The average enrichment scores (S_{AS}) (Equation 1) of the different algorithms are shown in Figure 1(b). The Modularization algorithm is found to be performing best among all the algorithms considered here, courtesy this figure. The algorithm creates partitions with average functional enrichment score of 163.22, 258.37, and 274.19 with respect to 'BP', 'CC' and 'GF' as background databases. Kernighan-Lin's algorithm creates partitions with the least average functional enrichment score (17.66, 18.93 and 23.14 respectively) preceded by Newman's algorithm (34.45, 28.08 and 44.79 respectively), although, both the algorithms have created partitions for which maximum number of valid attributes are found to be associated (Figure 1(a)). It proves that only counting valid attributes associated with a partition is not a proper measure to deem that partition as good. Among the valid attributes, an association index must be established. Functional enrichment scores reflect such association index. Among Greedy and Farhat's algorithms,

Table 3: Module information of species-specific Wnt signaling pathways. Details about the individual Wnt signaling pathway modules of different species are given in this table [n: number of connected nodes in a species-specific pathway; r: number of relations present the connected component of a species-specific pathway; t: total number of modules created from a species-specific pathway]. The modules have been created for $c = 3$. The table throws light on the developmental trend of Wnt signaling pathway among the taken set of species. Number nodes present in each module is listed along side it in parentheses.

c	n	r	t	WNT	(DVL)1	Axin	β -catenin	TCF	TP53	(DVL)2	PLC
hsa	60	70	8	WNT [8]	(DVL)1 [7]	Axin [4]	β -catenin [8]	TCF [14]	p53 [2]	(DVL)2 [10]	PLC [7]
mmu	60	70	8	WNT [8]	(DVL)1 [7]	Axin [4]	β -catenin [8]	TCF [14]	p53 [2]	(DVL)2 [10]	PLC [7]
rno	59	69	8	WNT [7]	(DVL)1 [7]	Axin [4]	β -catenin [8]	TCF [14]	p53 [2]	(DVL)2 [10]	PLC [7]
bta	58	68	8	WNT [7]	(DVL)1 [6]	Axin [4]	β -catenin [8]	TCF [14]	p53 [2]	(DVL)2 [10]	PLC [7]
cfa	58	68	8	WNT [8]	(DVL)1 [7]	Axin [4]	β -catenin [7]	TCF [13]	p53 [2]	(DVL)2 [10]	PLC [7]
ptr	58	67	8	WNT [8]	(DVL)1 [7]	Axin [4]	β -catenin [8]	TCF [13]	p53 [2]	(DVL)2 [10]	PLC [6]
mcc	55	63	8	WNT [7]	(DVL)1 [6]	Axin [4]	β -catenin [8]	TCF [13]	p53 [2]	(DVL)2 [8]	PLC [7]
mdo	54	64	7	WNT [8]	(DVL)1 [7]	Axin [2]	β -catenin [9]	TCF [11]	-	(DVL)2 [10]	PLC [7]
gga	54	63	8	WNT [7]	(DVL)1 [6]	Axin [3]	β -catenin [8]	TCF [11]	p53 [2]	(DVL)2 [10]	PLC [7]
dre	52	60	7	WNT [8]	-	Axin [4]	β -catenin [7]	TCF [13]	p53 [2]	(DVL)2 [11]	PLC [7]
xla	43	45	6	WNT [7]	-	-	β -catenin [8]	TCF [11]	p53 [2]	(DVL)2 [8]	PLC [7]
spu	39	45	6	-	(DVL)1 [7]	Axin [2]	β -catenin [5]	TCF [10]	-	(DVL)2 [9]	PLC [6]
xtr	37	36	6	WNT [3]	-	-	β -catenin [7]	TCF [6]	p53 [2]	(DVL)2 [12]	PLC [7]
dme	36	42	7	WNT [6]	(DVL)1 [5]	Axin [2]	β -catenin [6]	TAK1 [2]	-	(DVL)2 [9]	PLC [6]
ecb	36	38	7	(Frizzled)1 [5]	(DVL)1 [3]	-	β -catenin [9]	TAK1 [2]	p53 [2]	(DVL)2 [8]	PLC [7]
nve	32	33	6	(Frizzled)1 [5]	-	Axin [5]	β -catenin [6]	TAK1 [2]	-	(DVL)2 [7]	PLC [7]
ame	30	32	5	-	(DVL)1 [4]	-	β -catenin [9]	TAK1 [2]	-	(DVL)2 [8]	PLC [7]
dpo	28	30	4	(Frizzled)1 [8]	-	-	β -catenin [7]	-	-	(DVL)2 [8]	PLC [5]
tca	26	27	4	(Frizzled)1 [7]	-	-	β -catenin [7]	-	-	(DVL)2 [6]	PLC [6]
aag	24	24	4	(Frizzled)1 [4]	-	-	β -catenin [4]	-	-	(DVL)2 [10]	PLC [6]
oaa	22	22	4	WNT [2]	-	-	β -catenin [7]	-	-	(DVL)2 [8]	PLC [5]
cel	22	20	3	-	-	-	β -catenin [10]	-	-	RhoA [6]	PLC [6]
aga	20	18	3	(Frizzled)1 [11]	-	-	β -catenin [4]	-	-	-	PLC [5]
ssc	19	16	4	FRP [2]	-	-	β -catenin [5]	TCF [7]	-	-	PLC [5]
bfo	18	16	3	-	-	-	β -catenin [9]	-	-	(DVL)2 [5]	PLC [4]
cin	17	14	3	-	(DVL)1 [7]	-	-	-	-	(DVL)2 [5]	PLC [5]
dan	16	12	4	-	(DVL)1 [2]	-	β -catenin [4]	-	-	(DVL)2 [5]	PLC [5]
bmy	13	11	3	-	(DVL)1 [4]	-	-	-	-	(DVL)2 [5]	PLC [4]
api	13	10	3	-	(DVL)1 [4]	-	-	-	-	(DVL)2 [5]	PLC [4]
tad	6	4	2	-	-	-	-	-	-	Rac [2]	PLC [4]
cbr	4	3	1	-	(DVL)1 [4]	-	-	-	-	-	-

the later performs better for the background databases 'BP' (103.65), while Greedy algorithm creates partitions that are found to be associated with more attributes of 'CC' and 'GF' databases (197.35 and 260.06).

3.2 Comparative analysis to find conserved modules

Here, modules of 31 species-specific Wnt signaling pathways (aag, aga, ame, api, bfo, bmy, bta, cbr, cel, cin, cfa, dan, dmw, dpo, dre, ecb, gga, hsa, mcc, mdo, mmu, nve, oaa, ptr, rno, ssc, spu, tad, tca, xla and xtr) were analyzed and subjected for comparative analysis. Module details of these 31 species are given in Table 3. It is important to mention here that the Wnt signaling pathway of each species may vary in terms of nodes, relations and topology. More number of absent nodes depict a pathway's lower level of development. Likewise more number of isolated nodes indicate towards poorly developed architecture of a pathway. But, in some cases nodes or relations *absentia* do mean lack of information to indicate their presence in a pathway. Table 3 gives individual details (number of nodes, relations and modules) of all the pathways considered here.

Wnt signaling pathways of the aforementioned species

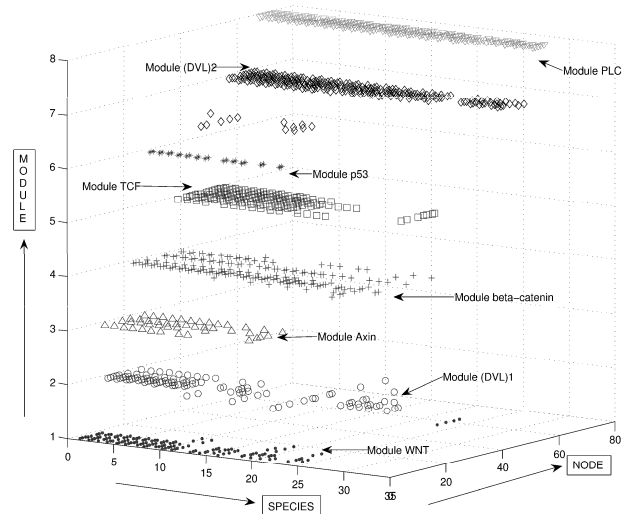


Fig. 3: One to one module wise comparison of Wnt signaling pathway of 31 different species

were subjected to modularization for $c = 3$ as for the same c -value meaningful modules were found in human Wnt

signaling pathway. We were getting 2 to 8 modules for each species that vary in their size (number of nodes present in the module) as shown in Table 3. Modules *Wnt* and β -catenin were found to be conserved in 9 species (hsa, mmu, rno, bta, cfa, ptr, mcc, mdo and gga), module *TCF* was found to be conserved in 5 species (hsa, mmu, rno, bta and cfa); module *Tp53* was observed in altogether 12 species (hsa, mmu, rno, bta, cfa, ptr, mcc, gga, dre, xla, xtr and ecb) and it was conserved by size and topology in all these species; module (*DVL*)₂ remained conserved in 11 species (hsa, mmu, rno, bta, cfa, ptr, mdo, gga, dre, spu and dme); module *PLC* turned out to be the most conserved module, found in a maximum number of 17 species (hsa, mmu, rno, bta, cfa, ptr, mcc, mdo, gga, dre, xla, spu, xtr, dme, ecb, nve and ame). Conservation patterns are shown in Figure 3.

4. Conclusions

Modularization algorithm is a better algorithm to create modules from human Wnt signaling pathway. A new GO attribute based score (Functional enrichment score) is designed for validating these modules. The score establishes a validity index among GO attributes and can be extended for performance measurement of any kind of partitions/clusters/modules created from biological networks. A comparative study of 31 species-specific Wnt signaling pathway modules is done by utilizing this algorithm. Module *PLC* is found to be the most conserved module, found in a maximum number of 17 species. Wnt signaling pathway is found to be intrinsic in many diseases; being a major player in the human cancer arena. Hence, knowledge about conserved modules can be utilized in laboratory experiments when a particular module is found to be associated with the background mechanism of a disease.

References

- [1] G. Chartrand and O. R. Oellermann, *Applied and algorithmic graph theory*. New York: McGraw Hill, 1993.
- [2] C. Farhat, "A simple and efficient automatic FEM domain decomposer," *Computers and Structures*, vol. 28, pp. 579–602, 1988.
- [3] B. W. Kernighan and S. Lin, "An Efficient Heuristic Procedure for Partitioning Graphs," *The Bell System Technical Journal*, vol. 49, pp. 291–307, 1970.
- [4] M. E. J. Newman, "Modularity and community structure in networks," in *Proc Natl Acad Sci.* PNAS, USA, 2006, pp. 8577–8482.
- [5] E. A. Leicht and M. E. J. Newman, "Community Structure in Directed Networks," *Physical Review Letters*, vol. 100, no. 118703, 2008.
- [6] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.
- [7] L. Nayak and R. K. De, "An algorithm for modularization of MAPK and calcium signaling pathways: Comparative analysis among different species," *Journal of Biomedical Informatics*, vol. 40, pp. 726–749, 2007.
- [8] J. Saez-Rodriguez, S. Gayler, M. Ginkel, and E. D. Gilles, "Automatic decomposition of kinetic models of signaling networks minimizing the retroactivity among modules," *Bioinformatics*, vol. 24, pp. i213–i219, 2008.
- [9] E. Grafahrend-Belau, F. Schreiber, M. Heiner, A. Sackmann, B. H. Junker, S. Grunwald, A. Speer, K. Winder, and I. Koch, "Modularization of biochemical networks based on classification of Petri net t-invariants," *BMC Bioinformatics*, vol. 9, no. 90, 2008.
- [10] P. H. Lee and D. Lee, "Modularized learning of genetic interaction networks from biological annotations and mRNA expression data," *Bioinformatics*, vol. 21, pp. 2739–2747, 2005.
- [11] R. L. Chang, F. Luo, S. Johnson, and R. H. Scheuermann, "Deterministic Graph Theoretic Algorithm for Detecting Modules in Biological Interaction Networks," *International Journal of Bioinformatics Research and Application*, vol. 6, no. 6, pp. 101–119, 2010.
- [12] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, pp. 166–176, 2003.
- [13] K. Macropol, T. Can, and A. K. Singh, "RRW: repeated random walks on genome-scale protein networks for local cluster discovery," *BMC Bioinformatics*, vol. 10, p. 283, 2009.
- [14] W. S. Verwoerd, "A new computational method to split large biochemical networks into coherent subnets," *BMC Systems Biology*, vol. 5, no. 25, 2011.
- [15] S. Sun, X. Dong, Y. Fu, and W. Tian, "An iterative network partition algorithm for accurate identification of dense network modules," *Nucleic Acids Research*, vol. 40, no. 3, p. e18, 2012.
- [16] M. Mete, F. Tang, X. Xu, and N. Yuruk, "A structural approach for finding functional modules from large biological networks," *BMC Bioinformatics*, vol. 9(Suppl 9), no. S19, 2008.
- [17] R. Guimera and L. A. N. Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, 2005.
- [18] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. Deustachio, E. Schmidt, B. D. Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Research*, vol. 33, pp. D428–D432, 2005.
- [19] D. Nishimura, "A view from the Web: Biocarta," *Biotech. Software and Internet Report*, vol. 2, no. 3, pp. 117–120, 2001.
- [20] C. F. Schaefer, K. Anthony, K. S. J. Buchoff, M. Day, H. T. and B. K. H., "PID: The Pathway Interaction Database," *Nucleic Acids Res.*, vol. 37, pp. D674–D679, 2009.
- [21] K. Kandasamy, S. S. Mohan, R. Raju, S. Keerthikumar, G. S. S. Kumar, A. K. Venugopal, D. Telikicherla, J. D. Navarro, S. Mathivanan, C. Pecquet, S. K. Gollapudi, S. G. Tattikota, S. Mohan, H. Padhukasahasram, Y. Subbannayya, R. Goel, H. K. C. Jacob, J. Zhong, R. Sekhar, V. Nanjappa, L. Balakrishnan, R. Subbaiah, Y. L. Ramachandra, B. A. Rahiman, T. S. K. Prasad, J. Lin, J. C. D. Houtman, S. Desiderio, J. Renauld, S. N. Constantinescu, O. Ohara, T. Hirano, M. Kubo, S. Singh, P. Khatri, S. Draghici, G. D. Bader, C. Sander, W. J. Leonard, and A. Pandey, "NetPath: a public resource of curated signal transduction pathways," *Genome Biology*, vol. 11, p. R3, 2010.
- [22] K. I. Fukuda and T. Takagi, "Knowledge Representation of Signal Transduction Pathways," *Bioinformatics*, vol. 17, pp. 8290–837, 2001.
- [23] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 2000.
- [24] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks," *Bioinformatics*, vol. 21, pp. 3448–3449, 2005.
- [25] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, 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, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.
- [26] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Res.*, vol. 13, pp. 2498–2504, 2003.
- [27] A. Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," *Bioinformatics*, vol. 24, pp. 1359–1366, 2008.

Identification and Bioinformatic Analyses of Vanillate Operon in Cyanobacterium *Synechococcus* sp. IU 625

Robert Newby, Jr., and Tin-Chun Chu

Department of Biological Sciences, Seton Hall University, South Orange, NJ, USA

Abstract - Vanillate is a byproduct of saprotrophic digestion of plant lignin. We demonstrated that *Cyanobacterium Synechococcus* sp. IU 625 (*S. IU 625*) is capable of utilizing vanillate as a sole carbon source when deprived of light for photosynthesis. Result indicated that *S. IU 625* is capable of utilizing vanillate as a sole carbon source at 0.5 and 1.0 mM concentrations when light is not provided for photosynthesis. Using the sequences obtained from *Caulobacter crescentus* NA1000, analysis of potential genes for vanillate utilization has been carried out. Of the three proteins which are coded for in the vanillate utilization operon in *C. crescentus*, *VanA*, *VanB*, and *VanR*; homology from *VanB* has been found to two unique unidentified proteins in a related species of cyanobacteria *Synechococcus elongatus* PCC 7942. This study was aimed at identification of the genes which compose vanillate operon in cyanobacteria to better understand potential cyanobacterial heterotrophic growth.

Keywords: Cyanobacteria, Vanillate operon, Sequence alignment

1 Introduction

Cyanobacteria (formally known as blue-green algae) are photosynthetic prokaryotes of great importance in many ecological settings. They affect water quality and play a huge role in global biogeochemical cycles [1]. Harmful algal blooms (HAB) caused by eutrophication have been reported in nearly every industrialized nation [2, 3]. *Synechococcus* sp. IU 625 (*S. IU 625*) is a non-toxin producing freshwater unicellular cyanobacterium which has been reported to cause HAB previously.

Heavy metals such as iron, copper, nickel, cobalt, and manganese are important trace nutrients and are often added to fertilizers to enhance plant growth [4]. These heavy metals are also released by unregulated industrial waste water effluent [5]. Several of these heavy metals have been designated by the US Environmental Protection Agency (EPA) as potential threats. These EPA targeted heavy metals include zinc, nickel, cobalt, iron and manganese.

HAB appear to be enhanced by not only eutrophication, but also by the presence of several EPA target heavy metals. Research has shown that a natural predator of cyanobacteria, cyanophage, is inhibited by the presence of high concentrations of heavy metals [6]. Cyanophages are important in regulating the growth of cyanobacteria [7]. Heavy metals appear to further alter their environment by altering the pH, changing the oxidative state of nutrients, and in some cases as reported with mercury, will bioaccumulate in the ecologic food web. This alteration of the environment is creating a niche for heavy metal resistant cyanobacteria to flourish and potentially bloom [8].

Cyanobacteria were originally generally considered obligate photoautotrophs [9]. Evidence provided in this study and other experimental studies shows that *S. IU 625* and other cyanobacteria can be photoheterotrophs under certain conditions [10]. Photoheterotrophic growth of cyanobacteria can be used as a potential tool for molecular studies of cyanobacteria much similar to that of other bacteria; such as lactose based expression in *E. coli* and vanillate based expression in *Caulobacter crescentus* (*C. crescentus*). Vanillate is the byproduct of the saprotrophic digestion of plant lignin [11]. It is a phenolic compound, which can be used a sole carbon source in several prokaryotic species. Typically the vanillate operon (*Van*) consists of three genes: *vanA*, a monooxygenase; *vanB*, a phenolic demethylase; and *vanR*, a transcriptional repressor [12]. Work has previously been done in *C. crescentus* NA1000, for the characterization and cloning of the promoter from the *Van* region; and a set of molecular vectors has been established linking the expression of a gene of interest to the *Van* promoter [12]. This study focuses on identification, cloning and characterization of *Van* operon in *S. IU 625* using bioinformatics analysis for screening *S. IU 625* for the ability to utilize vanillate as a sole carbon source.

2 Materials and Methods

2.1 Growth conditions of *Synechococcus* sp. IU 625

The unicellular Cyanobacterium *S. IU 625* was obtained from American Type Culture Collection (ATCC);

Manassas, VA) and was maintained in sterile Mauro's Modified Medium (3M) at a pH 7.9. The cells were grown at 26°C, with constant fluorescent light and continuous agitation at 100 rpm in a Gyromax 747R incubator shaker (Amerex Instruments; Lafayette, CA). *S. IU 625* was maintained in sterile 250 ml Erlenmeyer flasks.

2.2 Bioinformatic analysis

Known sequences from the van operon in *C. crescentus* were obtained from GenBank on the National Center for Bioinformatics (NCBI) website. Using the sequences BlastP searches were carried out into *Synechococcus elongatus* PCC 7942 and queries of high match were recorded.

2.3 DNA isolation and purification

The exponentially growing *S. IU 625* cells were collected for DNA isolation. Sambrook and Russell's DNA extraction and purification protocol was followed with minor modifications [13]. The concentration and the purity of the isolated DNA were determined by NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE).

2.4 Primer design

PCR primers were designed with National Center for Bioinformatics (NCBI) Primer-BLAST and Integrated DNA Technology's (IDT) PrimerQuestSM software. Designed primers were analyzed with IDT's OligoAnalyzer 3.1 program. A closely related species of cyanobacterium, *Synechococcus elongatus* PCC 7942, whose sequence is known, was used as a template. Primers were obtained from Eurofins MWG-Operon (Huntsville, AL), and were resuspended in sterile diH₂O to a final concentration of 100 µM following manufactures recommendations. Primers were designed to encompass the entire hypothetical vanillate operon (*vanR*, *vanA*, and *vanB*), the size of the amplicons ranging from 300-800 bp. No primer was designed to amplify more than 850 bp. Each oligo was designed to be under 30nt in length with melting temperatures of the oligos 60°C or above.

2.5 PCR-based assay, gel electrophoresis and sequencing

PCR-based assays were carried out with all the designed primers. For each 25 µl reaction tube, it contains the following, 1 µl of genomic DNA template, 12.5 µl of 2X GoTaq® Hot Start Green Master Mix (Promega; Madison, WI), 1 µl of forward and 1 µl reverse primer (both 10 µM), 2

µl DMSO and 7.5 µl of nuclease-free H₂O. The general run method of reaction was activation of the Hot Start polymerase at 95°C for 2 minutes, followed by denaturation for 30 seconds, lowest T_m of primer group for 30 seconds, 72°C for 30 seconds, for 35 cycles. A final extension step was done at 72°C for 5 minutes. Followed by the PCR, 1% agarose gel electrophoreses were carried out and the PCR products were also sent out for sequencing (Genewiz, Inc., South Plainfield, NJ).

2.6 *S. IU 625* growth monitoring with or without vanillate

Vanillate (50 mM) were purchased from Sigma-Aldrich. The stock solution was prepared with sodium hydroxide and was filtered through a 0.2 µm filter. Six sterile Erlenmeyer flasks for both light and dark sets (3 flasks each) were autoclaved and labeled accordingly. A *S. IU 625* culture with OD_{750nm} approximately 1.0 was used. Each flask contains 5 ml *S. IU 625* and 95 ml 3M media. Vanillate was then added into the flasks in duplicates of 0 (control), 0.5 and 1 mM final vanillate concentrations, respectively. The light set of flasks was grown under the standard growth conditions while the dark set of the flasks were wrapped with aluminum foil to prevent the light exposure. The cell growths were monitored with Ultrospec III (Pharmacia LKB, Sweden). Vanillate utilization for all 6 flasks was also monitored with NanoDrop ND-1000 (Thermo Fisher Scientific, Wilmington, DE).

3 Results

Cyanobacteria Using the BlastP of the VanA in *C. crescentus* into *Synechococcus elongatus* PCC 7942, a comprehensive set of primers was designed using NCBI Primer-Blast and PCR-based assays were carried out. Prior to primer design, bioinformatics work was carried out to try to elucidate the location of a potential vanillate response operon in cyanobacteria. Using the known sequence from *C. crescentus*, Figure 1A shows the bioinformatic analyses of the protein sequence of VanA in *C. crescentus* compared with *S. elongatus* PCC 7942. Proposed Van operons in *S. IU 625* and *S. elongatus* PCC 7942 are shown in Figure 1B. Figure 1C shows a partial order alignment visualizer (POAVIZ) of the VanA protein in several species of bacteria compared to the hypothetical VanA sequence in *S. elongatus* PCC 7942. Tables 1 listed out all the designed primers used in this study. Gel electrophoreses for the PCR products were carried out and selected results are shown in Figure 2. PCR products were sent out for sequencing and BLAST results were performed on the obtained sequences.

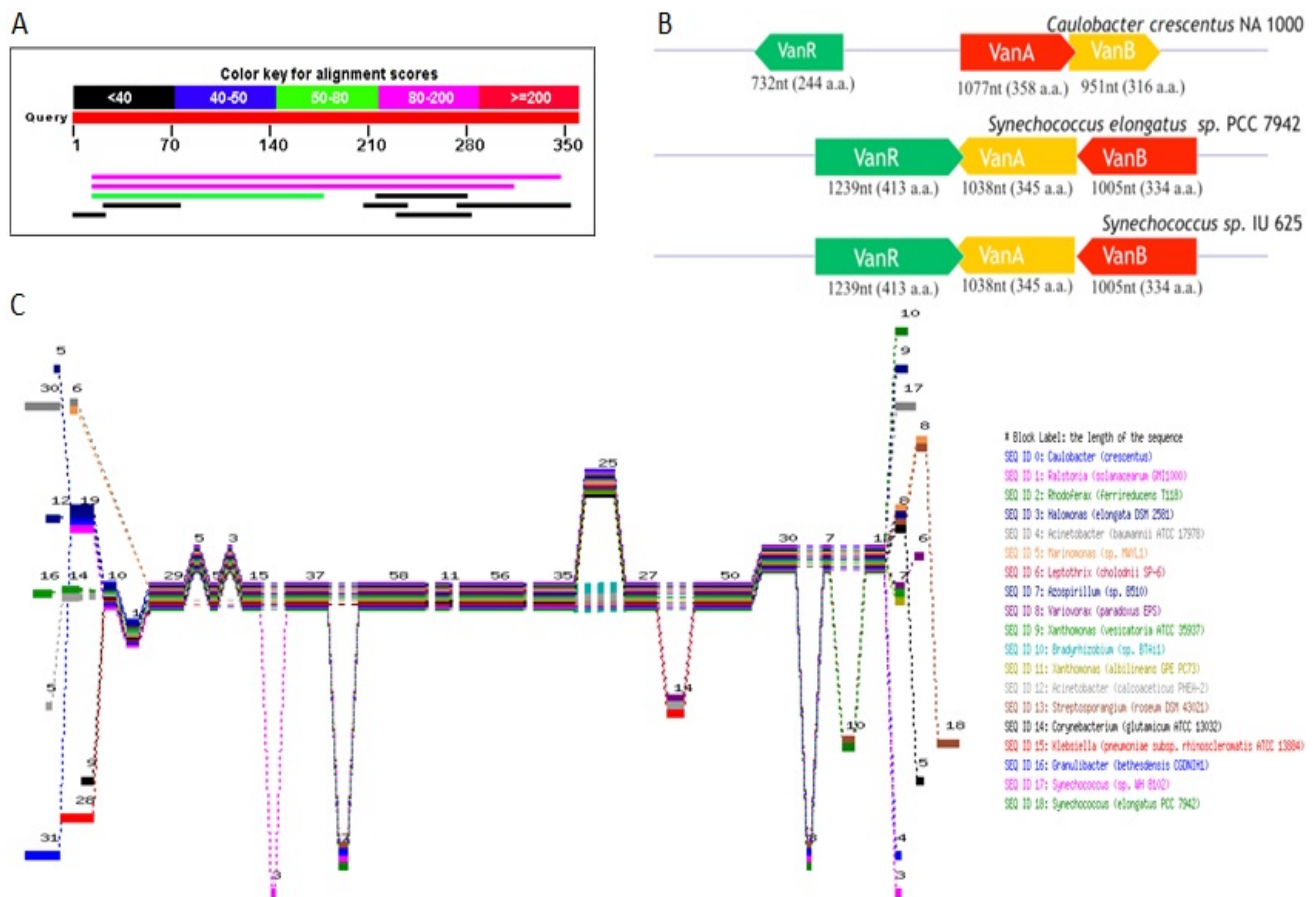


Figure 1. A) BlastP results for VanA proteins. BlastP matches are shown using the VanA protein sequence from *C. crescentus* into the genome of *S. elongatus* PCC 7942. The top two results are SynPCC7942_2035 and SynPCC7942_2036 respectively. B) Proposed Van operons in *S. IU 625* and *S. elongatus* PCC 7942. C) POAVIZ result of known bacterial VanA compared with *S. elongatus* PCC 7942 (Smooth: 2)

Primer	Sequence 5' → 3'	Tm (°C)	Size (bp)	Primer	Sequence 5' → 3'	Tm (°C)	Size (bp)
VanA_3F	CCTTGGGCGACGAGGGAGGA	68.6	551	VanB_1F	TCAGCGGCGACGTGGGTTTC	66.6	572
VanA_3R	GCGGGCAGACTATCGTCGCGT	66.6		VanB_1R	GATCGTCCACCCTTGCCGCC	68.6	
VanA_4F	TGCGTTTGC GTGCCACCA	64.5	683	VanB_2F	TGGTGACGGTACGTGGGGCA	66.6	495
VanA_4R	GGCTTCGCTAGCCTGCGGTC	64.6		VanB_2R	GCGCTGGCTCAGAAGGTCGG	66.6	
VanA_5F	CTGCCGTC AAGAAATTGCGGACAT	64.6	582	VanR_1F	TGATTGCGACCCCCACACC	66.6	762
VanA_5R	ATCTCGATGGTCATTGCGGCTCAGA	64.6		VanR_1R	GCGGAAACCAAGGGCTCGCA	66.6	
VanA_6F	GAGCCAGCGCCACGGGATAC	68.6	639	VanR_2F	GCGCAGCTGCTAGGCCAGAT	66.6	818
VanA_6R	ATGACCAAGCCGGTTGGCGG	66.6		VanR_2R	AGTTCTGCGACCCGCTTGCC	66.6	
VanA_7F	GCGACGATAGTCTGCCGCC	68.6	674	VanR_3F	AACAGCTGGTGCGCAGAGGG	66.6	787
VanA_7R	TCGCCAATGTTGGTGACCGGC	66.6		VanR_3R	AGTGAACCCACGTCGCCGC	66.6	

Table 1. Selected primers designed with NCBI Primer-BLAST and verified with IDT OligoAnalyzer 3.1 program. Ten primer sequences, Tm and the amplicon sizes are listed.

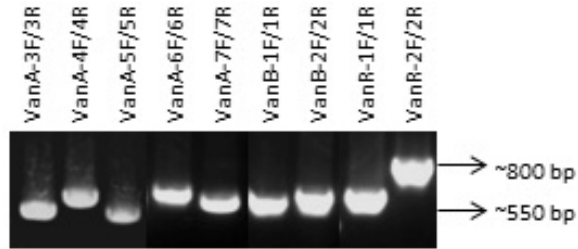


Figure 2. Selected gel electrophoresis results for the PCR products. All the sizes of the PCR products correspond to the estimated amplification size.

The results obtained indicated that *S. IU 625* may be able to utilize vanillate as a sole carbon source when photosynthesis is unavailable. By monitoring growth of *S. IU 625* cultures inoculated with 0, 0.5 and 1.0 mM concentrations of vanillate, as well as presence or absence of light, the ability of *S. IU 625* to utilize vanillate was experimentally shown. Once the monitoring study was completed the information from the study was compiled and graphed to show the growth patterns of *S. IU 625* when exposed to vanillate. Figure 3 shows the growth of the *S. IU 625* cells in the light and the dark condition.

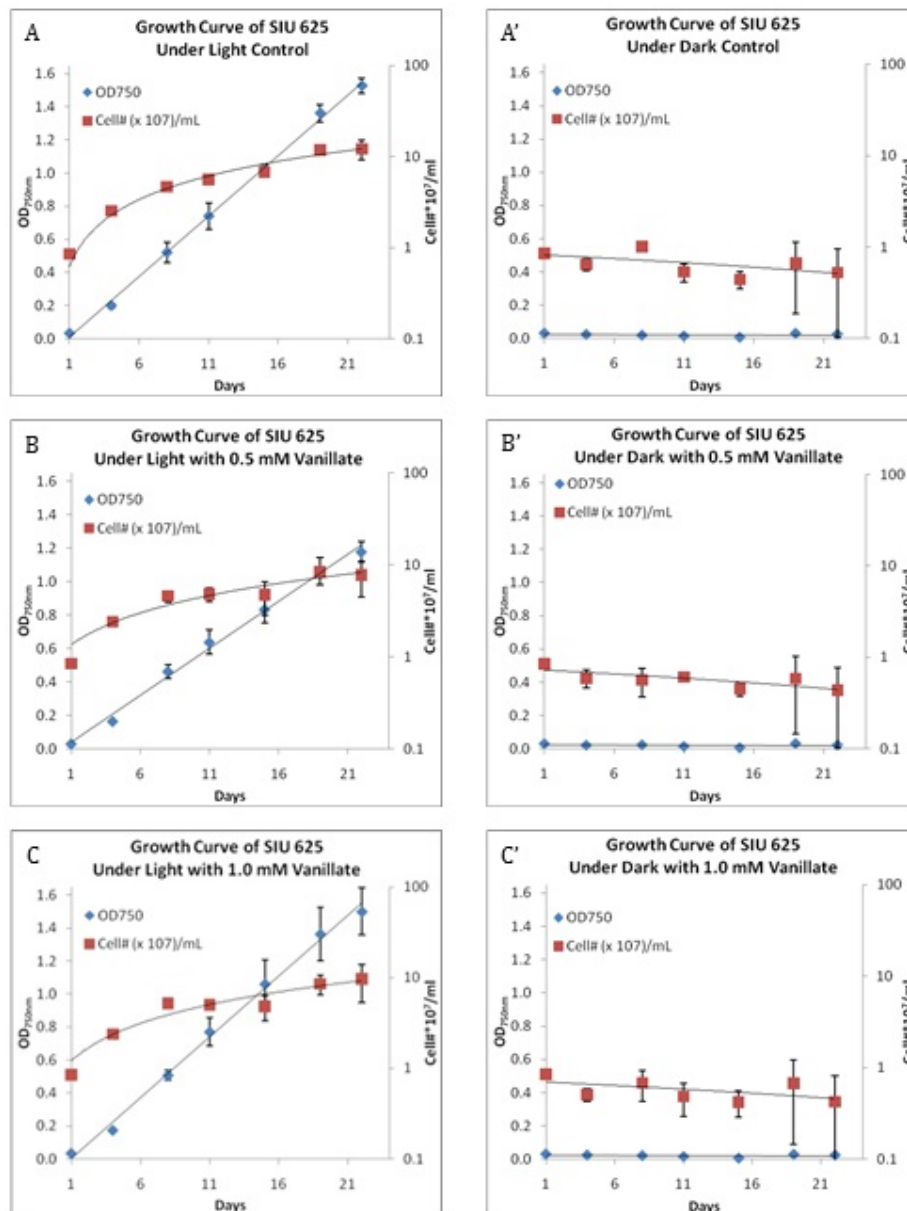


Figure 3. The growth curves of *S. IU 625* with 0, 0.5 and 1.0 mM vanillate under light (A-C) and dark (A'-C') conditions. The cells in the light set had similar growth with or without vanillate. The cells in the dark set had minimal growth for all three conditions (0, 0.5 and 1.0 mM vanillate).

The absorbance at OD₂₈₆ was used to measure the degradation and utilization of vanillate in media. Based on work in Thanbichler (2007), this was the measured non-interfering absorbance of vanillate in supernatant. By comparing the readings at OD₂₈₆ we were able to estimate the relative concentration of vanillate in the supernatant. Figure 4 shows the measured vanillate concentration in the supernatant for both the light and dark sets of cells. Another indication that vanillate was being utilized is apparent for

the dark set. On day 22, the cultures were centrifuged to separate the cells and the supernatant. The color of the supernatant in the flask with 1.0 mM vanillate under the light was much darker compared with the flask with 0.5 mM vanillate; while all three flasks in the dark set contain colorless supernatant after centrifugation. It indicated the *S. IU 625* cells were able to utilize vanillate as carbon source when the light is absent.

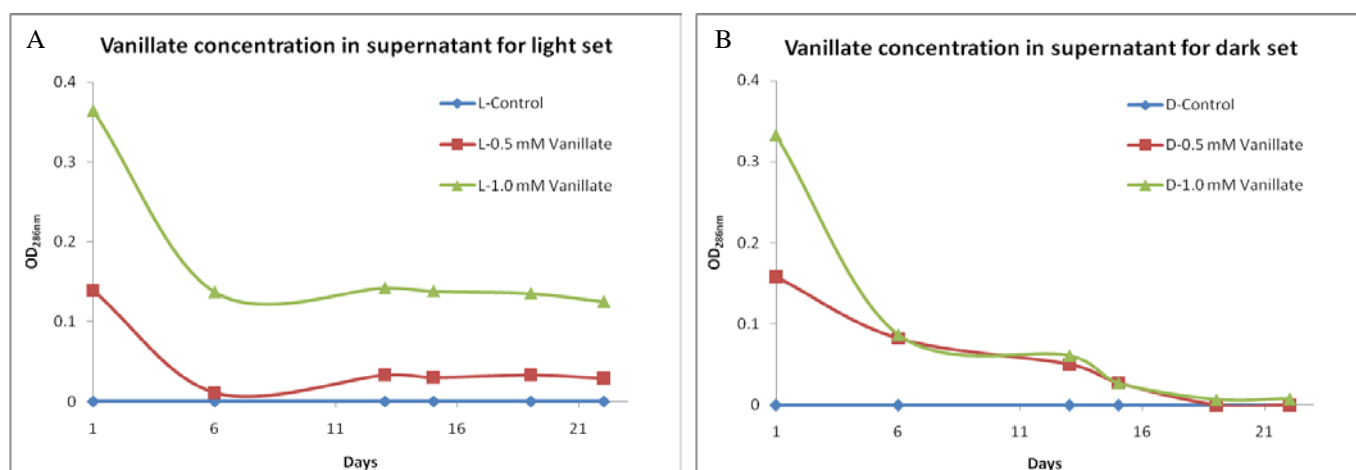


Figure 4. Degradation of vanillate (0, 0.5 and 1.0 mM) under (A) light and (B) dark conditions. One ml of the culture in each flask was collected at 6 time points throughout 22 day period. Vanillate concentration in supernatant were measured in OD_{286nm}. The degradation curve indicates that the cells in the light set were able to utilize some vanillate while the cells in the dark set utilized the vanillate completely by day 20.

4 Discussions

Vanillate has been experimentally shown to encourage heterotrophic growth of *Synechococcus* sp. *IU 625*. This is a novel study indicated that the obligated photoautotroph *S. IU 625* can also survive and grow in the dark condition with supplemented vanillate. Since experimental evidence exist which show the ability of *S. IU 625* to utilize vanillate as a sole carbon source, a cluster of genes for its regulation and processing must exist. In our experiments we showed that a high homology between vanillate response genes in other freshwater oligotrophs, such as *Caulobacter crescentus*, has homology to a cluster of genes in *S. IU 625*. Sequencing of the operon, which was undertaken using *Synechococcus elongatus* PCC 7942 as a template for design, gave insight into the genome of *S. IU 625* not previously seen. One potential purpose of the genes might be to allow the cell to survive when nutrient deprived or in an area where sunlight is blocked by natural foliage. The ability for *S. IU 625* to utilize vanillate is an important finding since it allows a better understanding of how cyanobacteria respond to external sources of nutrients. Unpublished data shows that this cluster of genes exists in several other species of

cyanobacteria. Based on the work in Thanbichler, 2007, vanillate is degraded by the enzyme complex VanAB. The VanAB enzyme breaks down vanillate to pyruvate. However the exact mechanism is not known in *S. IU 625* and is an interest for further studies for this project.

This study sought to explore the use vanillate inducible gene expression to measure gene expression. By cloning the hypothetical promoter from the vanillate operon, we sought to make a molecular switch, similar to the LacZ based promotion in *E. coli*. This work has been previously described in great detail in *C. crescentus*, and a set of vanillate inducible plasmids exists based on work by Thanbichler, 2007. These vectors are similar to the xylose inducible vectors created for *Staphylococci* spp. in Wieland et al., 1995 [14].

Use of a carbon based inducible vectors has not been shown previously for cyanobacteria. This study represents the first of its kind to show that *S. IU 625* is capable of both heterotrophic and phototrophic growth. Based on the results seen in Figures 3, *S. IU 625* is capable of surviving up to 22 days when supplemented with vanillate. However, vanillate

degradation is not an optimal source of energy to sustain the continued growth of *S. IU 625*.

5 References

- [1] Roderick Oliver and George Ganf. "Freshwater Blooms". p. 149-194. In B. A. Whitton and M. Potts (ed.), "The Ecology of Cyanobacteria". Kluwer Academic Publishers, 2000.
- [2] Hans W. Paerl, Rolland S. Fulton, III., Pia Moisaner, and Julianne Dyble. "Harmful freshwater algal blooms, with an emphasis on cyanobacteria". *Sci. World*, 1: 76-113. Apr. 2001.
- [3] Joan L. Slonczewski and John W. Foster. "Microbiology: an evolving science". W. W. Norton & Co, 2008.
- [4] Hendrik Küpper, Frithjof Küpper and Martin Spiller. "Environmental relevance of heavy metal substituted chlorophylls using the example of water plants"; *Journal of Experimental Botany*, 47 (295): 259-266, Feb 1996.
- [5] Stéphane Audry, Jörg Schäfer, Gérard Blanc and Jean-Marie Jouanneau. " Fifty-year sedimentary record of heavy metal pollution (Cd, Zn, Cu, Pb) in the Lot River reservoirs (France)"; *Environ. Pollut.*, 132 (3): 413-426, Dec 2004.
- [6] Lee H Lee, Doris Lui, Patricia J. Platner, Shi-Fang Hsu, Tin-Chun Chu, John J. Gaynor, Quinn C Vega and Bonnie K Lustigman. "Induction of temperate cyanophage AS-1 by heavy metal - copper"; *BMC Microbiology.*, 6: 17, Feb 2006.
- [7] Martin Mühling, Nicholas J. Fuller, Andrew Millard, Paul J. Somerfeld, Dominique Marie, William H. Wilson, David J. Scanian, Anton F. Post, Ian Joint, and Nicholas H. Mann. "Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton"; *Environ. Microbiol.*, 7 (4): 499-508, Apr 2005.
- [8] N. Thajuddin, and G. Subramanian. "Cyanobacterial biodiversity and potential applications in biotechnology"; *Current Science*, 89 (1), 1-14, Jul 2005.
- [9] William A. Kratz and Jack Myers. "Nutrition and Growth of Several Blue-Green Algae"; *American Journal of Botany*, 42 (3): 282-287, Mar 1955.
- [10] Shawn L. Anderson and Lee McIntosh. "Light-Activated heterotrophic growth of the cyanobacterium *Synechocystis* sp. Strain PCC 6803: a blue-light-requiring process"; *J. Bacteriol.*, 173(9): 2761-2767. May 1991.
- [11] Francoise Brunel and John Davidson. "Cloning and Sequencing of *Pseudomonas* Genes Encoding Vanillate Demethylase"; *J. Bacteriol.*, 170(10): 4924-4930, Oct 1988.
- [12] Martin Thanbichler, Antonio A. Iniesta and Lucy Shapiro. "A comprehensive set of plasmids for vanillate and xylose-inducible gene expression in *Caulobacter crescentus*"; *Nucleic Acids Research*, 35(20): e137, Oct 2007
- [13] Joseph Sambrook and David W. Russell. "Molecular Cloning: A Laboratory Manual". 3rd ed. Cold Spring Harbor Laboratory Press. 2001.
- [14] Karsten-Peter Wieland, Bernd Wieland and Friedrich Götz. "A promoter-screening plasmid and xylose-inducible, glucose-repressible expression vectors for *Staphylococcus carnosus*"; *Gene*, 158(1): 91-96. May 1995.

Evolutionary Approach of Ligand Design for Protein-Ligand Docking Problem Using Artificial Bee Colony Optimization

Pratyusha Rakshit¹, Papia Das², Archana Chowdhury¹, Amit Konar¹, Mita Nasipuri², Atulya K. Nagar³

¹ETCE, ²CSE Dept., Jadavpur University, Kolkata, India

³Department of Math & Computer Science, Liverpool Hope University, Liverpool, UK

Abstract- *The paper addresses an interesting approach to protein-ligand docking problem using Artificial Bee Colony optimization algorithm. In this work, protein-ligand docking is formulated as an optimization problem. The docking energy is used as a scoring function for the solutions. Results are demonstrated for six different target proteins both numerically and pictorially. Experimental results reveal that the proposed method outperforms Variable string-length Genetic Algorithm based ligand design method considering the intra- and inter-molecular energies of the evolved molecules.*

Keywords- ligand; artificial bee colony optimization algorithm; fragment based approach; CHARMM energy; active site .

1 Introduction

Proteins are macromolecules consisting of two or more amino acids. They are greatly responsible for structural and functional characteristics of cells, and communication of biological signals among cells. Active sites in protein molecules refer to a part of the molecule primarily responsible for its functioning. They are usually hydrophobic pockets involving side chain atoms. All protein molecules available in the nature are not equally useful for the living organisms. On the contrary, there are evidences of proteins, causing fatal or infective diseases. Researchers are taking keen interest to selectively identify the right candidate structure that fits well in the active site of a protein. These molecules capable of binding at the active site and thereby changing the functional behavior of the protein are called ligands.

Docking is of extreme relevance in cellular biology, where function is accomplished by proteins interacting with themselves and with other molecular components [6]. It is the key to oriental drug design. The result of docking can be used to find inhibitors for specific target proteins and thus to design new drugs. Protein-ligand docking is an energy minimization search problem with the aim to find the best ligand conformation and orientation relative to the active site of a target protein.

Research aimed at solving the protein-ligand docking problem considers designing interesting algorithms to balance the efficient search for fitting the ligand optimally with the target protein with the order of complexity required to execute the algorithm. Application of evolutionary computation for ligand molecule discovery by searching the

vast organic space of active site of receptor protein thus is apparent.

In this paper, we study the scope of the well-known Artificial Bee Colony (ABC) optimization algorithm [3] to judiciously determine the ligand structure to be docked at the active site of a protein. The choice of ABC in the present context is inspired partly heuristically because of the background of the algorithm in the topic, and partly because of its established performance in the literature [3, 5].

Performance of an evolutionary algorithm in engineering search or optimization problem greatly depends on the data structure used to represent 'evolvable trial solutions'. We observed that one interesting time-efficient solution to the ligand docking problem can be realized by selecting a tree-structure for the ligand. The tree structure helps in connecting primitive fragments or radicals to determine the right candidate solution for the ligand that best suits to the active site of the protein. The swarm evolutionary algorithm to be employed randomly connects the radicals and then filters unwanted connection by providing higher penalty to the resulting trial solutions. The process of random selection is continued until an appreciably good ligand structure is selected based on its 'fitness' measure.

A 'fitness function' is generally introduced in a meta-heuristic algorithm to determine the desired solutions for an optimization problem. Naturally, the better the formulation of the fitness function, the better is the expected quality of the trial solutions. In the present context of the ligand docking problem, optimal selection of the ligand is inspired by minimization of an energy function that determines the stable connectivity between the protein and the ligand. So, the fitness function here is an energy function, whose minimization yields trial solutions to the problem.

In this paper, the CHARMM energy function [7] is used as a scoring function to evaluate the affinity between the ligand and the protein. It is based on decomposition of the ligand binding energy into individual interaction terms such as van der Waals energies, electrostatic energies, bond stretching, bending, and torsional energies, etc., using a set of derived force-field parameters.

In [1], fixed length genetic algorithm (GA) is used to evolve molecular structure of possible ligands that bind to a given target protein. The molecules are represented by tree-like structure, composed of atoms at the nodes and the bonds as links. Evidently, an a priori knowledge of the size of the tree is difficult to obtain. Another approach for ligand design, which is based on variable length representation of trees on

both sides of the pharmacophore, was studied by Bandopadhyay *et al.* [2]. However, the approach is restricted to build the ligand in two-dimensional space from a small suite of seven functional groups. Furthermore, the fitness value of the ligand is determined by the measure of van der Waals force only.

This paper has significantly improved the work proposed in [2] because of the following reasons. 1) It uses a new representation for the ligand utilizing dynamic memory allocation technique, 2) It constructs the ligand using a larger suite of fragments, 3) It optimizes both intra- and inter-molecular docking energies of the docked molecule, and 4) ABC with its ability to handle combinatorial explosion appears to be very promising to the ligand design problem addressed here. It can be seen from the results that, ABC based optimization model has performed quite satisfactorily in comparison with Variable string-length Genetic Algorithm (VGA) as proposed in [2].

The rest of the paper is organized as follows. In section 2, we explain the formulation of protein- ligand docking problem. Section 3 depicts the principles used to predict the ligand structures. In section 4, we describe the artificial bee colony optimization algorithm used to find the best ligand structure. The pseudo-code for solving the given constrained optimization function is scripted in Section 5. We present the experimental results for six proteins in section 6. Section 7 concludes the paper.

2 Formulation of the Problem

In protein- ligand docking problem, the objective is to minimize the energy. Firstly the internal energy of the ligand should be minimized for better stability of the ligand. This intra-molecular energy calculation is based on the calculation of interaction energy of the different functional groups within the ligand and incorporates the bond stretching, angle bending, and torsion terms. The inter-molecular energy value, which is thereafter optimized, is the interaction energy between the ligand and the active site of the receptor protein. This energy calculation is based on the proximity of the different residues in the active site of the receptor protein to the closest functional groups in the ligand and their chemical properties. The inter-molecular non bonding interaction energy is computed in terms of the van der Waals energy and the electrostatic energy.

Hence, in order to perform a qualitative analysis of the conformation of ligands in large space, there is a need of some cost or energy functions, commonly known as force fields. In this work the CHARMM force fields are considered to evaluate the cost of the conformations which is commonly known as Chemistry at HARvard Macromolecular Mechanism. CHARMM models the dynamics and mechanism of macromolecular system using empirical and mixed empirical quantum mechanical force fields. CHARMM uses potential functions that approximate the total potential as a sum of bond stretching, bond bending, bond twisting, improper potentials which are used to maintain planar bonds, plus potentials representing the nonbonded van der Waals and electrostatic interactions.

The energy of the bond stretching is approximated as

$$V_{\text{bond}} = K_b (b - b_0)^2 \quad (1)$$

where K_b is a constant that depends on the identity of the two atoms sharing the bond in a ligand, b is the length of the bond and b_0 is the unstrained bond length in equilibrium.

The energy of the bond bending is approximated as

$$V_{\text{angle}} = K_\theta (\theta - \theta_0)^2 \quad (2)$$

where K_θ is a constant that depends on the three atoms defining the angle θ within a ligand, θ is the angle between the atoms and θ_0 is the unstrained angle in equilibrium.

Determination of the energy of bond twisting (dihedral energy) requires four atoms of a ligand to define the bond and the amount it is twisted. It is approximated as

$$V_{\text{dihedral}} = K_\chi (1 + \cos(n\chi - \delta)) \quad (3)$$

where K_χ and δ are constants that depend on the adjacent atoms, n is an integer that depends on the number of bonds made by atoms, and χ is the value of the dihedral angle.

Improper forces or potentials are artificial forces or potentials that are used to hold a group consisting of one central atom that is bonded to three others in a particular configuration. The potential that is used in CHARMM for improper dihedrals of a ligand is

$$V_{\text{improper}} = K_\psi (\psi - \psi_0)^2 \quad (4)$$

where K_ψ is a constant and ψ is the improper angle that depends on the coordinates of the atoms and ψ_0 is the equilibrium improper angle.

More elaborate force field may include the Urey-Bradley term given as

$$V_{\text{Urey-Bradley}} = K_{UB} (S - S_0)^2 \quad (5)$$

where K_{UB} is the Urey-Bradley force constant, S is the distance between two atoms separated by two covalent bonds (1, 3 distance) and S_0 is the equilibrium distance.

Therefore, the intra-molecular energy of a ligand or bonded energy is given by

$$V_{\text{bond}} = \sum_{\text{bond}} K_b (b - b_0)^2 + \sum_{\text{angle}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedral}} K_\chi (1 + \cos(n\chi - \delta)) + \sum_{\text{improper}} K_\psi (\psi - \psi_0)^2 + \sum_{UB} K_{UB} (S - S_0)^2 \quad (6)$$

Van der Waals interactions between two atoms within the active site are approximated with a Lennard-Jones potential as

$$V_{\text{Lennard-Jones}} = \varepsilon_{i,j} \left[\left(\frac{R_{\text{min},i,j}}{r} \right)^{12} - 2 \left(\frac{R_{\text{min},i,j}}{r} \right)^6 \right] \quad (7)$$

where $\varepsilon_{i,j}$ is the Lennard-Jones well depth, r is the distance between atoms i and j , $R_{\text{min},i,j}$ is the minimum interaction radius.

The electrostatic interaction between two atoms is

$$V_{\text{electrostatic}} = \frac{q_i q_j}{4\pi\epsilon r} \quad (8)$$

fitness of the associated solution. The number of employed bees and onlooker bees is equal to the number of solutions in the population. ABC consists of following steps:

4.1 Initialization

ABC generates a randomly distributed initial population P of N_p solutions (food source positions) where N_p denotes the size of population. Each solution X_i ($i=0, 1, 2, \dots, N_p - 1$) is a D dimensional vector.

4.2 Placement of employed bees on the food sources

An employed bee produces a modification on the position in her memory depending on the local information (visual information) as stated by equation (13) and tests the nectar amount of the new source. Provided that the nectar amount of the new one is higher than that of the previous one, the bee memorizes the new position and forgets the old one. Otherwise she keeps the position of the previous one in her memory.

4.3 Placement of onlooker bees on the food sources

An onlooker bee evaluates the nectar information from all employed bees and chooses a food source depending on the probability value associated with that food source, p_i , given as

$$p_i = \frac{\text{fit}_i}{\sum_{j=0}^{N_p-1} \text{fit}_j} \quad (12)$$

where fit_i is the fitness value of the solution i evaluated by its employed bee. After that, as in case of employed bee, onlooker bee produces a modification on the position in her memory and memorizes the position of better food source only.

In order to find a solution X_i' in the neighborhood of X_i , a solution parameter j and another solution X_k are selected on random basis. Except for the value of chosen parameter j , all other parameter values of X_i' are same as in the solution X_i , for example, $X_i' = (X_{i0}, X_{i1}, \dots, X_{i(j-1)}, X_{ij}', X_{i(j+1)}, \dots, X_{i(D-1)})$. The value of x_{ij}' parameter in X_i' solution is computed as follows:

$$x_{ij}' = x_{ij} + u(x_{ij} - x_{kj}) \quad (13)$$

where u is a uniform random variable in $[-1, 1]$ and k is any number between 0 to N_p-1 but not equal to i .

4.4 Placement of scout bee on the abandoned food source

In the ABC algorithm, if a position cannot be improved further through a predefined number of cycles called 'limit', the food source is abandoned. This abandoned food source is replaced by the scout by randomly producing a position.

After that again steps (B), (C) and (D) will be repeated until the stopping criteria is met.

5 Solving the Constraint Optimization Problem using ABC

In this section we propose a solution to the ligand design using ABC. A potential ligand is encoded by a food source in ABC. In every step of the optimization algorithm, bond length, bond angles and dihedral angles are calculated for

every encoded ligand for fitness evaluation. An algorithm outlining the scheme is discussed below:

Pseudo Code:

Input: Coordinates of active site of receptor target protein P (active_site_P).

Output: Desired ligand structure L for receptor target protein P.

Begin

 Call ABC (active_site_P);

End.

Procedure ABC (active_site_P)

 Begin

 Initialize all the food sources X_i and $\text{trial}_i=0$, for $i=0, 1, \dots, N_p-1$, as in Fig.1 & 3 within active_site_P using fragments from Fig.2 and algorithm parameters like "limit".

 Evaluate the fitness ($\text{fit}(X_i)$) of the population using (11) after decoding X_i .

 For Iter=1 to Maxiter do

 Begin

 For each employed bee do

 Begin

 Produce a new solution X_i' from (13);

 Calculate its fitness value $\text{fit}(X_i')$ using (11) after decoding X_i' as in Fig.3;

 If $\text{fit}(X_i') > \text{fit}(X_i)$ Then $X_i \leftarrow X_i'$; $\text{trial}_i=0$; Else $\text{trial}_i = \text{trial}_i + 1$;

 End If;

 End For;

 End For;

 For each onlooker bee do

 Begin

 Select the food source X_i depending on p_i as in (12);

 Produce new solution X_i' from (13);

 Calculate its fitness value $\text{fit}(X_i')$ using (11) after decoding as X_i' in Fig.3;

 If $\text{fit}(X_i') > \text{fit}(X_i)$ Then $X_i \leftarrow X_i'$; $\text{trial}_i=0$; Else $\text{trial}_i = \text{trial}_i + 1$;

 End If;

 End For;

 Memorize the best solution best_sol obtained so far;

 Set $\text{index} \leftarrow \arg(\max(\text{trial}_i))$;

 If $\text{trial}_{\text{index}} > \text{limit}$ Then reinitialize X_{index} by scout bee;

 End If;

 End For;

 Update: $L \leftarrow \text{best_sol}$;

 Return.

6 Experiments and Results

The experiment was carried out on a simulated environment on Intel Core 2 Duo processor architecture with clock speed of 2GHz using MATLAB. Population size for ABC is taken to be 50 and the algorithm is run for 200 generations. In each generation, each of the food sources is decoded to obtain the corresponding ligand structure. The three dimensional structure of this ligand is obtained using ChemSketch software. The ligand thus designed is docked with the corresponding receptor protein using PatchDock [<http://bioinfo3d.cs.tau.ac.il/PatchDock/>]. Results are taken for different possible positions of the ligand within the active site, and the evolved ligand having the lowest energy value is taken as the solution. The evolvable structures of the ligand for the breast cancer type 1 susceptibility protein (BRCA1) are presented in Fig. 4.

For the experiments, five more different proteins are considered. These are HIV-I Nef, HIV-I Integrase, HIV-I Capsid, HIV-I Protease and thrombin. The active sites conformations of these proteins are obtained from Protein Data Bank [<http://www.rcsb.org/pdb/home/home.do>] and

Active Site Prediction Server [<http://www.scfbio-itt.res.in/dock/ActiveSite.jsp>]. HIV proteins are responsible for the acquired immunodeficiency syndrome (AIDS), a condition in humans in which progressive failure of the immune system allows life-threatening opportunistic infections and cancers to thrive. Thrombin is a serine protease essential for blood coagulation. It hydrolyzes fibrinogen to fibrin for activating platelets to form the clot.

Fig. 5(a)-9(a) show the two-dimensional structure of the ligands evolved using VGA as proposed in [2] for the five different proteins. Fig. 5(b)-9(b) represent the three dimensional structure of the complex obtained in PyMOL after docking of the designed ligand into the active site of corresponding receptor protein molecule.

The two-dimensional structure of ligand molecule evolved using ABC are pictorially represented in Fig. 5(c)-9(c). Fig. 5(d)-9(d) show the three-dimensional geometries of the protein- ligand docked molecules for the corresponding five proteins. Fig. 10(a)-(e) show the active site of the proteins after docking as obtained from PyMOL software. As is evident from the figures, the designed molecules using

ABC are found to fill up the active site reasonably well. For the sake of comparison, the energy values of the ligands (obtained by the VGA based method and ABC based method) as well as those of the ligand-protein complexes are computed and are presented in Tables I and II.

Lower internal energy value of ligands suggests better stability of the ligand. As seen from Table I, in all the cases ABC provides more stable ligands that are associated with lower energy values except for HIV-1 Protease. Significantly lower energy values of the molecules designed using the proposed ABC algorithm indicates more stable receptor ligand complexes as evident from TABLE-II. The ligands designed by the proposed algorithm are generally smaller and comprise of less aromatic groups (Fig. 4(c)-8(c)), causing less steric hindrance and better interaction with the target protein, in comparison to the ligands designed by the other method.

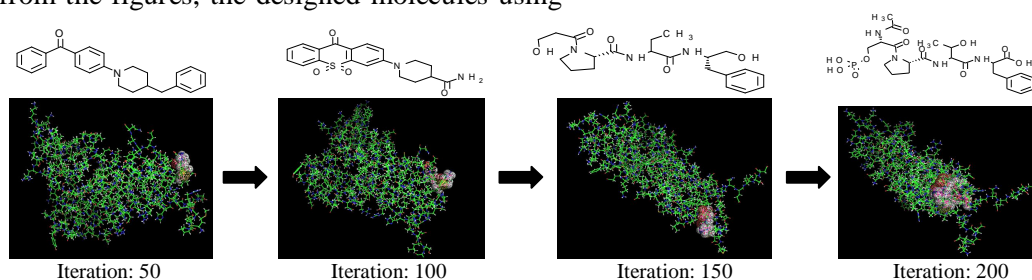


Fig. 4. Evolvable ligand structure for breast cancer type 1 susceptibility protein in each step of ABC- based simulation

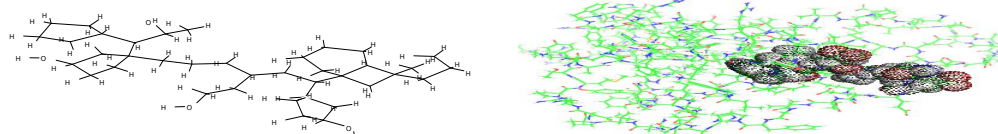


Fig. 5(a) & (b). Using the VGA based method for HIV-1 Nef protein (a) Structure of the evolved ligand molecule (b) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

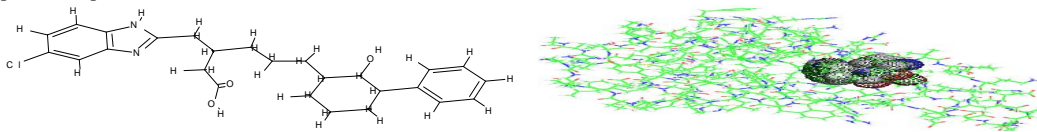


Fig. 5(c) & (d). Using the ABC- based method for HIV-1 Nef protein (c) Structure of the evolved ligand molecule (d) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

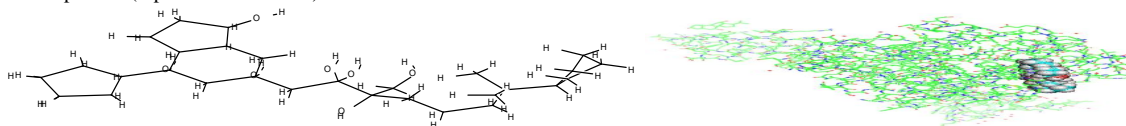


Fig. 6(a) & (b). Using the VGA based method for HIV-1 Integrase protein (a) Structure of the evolved ligand molecule (b) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

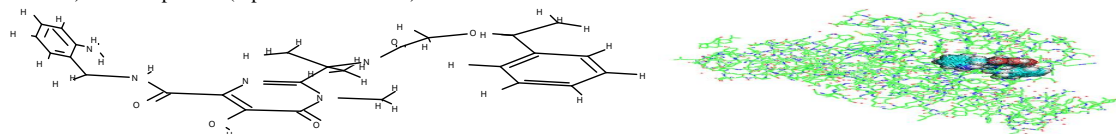


Fig. 6(c) & (d). Using the ABC- based method for HIV-1 Integrase protein (c) Structure of the evolved ligand molecule (d) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

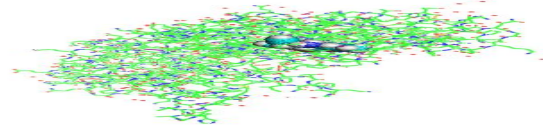
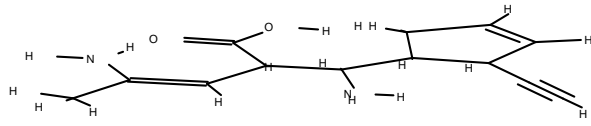


Fig. 7(a) & (b). Using the VGA based method for HIV-1 Capsid protein (a) Structure of the evolved ligand molecule (b) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

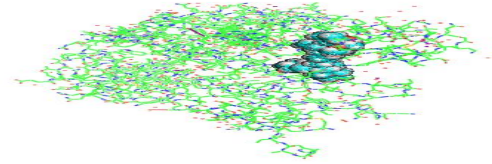
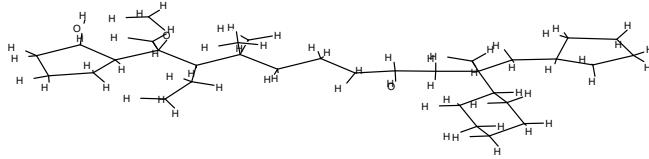


Fig. 7(c) & (d). Using the ABC- based method for HIV-1 Capsid protein (c) Structure of the evolved ligand molecule (d) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

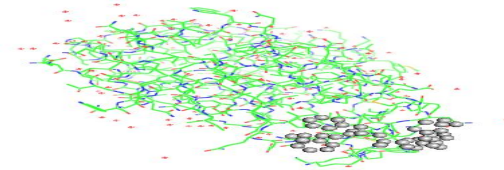
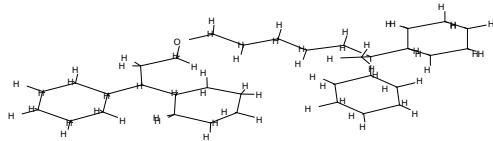


Fig. 8(a) & (b). Using the VGA based method for HIV-1 Protease protein (a) Structure of the evolved ligand molecule (b) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

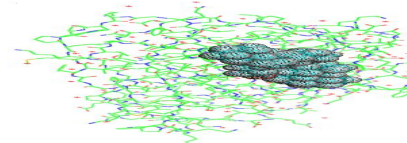
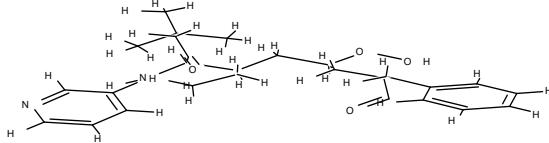


Fig. 8(c) & (d). Using the ABC- based method for HIV-1 Protease protein (c) Structure of the evolved ligand molecule (d) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

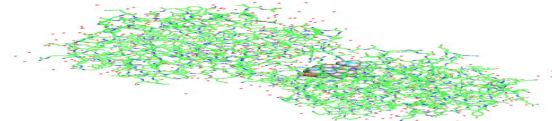
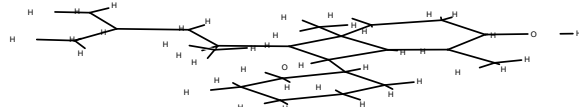


Fig. 9(a) & (b). Using the VGA based method for Thrombin protein (a) Structure of the evolved ligand molecule (b) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

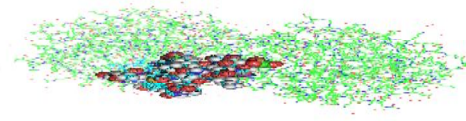
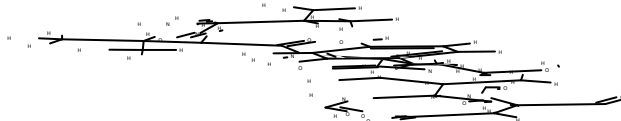


Fig. 9(c) & (d). Using the ABC- based method for Thrombin protein (c) Structure of the evolved ligand molecule (d) Interaction of the ligand (represented as dots) with the protein (represented as sticks)

TABLE-I

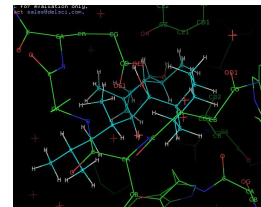
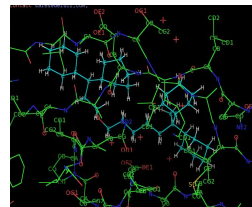
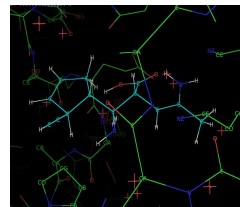
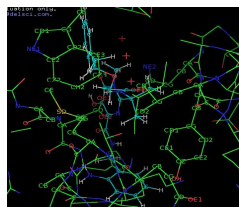
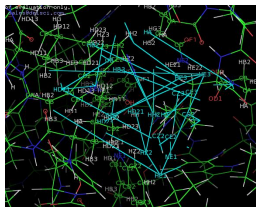
Energy values of the ligands corresponding to target proteins (Kcal/mol)

Process	HIV-1 Nef	HIV-1 Integrase	HIV-1 Capsid	HIV-1 Protease	Thrombin
VGA	4.04	3.35	3.51	2.84	-2.09
ABC	-5.53	-9.23	2.84	4.76	-8.68

TABLE-II

Interaction energy values of the ligands with protein targets(Kcal/mol)

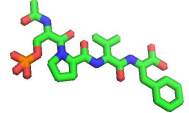
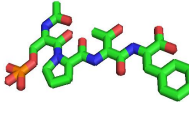
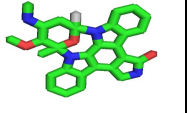
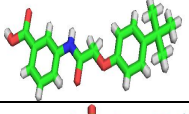
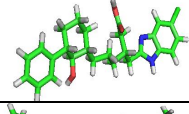
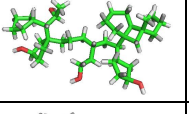
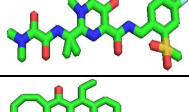
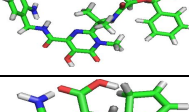
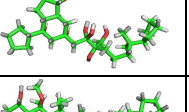
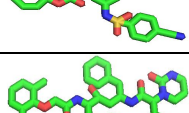
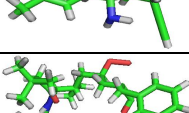
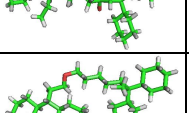
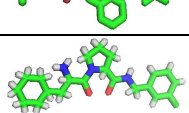
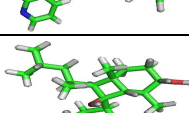
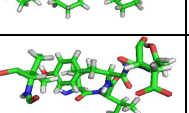
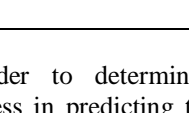
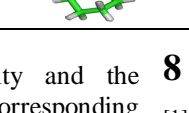
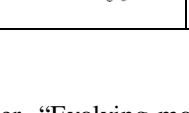
Process	HIV-1 Nef	HIV-1 Integrase	HIV-1 Capsid	HIV-1 Protease	Thrombin
VGA	2.86	3.05	-5.97	2.06	1.97
ABC	-12.27	1.36	-4.01	-10.34	-10.09



(a) (b) (c) (d) (e)

Fig. 10(a) to (e). Interaction between ligand molecule (represented by cyan) within active site after docking using the ABC- based method for (a)HIV-1 Nef (b) HIV-1 Integrase (c) HIV-1 Capsid (d) HIV-1 Protease (e) Thrombin

TABLE-III
Comparison of structures, intra- and inter-molecular energies of ligands obtained from Binding Database, ABC and VGA-based simulations

Proteins	Ligand conformation								
	Binding Database			ABC- based simulation			VGA- based simulation		
	Structure	Intra-molecular	Inter-molecular	Structure	Intra-molecular	Inter-molecular	Structure	Intra-molecular	Inter-molecular
BRCA1		-5.28	-12.48		-2.57	2.08		3.56	2.23
HIV-1 Nef		-10.43	-23.14		-5.53	-12.27		4.04	2.86
HIV-1 Integrase		-15.04	-25.25		-9.23	1.36		3.35	3.05
HIV-1 Capsid		-9.77	-10.82		2.84	-4.01		3.51	-5.97
HIV-1 Protease		-7.76	-21.27		4.76	-10.34		2.84	2.06
Thrombin		-10.81	-41.35		-8.68	-10.09		-2.09	1.97

In order to determine the synthesizability and the correctness in predicting the ligand structure corresponding to the target protein, we refer to the Binding Database (<http://www.bindingdb.org/bind/index.jsp>). The structures of the ligand molecules obtained from the Binding Database, VGA and ABC- based simulations are listed in Table- III along with their corresponding intra and inter-molecular energy values, corresponding to a fixed target protein. As evident, the structures as well as the docking energy values of the ligands designed using ABC are closer to those proposed by the database. This indicates in general the ligands conformations obtained by ABC- based simulation are more stable.

7 Conclusion

A novel method for protein ligand docking using dynamic memory allocation with doubly linked list node representation is proposed. The proposed algorithm is observed to optimize the energy of protein-ligand compound close to the benchmark and also better than that of VGA as verified by experimental results. The proposed technique can be used to provide a powerful exploratory tool for the medicinal synthetic chemist, to evolve molecular structure once the functional protein is given.

8 References

- [1] G. Goh, and J.A. Foster. "Evolving molecules for drug design using genetic algorithm via Molecular Tree". In Int. Conf. Genet. Evol. Comput., pages(27-33), USA 2000.
- [2] S. Bandyopadhyay, A. Bagchi, and U. Maulik. "Active Site Driven Ligand Design: An Evolutionary Approach". Journal of Bioinformatics and Computational Biology. 3(5):1053-1070, October 2005.
- [3] B. Basturk, and Dervis Karaboga. "An Artificial Bee Colony (ABC) Algorithm for Numeric function Optimization". IEEE Swarm Intelligence Symposium 2006, May 12-14, 2006, Indianapolis, Indiana, USA.
- [4] O. Güner. Pharmacophore Perception, "Development and Use in Drug Design". International University Line: La Jolla, CA, 2000.
- [5] Preetha Bhattacharjee, Pratyusha Rakshit, Indrani Goswami, Amit Konar, Atulya K. Nagar. "Multi-robot path-planning using artificial bee colony optimization algorithm". NaBIC 2011: 219-224.
- [6] N Moitessier, P Englebienne, D Lee, J Lawandi, and C R Corbeil. "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go". Br J Pharmacol. 2008 March; 153(S1): S7-S26. Published: 2007 Nov 26.
- [7] Datta Ayan, Talukdar Veer, Konar Amit and Jain Lakshmi C. "Neuro-swarm hybridization for protein tertiary structure prediction". International Journal of Hybrid Intelligent Systems 5 (2008) 153-159.
- [8] Protein-Ligand Docking: "A Critical Review of Molecular Dynamics". Robotics, and Rotamer Library Methods. -by Thomas Butler.

Protein Query Language: A Novel Approach

Sherif Elfayoumy and Paul Bathen

School of Computing, University of North Florida, Jacksonville, FL, USA

Abstract - This paper introduces a Protein Query Language (PQL) for querying protein structures in an expressive yet concise manner, utilizing the work of Patel [1] and introducing constructs in principal similar to those in Roldan-Garcia [2]. One of the objectives of the paper is to demonstrate how such a language would be beneficial to protein researchers to obtain in-depth protein data from a relational database without extensive SQL knowledge. The language features options such as limiting query results by key protein characteristics such as methyl donated hydrogen bond interactions, minimum and maximum phi and psi angles, repulsive forces, CH/Pi calculations, and other pertinent factors. In addition, front end applications can be developed to support retrieving, transforming, and preprocessing of information from the Research Collaboratory for Structural Bioinformatics (RCSB) [3] into the backend data repository.

Keywords: Proteomics; Bioinformatics; Query Languages; Protein Query Language.

1 Background

One of the major challenges facing biology and biochemistry researchers is the ability to view relationships among protein data, structures, functions, and pathways in a single query or at least in a concise and expressive manner [4]. For example, biochemists are performing cutting edge research into carbon-donated hydrogen bonds and their effect on protein structures [5]. To do so, they require data at the atomic level of the protein to perform calculations such as determining methyl-donated hydrogen bonds, repulsive forces, and CH/Pi interactions. Yet no online database is known to exist which supplies experimental data in an easy-to-use format at the atomic level without parsing the data manually, nor do tools exist to facilitate the calculations once data is parsed. To support their research, chemists have been downloading files from the RCSB in Protein Data Bank (.pdb) format, parsing data manually, and loading data into spreadsheets to perform calculations. This approach is tedious and potentially error prone, and spreadsheet limitations as well as other limiting factors obviate the need for a more efficient solution. For example, it is complicated in spreadsheets to answer the question "find all 'acceptor' atoms (i.e., oxygen, nitrogen, sulfur, or carbon) in a given model and chain of a protein within +/- 5 angstroms of a hydrogen atom which is considered potential methyl-donated hydrogen, and calculate the distance between the two atoms." In

addition, the number of atoms alone in large proteins may not fit within older spreadsheet program row limitations.

1.1 Dataset Repositories

Research chemists around the world do have access to various public protein data sources, but the access is not designed to support processing and retrieval at the atomic level. Online 'databases' supporting biochemistry research include Genbank, EMBL Data Library in the UK, the DNA Data Bank of Japan (DDBJ), and COLUMBA [6]. In essence, the only known public access to these databases is via a supplied front-end, and the returned data is formatted for user reading rather than for storing the data into a database for further processing and analysis.

Genbank provides meta level information about research performed on a given protein. Researchers can view what chains have been investigated, and download the meta data in a variety of formats including variations of XML. Researchers may view sequence alignments, and view pictures of what the protein chain looks like. However, this meta level is of insufficient detail for performing the atomic level research required.

Protein research data can be obtained via sites like the EMBL Data Library in the UK and the RCSB. The protein data is submitted to the RCSB by research investigators performing atom-level X-Ray crystallography, Nuclear Magnetic Resonance (NMR), and other types of studies. From these sites, researchers can download text files in a variety of formats (.pdb, .mmCIF, .xml and others) containing the detailed information required for research, yet the researcher has to parse the files to obtain the required detail data. Doing so for multiple files is a laborious process. In addition, error checking within the file and across multiple files for conditions which would preclude the researcher from using the protein is again a laborious process. These error conditions include but are not limited to: 1) atoms too close together in a given residue based on the van der Waals radius and/or in comparison to the average distance from one atom to another over multiple proteins; 2) atoms too close together across residues within a given protein under similar considerations as #1. In this research relational queries against a preliminary data model have already been written to find or confirm several active protein files on the RCSB where data is in an obvious error state.

Trissl presents a high level data model [6] as seen in Figure 1. This includes data from PDB, KEGG, SWISSPROT, CATH, SCOP, and others.

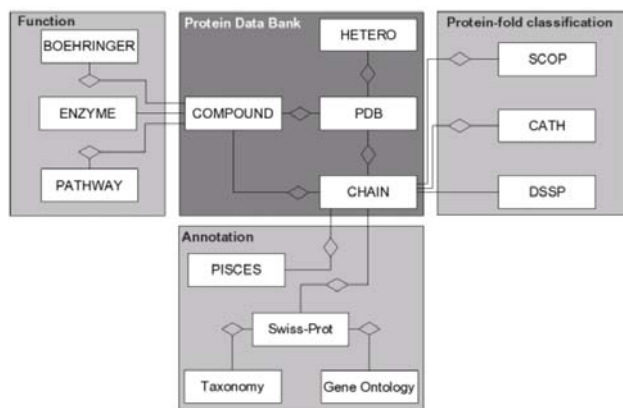


Figure 1: COLUMBA high-level schema

BioMolQuest [7] and iProt [4] both discuss importing protein data at the atomic level into a relational database. Also, Pryor and Fetrow discuss a relational database named PDB-SQL built in MySQL to handle protein-specific data at the atomic level [8]. The high level data model of PDB-SQL can be seen in Figures 2 and 3. In addition, this application was built for a specific purpose and had limited support for atom level data. Pryor proposed an extension to this model with fifty one (51) atom-level tables, one for each type of atom that may be included in a protein file.

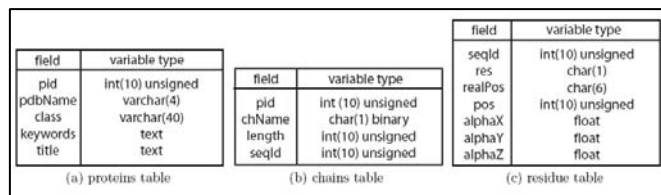


Figure 2: Base schema. The residue table would store the only atom-level detail, and only on carbon atoms.

Eltabakh et al discussed an extensible database engine for biological databases [9][10]. The proposed engine “extends the functionalities of current DBMSs with (1) annotation and provenance management including storage, indexing, manipulation, and querying of annotation and provenance as first class objects in dbms, (2) local dependency tracking to track the dependencies and derivations among data items, (3) update authorization to support data curation via content-based authorization, in contrast to identity-based authorization, and (4) new access methods and their supporting operators that support pattern matching on various types of compressed biological data types.” While interesting, the focus of these researches is on annotation and provenance tracking, and is not the primary focus of this paper. However

the concepts of annotation and provenance should be kept in mind during database design and maintenance. As demonstrated, there have been multiple attempts to import detailed protein data into relational databases. Yet the resulting database and data is either hidden by a front-end interface or is not available to the general public.

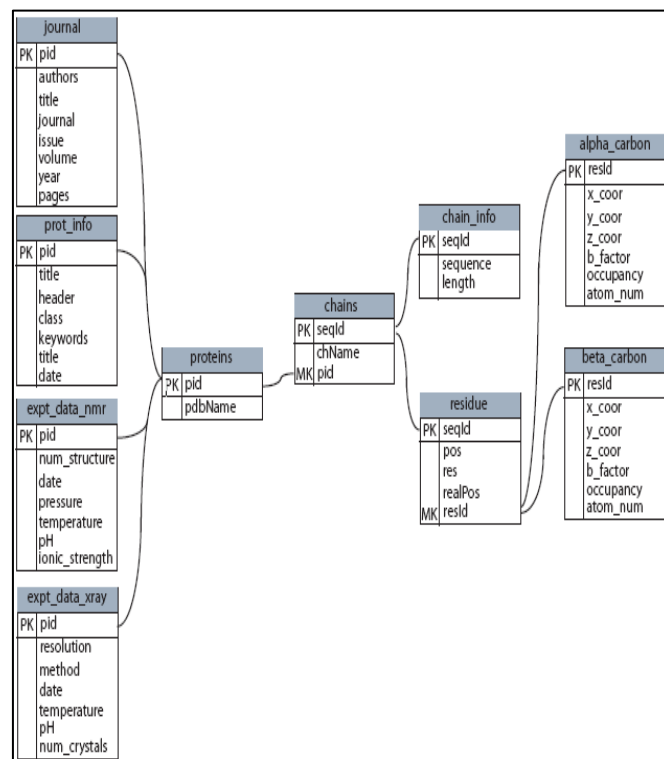


Figure 3: Extended schema: 'alpha_carbon' and 'beta-carbon' are representatives of the 51 atom-level tables.

1.2 Protein Data Parsing

As mentioned above, protein data obtained experimentally does exist and can be downloaded in .pdb, .mmCIF, and .xml formats. The .pdb format is designed for easy human reading. The .mmCIF and .XML formats are more structured in design, and use a data dictionary infrastructure [11]. These files may be downloaded from the RCSB via a web interface or via FTP. Software tools to parse the files include, but are not limited to, BioJava, BioPython, and CIFPARSE-OBJ. Software tools to facilitate loading the data into relational databases include but are not limited to BioJava, BioSql, and “Db Loader”. A list of tools is available at the RCSB site [12] and the Bioinformatics Link Directory (Biolinks). BioJava was chosen to parse .pdb files. More detail on this approach, including strengths and weaknesses, is expanded upon later.

1.3 Query Tools

Online tools exist to find primary and secondary protein sequences, and to compare related primary sequences across

proteins [13][14]. However, there are no known tools which support querying the secondary structure of proteins, one of the goals of this research. And, again, there are no known tools for querying data at the atomic level.

The current dearth of tools for querying protein data can be likened to the banking industry some thirty years ago, where procedural rather than declarative code was written to access data. This made for fragile code from an architectural standpoint. This research provides a declarative query language against standard relational databases and investigates the strengths and impedances of this approach in consideration of the vast data sets involved and the various query types that may be necessary.

Patel describes a declarative protein secondary structure query language as well as an efficient implementation using histograms and optimization [1]. Since secondary structures at their simplest form can be described using an alphabet of three character (h=helices, e=beta-sheets, and l=turns or loops), a secondary structure for a given protein might be "eeeeeeellllllh." Patel suggests expressing the secondary structure sequences as a series of segments. The above example secondary structure would be represented as nine (9) e's, five (5) l's, and two (2) h's. Patel then suggests a query language based on a triplet predicate of the form $\langle type, min\ length, max\ length \rangle$ where *type* refers to whether the segment is a helix, a beta-sheet, or a loop/turn, and the *min length* and *max length* refer to the length of the segment being queried. In addition, the *type* can be a wildcard. So, for example, a query such as $\langle e\ 8\ 10 \rangle \langle ?\ 3\ 5 \rangle \langle h\ 2\ 2 \rangle$ would match the example, since the first segment is of type 'e' and has length 9 (between 8 and 10), the second segment type matches the wildcard and is of length 5 (between 3 and 5 units long, inclusive), and the last segment matches the type 'h' and is 2 units long. The query would then be translated into SQL and executed against the database. Assuming the full sequence data is stored in the table *protTbl* and the segment data is stored in the table *segTbl*, the equivalent SQL code would be:

```
SELECT *
From protTbl p, segTbl s1, segTbl s2
WHERE s1.type = 'e'
AND s1.length BETWEEN 8 AND 10
AND s2.type = 'h'
AND s2.length = 2
AND s1.id = s2.id
AND s1.id = p.id
AND s2.start pos-(s1.start pos+s1.length)<=5
AND s2.start pos-(s1.start pos+s1.length)>=3
```

Note how the second segment in the query needs to be written as a relation between the first and third segment. For anyone unfamiliar with SQL, writing such a query might be a daunting task, and subject to error. As such, the proposed query language is an elegant way for researchers to query secondary structures of proteins. The language as developed, however, does not extend into the atomic level. In addition, this language does not operate against multiple columns in the same table or across multiple tables.

Although the subject area is the Semantic Web instead of a relational database, Roldan-Garcia proposed an interesting logic-based language named Extended Conjunctive Queries (ECQ) [2]. An example stated in the paper is:

```
ans(?x,?y)←fullprofessor(?x)
OR assistantprofessor(?x)
AND worksfor(?x,%university%)
AND >=3 teacherof(?x,?y) AND
ALL course(?y)
```

which after processing would translate to the SQL query:

```
SELECT distinct u1.url, u2.url
FROM uri index u1, uri index u2, worksfor 1 p w1,
teacherof 1 p t1, course 1 c c1
WHERE u2.id=t1.object and u1.id=w1.subject
AND w1.subject=t1.subject AND (w1.subject in
(SELECT url FROM fullprofessor 1 c) OR
w1.subject IN (SELECT url FROM
assistantprofessor 1 c )) and
t1.object=c1.id w1.object in (SELECT ua.id
FROM uri index ua, worksfor 12 p w1a WHERE
ua.id=w1a.object AND ua.url LIKE
'%university%') AND t1.subject IN (SELECT
subject FROM teacherof 1 p GROUP BY subject
HAVING COUNT(DISTINCT object) >=3 ) AND
t1.object=c1.url AND t1.subject IN (SELECT
t1a.subject FROM teacherof 1 p t1a GROUP BY
t1a.subject HAVING COUNT(DISTINCT
object)=(SELECT count (DISTINCT object) FROM
teacherof 1 p WHERE subject=t1a.subject AND
object IN (SELECT url FROM course 1 c) GROUP
BY subject)) ORDER BY u1.url
```

An ECQ expression has the form:

$$\text{ans}(V_1, V_2, V_3, \dots, V_n) \leftarrow Q_1 \text{ AND } Q_2 \text{ AND } \dots \text{ AND } Q_n$$

where each Q_i can take the form:

1. $C(x)$
2. $P(x, y)$
3. $C(x) \text{ OR } D(x)$
4. $\text{ALL } C(x)$
5. $\leq n\ P(x, y)$
6. $\geq n\ P(x, y)$
7. $= n\ P(x, y)$

where C and D are class names, P is a property name, x and y are instance names or variables, and n is a natural number. The simplicity of ECQ's approach may be useful in the target query language.

Another language with interesting features is TQL, proposed by Conforti et al [15]. Again, TQL is a language for semi-structured data that can be used to query XML, but is built on set comprehension in the tradition of SQL and other languages. An example stated in the paper of a query in TQL would be:

```
FROM $Bib |= .bib[.book[.year[1991] And .title[$t]]]
SELECT title[$t]
```

which should be read: "there is a path *.bib[.book[]]* that reaches a place that matches *.year[1991] And .title[\$t]*, i.e. a

place where you find both a path `.year[]` leading to 1991 and a path `.title[]` leading to something, that you will call `$t`".

2 PQL

The Protein Query Language (PQL) is declarative in nature. Users of the language have access to the following features:

1. Users may utilize familiar terms when referring to proteins, models, chains, residues, atoms, and other chemistry terms. The underlying relational model is abstracted from the user.
2. The ability to use mathematical, boolean, and string functions as part of the language. However, constructs such as conditionals and looping are supported at this time.
3. The user shall be able to save PQL constructs for later utilization.

2.1 Grammar

The grammar for the PQL was developed in Backus-Naur Form (BNF) using the grammar constructs provided within a software package named Gold-Parser Builder. The SQL statement was used to calculate the potential methyl-donated hydrogen bonds for a given protein, and therefore represents a practical example in biochemistry research. It can be easily seen the PQL representation of this calculation may be much easier for a non-SQL expert to develop. Further explanation of the grammar follows below.

The grammar is divided into five (5) distinct areas, two (2) of which are required and three (3) of which are optional as described here:

1. **EQUIVALENCE** (optional): An example best illustrates the use of an equivalence statement. Say, for example, in some portions of their query a user would like to reference a hydrogen atom as 'h' for brevity, whereas in other sections it might be more instructive to reference that same hydrogen atom as 'hydrogenAtom'. A user may add the equivalence statement 'h hydrogenAtom' with the semantics 'h is the same object as hydrogenAtom'.
2. **INSTANCE**: At least one statement required. An instance statement allows a user to tie an instance variable to a 'table', or in user-terms a group of chemically related items. For example, the statement:

```
Protein(p).Model(?).Chain(?).Residue(r).Atom(a)
```

allows the user to tie the instance variable 'p' to a protein structure, 'r' to a residue structure, and 'a' to an atom structure. In the ASSIGNMENT and CONSTRAINT sections, the user can place stipulations on how these instance variables are bound. The '?' variables are used as wildcards. In addition, multiple statements using the same instance variable tie statements together. For example:

```
Protein(p).Model(m).Chain(c).Residue(r).Atom(c)
Protein(p).Model(m).Chain(c).Residue(r).Atom(h)
```

means in essence that atom instance variables 'c' and 'h' share the same protein, model, chain, and residue. Again, in the ASSIGNMENT and CONSTRAINT sections the user might further restrict 'c' to be a carbon atom, and 'h' to be a hydrogen atom. 'Fields' within the instance variable can then be accessed in the ASSIGNMENT, CONSTRAINT, and RESULTS sections. For example, an atom has an atom name and potentially X, Y, and Z coordinates (if it has 3-D data associated with it). In the above example, these fields within the hydrogen structure would be accessed as `h.atomName`, `h.xcoor`, `h.ycoor`, `h.zcoor`. The user would have a list of the accessible fields per structure.

3. **ASSIGNMENT** (optional): An assignment statement takes the form:

```
a. thetaAngle=ThetaAngle(chDist,cxDist,hxDist)
b. tempName = StringAdd('C',
    Substring(h.atomName,2, Len(h.atomName)-
    2),'%')
c. s = A AND B OR C
d. h.atomName = "H11"
```

Assignment can be made to a temporary variable (e.g., `tempName`) or to an instance variable's field (e.g., `h.atomName`). Assignments can include combinations of boolean, string, and mathematical expressions.

4. **CONSTRAINT** (optional): The user can constrain certain conditions on the resultant returned data. Examples of constraint statements include:

```
a. cxDist >= 4.2
b. thetaAngle BETWEEN 150.0 AND 210.0F
c. carbonAtom.atomName LIKE ('C' +
    Substring(h.atomName,2,Len(h.atomName)-2) +
    '%')
d. carbonAtom.atomName LIKE "CH11"
```

Constraints can include combinations of boolean, string, and mathematical expressions.

5. **RESULTS**: At least one statement is required. Result statements are where users specify what the returned dataset looks like and how it is sorted. Typical statements look like:

```
a. h.atomName ASC 1
b. tempString as carbonAtomName DESC 2
c. tempString2 DESC 3 NO OUTPUT
```

User can specify an instance variable field (e.g., `h.atomName`) or a temporary variable (e.g., `tempString`) as well as an optional output name for that variable (e.g., 'carbonAtomName' above) and an ascending or descending order (e.g., the 'ASC x' and 'DESC y' portions of the statements above). The result set is returned in the column order specified line-by-line, and sorted in the order

specified by the ASC (or ASCENDING) and DESC (or DESCENDING) sub-statements. Lastly, user can specify 'NO OUTPUT' to restrict a given Result statement from the output.

Users also have access to useful 'method' calls in the ASSIGNMENT and CONSTRAINT sections including methods such as:

1. Distance(atom1,atom2)
2. Distance(xcoor1,xcoor2,ycoor1,ycoor2,zcoor,zcoor2)
3. ThetaAngle(distance1to3,distance1to2,distance2to3)

3 Conclusion

As detailed above, the vast preponderance of computational tools available to protein researchers seem to concentrate on predictions at the amino acid residue level, including prediction of the secondary state. Important biochemistry research is being done at the atom level, yet little or no computing tools are publicly available to biochemists to support this work. The PQL language is an attempt to provide an intuitive declarative language within query application to researchers who are unfamiliar with SQL coding. The PQL query system allows users to interrogate a relational database containing protein data downloaded from the RCSB Protein Data Bank. Users can create queries to identify important research interactions between methyl-donated hydrogen bonds, amine repulsions, and CH/Pi interactions. We expect users of the new system to gain significant insight into research areas such as the tertiary structure of proteins.

4 References

- [1] J.M. Patel, D.P. Huddler, and L. Hammel. "Declarative and Efficient Querying on Protein Secondary Structures"; Data Mining in Bioinformatics, pp. 243-273, 2005.
- [2] M.M. Roldan-Garcia, J.J. Molina-Castro, and J.F. Aldana-Montes. "ECQ: A Simple Query Language for the Semantic Web"; Proceedings of DEXA, Turin, Italy, 2008.
- [3] Research Collaboratory for Structural Bioinformatics (RCSB). <http://www.rcsb.org>
- [4] M.I. Jaya, Z. Zainol, and N.H. Malim. "iProt – A Data Warehouse for Protein Database"; International Conference on Electrical Engineering and Informatics (ICEEI2007), Institut Teknologi Bandung, Indonesia, 2007.
- [5] K. Compaan, R. Vergenz, P. Von Rague Schleyer, and I. Arreguin. "Carbon-donated Hydrogen Bonding: Electrostatics, Frequency Shifts, Directionality, and Bifurcation"; International Journal of Quantum Chemistry, Vol. 108, No. 15, 2914–2923, 2008.

[6] S. Trissl, K. Rother, H. Mueller, T. Steinke, I. Koch, R. Preissner, C. Froemmel, and U. Leser. "Columba: an Integrated Database of Proteins, Structures, and Annotations"; BMC Bioinformatics, Vol. 8, No. 81, 2005.

[7] Y.V. Bukhman, and J. Skolnick. "BioMolQuest: Integrated Database-based Retrieval of Protein Structural and Functional Information"; Bioinformatics. Vol. 17, No. 5, 468-478, 2001.

[8] E.E. Pryor Jr., and J.S. Fetrow. "PDB-SQL: a Storage Engine for Macromolecular Data"; Proceedings of the 45th Annual Southeast Regional Conference, 2007.

[9] M.Y. Eltabakh, M. Ouzzani, M. Aref, "BDBMS- A Database Management System for Biological Data"; Third Biennial Conference on Innovative Data Systems Research (CIDR), 2007.

[10] M.Y. Eltabakh, M. Ouzzani, M. Aref, A.K. Elmagarmid, Y. Laura-Silva, M.U. Arshad, D. Salt, and I. Baxter. "Managing Biological Data Using BDBMS"; The IEEE 24th International Conference on Data Engineering, 2008.

[11] J. Westbrook, N. Ito, H. Nakamura, K. Henrick, and H.M. Berman. "PDBML: the Representation of Archival Macromolecular Structure Data in XML"; Bioinformatics, Vol. 21, No. 7, 988-992, 2005.

[12] RcsbSoftwareTools: <http://sw-tools.pdb.org/>

[13] BLAST: <http://blast.ncbi.nlm.nih.gov>

[14] FASTA: <http://fasta.bioch.virginia.edu/>

[15] G. Conforti, G. Ghelli, A. Albano, D. Colazzo, P. Manghi, and C. Sartiani. "The Query Language TQL"; The 5th International Workshop on Web and Databases (WebDB), 2002.

Purifying and Filtering the Coupling Matrix Approach for Protein-Protein Interaction Network Analysis

Ying Liu

¹Department of Mathematics and Information Sciences, University of North Texas at Dallas, Dallas, TX

Abstract - One of the most pressing problems of the post genomic era is identifying protein functions. Clustering Protein-Protein-Interaction networks is a systems biological approach to this problem. Traditional Graph Clustering Methods are crisp, and allow only membership of each node in at most one cluster. However, most real world networks contain overlapping clusters. Recently the need for scalable, accurate and efficient overlapping graph clustering methods has been recognized and various soft (overlapping) graph clustering methods have been proposed. In this paper, an efficient, novel, and fast overlapping clustering method is proposed based on purifying and filtering the coupling matrix (PFC). PFC is tested on PPI networks. The experimental results show that PFC method outperforms many existing methods by a few orders of magnitude in terms of average statistical (hypergeometrical) confidence regarding biological enrichment of the identified clusters.

Keywords: Protein-Protein Interaction networks; Graph Clustering; Overlapping functional modules; Coupling Matrix; Systems biology

1 Introduction

Homology based approaches have been the traditional bioinformatics approach to the problem of protein function identification. Variations of tools like BLAST [1] and Clustal [2] and concepts like COGs (Clusters of orthologous Groups) [3] have been applied to infer the function of a protein or the encoding gene from the known a closely related gene or protein in a closely related species. Although very useful, this approach has some serious limitations. For many proteins, no characterized homologs exist. Furthermore, form does not always determine function, and the closest hit returned by heuristic oriented sequence alignment tools is not always the closest relative or the best functional counterpart. Phenomena like Horizontal Gene Transfer complicate matters additionally. Last but not least, most biological Functions are achieved by collaboration of many different proteins and a proteins function is often context sensitive, depending on presence or absence of certain interaction partners.

A Systems Biology Approach to the problem aims at identifying functional modules (groups of closely cooperating

and physically interacting cellular components that achieve a common biological function) or protein complexes by identifying network communities (groups of densely connected nodes in PPI networks). This involves clustering of PPI-networks as a main step. Once communities are detected, a hypergeometrical p-value is computed for each cluster and each biological function to evaluate the biological relevance of the clusters. Research on network clustering has focused for the most part on crisp clustering. However, many real world functional modules overlap. The present paper introduces a new simple soft clustering method for which the biological enrichment of the identified clusters seem to have in average somewhat better confidence values than current soft clustering methods.

2 Previous Work

Examples for crisp clustering methods include HCS [4], RNSC [5] and SPC [6]. More recently, soft or overlapping network clustering methods have evolved. The importance of soft clustering methods was first discussed in [7], the same group of authors also developed one of the first soft clustering algorithms for soft clustering, Clique Percolation Method or CPM [8]. An implementation of CPM, called CFinder [9] is available online. The CPM approach is basically based on the “defective cliques” idea and has received some much deserved attention. Another soft clustering tool is Chinese Whisper [10] with origins in Natural Language Processing. According to its author, CW can be seen as a special case of the Random Walks based method Markov-Chain-Clustering (MCL) [11] with an aggressive pruning strategy.

Recently, some authors [12, 13] have proposed and implemented betweenness based [14] Clustering (NG) method, which makes NG’s divisive hierarchical approach capable of identifying overlapping clusters. NG’s method finds communities by edge removal. The modifications involve node removal or node splitting. The decisions about which edges to remove and which nodes to split, are based on iterated all pair shortest path calculations.

In this paper, we present a new approach, called PFC, which is based on the notion of Coupling matrix (or common neighbors). In the rest of the paper, we first describe PFC and compare its results with the best results achieved by the aforementioned soft approaches. The second part of this work

aims to illustrate the biological relevance of soft methods by giving several examples of how the biological functions of overlap nodes relate to biological functions of respective clusters.

3 PFC Method

The method introduced here is based on the purification and filtering of coupling matrix, PFC. PFC is a soft graph clustering method that involves only a few matrix multiplications/ manipulation. Our experimental results show that it outperforms the above mentioned methods in terms of the p-values for MIPS functional enrichment [15] of the identified clusters. The PPI networks we used in the paper are yeast PPI networks (4873 proteins and 17200 interactions).

Liu and Foroushani [16] proposed aPFC 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 comprise huge sets that are less likely to be all functionally related. Additionally, it has been shown that most functional modules are meso-scale [6]. There are also some fundamental physical 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 $_i$ th node in the $_i$ th row) has a particularly low (< 14) or particularly high (> 30) degree. Discard all clusters where the module size is too small (< 35) or particularly large (> 65).

The selected minimum and maximum values for degree of labeling nodes and module size are heuristically motivated.

The intervals can be easily changed to obtain or discard more clusters, but the enrichment results for these intervals seem reasonably good. The peak log value for the enrichment of selected clusters is at -91.00 and the average lies at -18.99 . Using this filter, by clustering yeast PPI networks, PFC yields 151 clusters from 52 different Functional categories. Figure 1 gives an example.

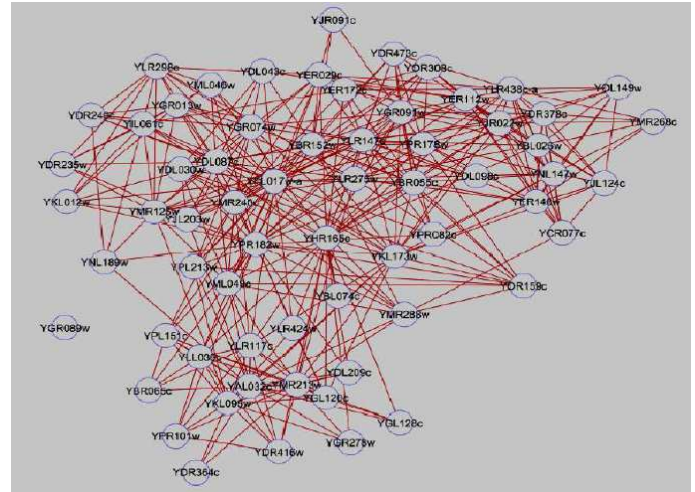


Figure 1 This Figure shows the community for the row labeled “YKL173w” in the purified coupling matrix of yeast PPI network. It is one of the clusters selected by PFC1. Out of the 63 proteins in this community, 58 belong to MIPS Funcat 11.04.03.01.

3.2 Filtering by Corroboration

Filtering by local criteria gives impressive 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 by a Line B if the majority of nonzero elements in A are also nonzero in B. The following summarizes this filter:

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.

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.

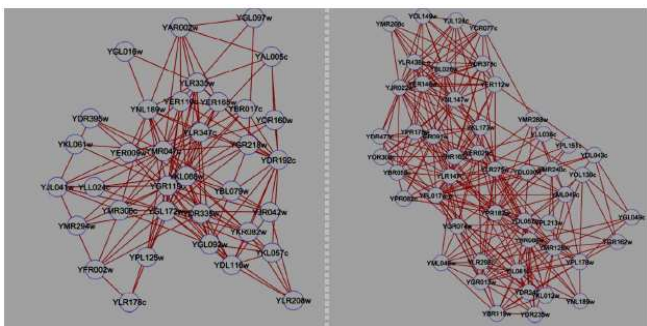


Figure 1: Two of the clusters selected by PFC2. The left Figure shows the selected community for the row labeled “YDR335w” in the purified coupling matrix. Out of the 35 proteins in this community, 29 belong to MIPS Funcat 20.09.01(nuclear transport). The right Figure shows the selected community for 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 is a clustering results of the PFC. The result demonstrates an important *advantage* of soft methods against crisp ones: They show how soft clustering can adequately mirror the fact that many proteins have context dependent functions, and how in some cases overlap nodes can act as functional bridges between different modules.

Table 1 Comparison of results from different methods

Method	Cluster Count	Average Cluster Size	Average Enrichment	Network Coverage	Diversity
Betweenness based	20	302.70	-15.11	0.58	19/20
Chinese Whisper	38	23.45	-12.11	0.17	32/38
C Finder	68	14.50	-15.70	0.19	48/68
PFC1	183	44.76	-19.35	0.31	55/183
PFC1	40	25.4	-19.40	0.17	36/40

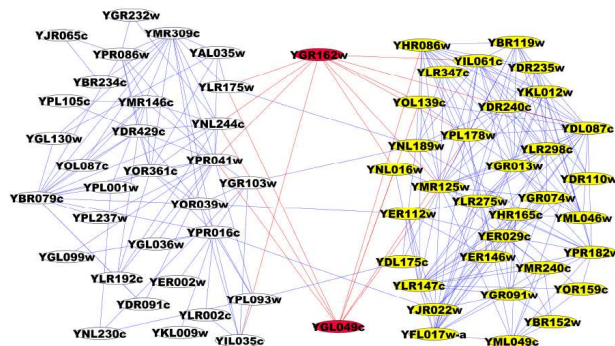


Figure 3. result #1: The dominant function for the left module is translation initiation (10 out of 31) for the right module, it is nuclear mRNA splicing (27 out of 33); both overlap nodes are involved in translation initiation and Protein-RNA complex assembly.

5 Conclusions

This paper introduced PFC, a new clustering concept based on purification and filtering of a coupling (common neighbor) matrix. It discussed a very different filtering method. PFC consists of only a few matrix multiplications and manipulations and is therefore very efficient. The PFC outperforms current soft clustering methods on PPI networks by a few orders of magnitude in terms of average statistical confidence on biological enrichment of the identified clusters. The paper illustrated the importance of soft clustering methods in systems biology by giving a few concrete examples of how the biological function of the overlap nodes relates to the functions of the respective clusters.

6 References

- [1] Altschul, SF, et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic acids research* 25, no. 17: 3389, 1997.
- [2] Thompson, JD, DG Higgins, and TJ Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic acids research* 22, no. 22: 4673-4680, 1994
- [3] Tatusov, R. L., E. V. Koonin, and D. J. Lipman. "A genomic perspective on protein families". *Science* 278, no. 5338: 631, 1997.
- [4] Hartuv, E., R. Shamir. "A clustering algorithm based on graph connectivity". *Information processing letters* 76, no. 4-6: 175-181, 2000.
- [5] King, A. D., N. Przulj, and I. Jurisica. "Protein complex prediction via cost-based clustering". *Bioinformatics* 20,; 3013-3020, 2004.
- [6] Spirin, V., L. A. Mirny. "Protein complexes and functional modules in molecular networks". *Proceedings of the National Academy of Sciences* 100, no. 21: 12123-12128, 2003.
- [7] Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society". *Nature* 435, no. 7043 (Jun 9): 814-818, 2005.
- [8] Derenyi, I., et al. "Clique percolation in random networks". *Physical Review Letters* 94, no. 16: 160202, 2005.
- [9] Adamcsek, B., G. et al. "CFinder: locating cliques and overlapping modules in biological networks". *Bioinformatics* 22, no. 8: 1021-1023, 2006.
- [10] Biemann, C. "Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems". In *Proceedings of the HLT-NAACL-06 workshop on textgraphs-06*, new york, USA, 2006.
- [11] Van Dongen, S. "A cluster algorithm for graphs". *Report- Information systems* , no. 10: 1-40, 2000.
- [12] Pinney, J. W., D. R. Westhead. "Betweenness-based decomposition methods for social and biological networks". In *Interdisciplinary statistics and bioinformatics*. Edited by S. Barber, P. D. Baxter, K. V. Mardia and R. E. Walls. Leeds University Press, 2000.
- [13] Gregory, S. "An algorithm to find overlapping community structure in networks". *Lecture Notes in Computer Science* 4702: 91, 2007.
- [14] Girvan, M., M. E. Newman. "Community structure in social and biological networks". *PNAS* 99: 7821-7826, 2002.
- [16] Chua, H. N. et al. "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions". *Bioinformatics* 22: 1623-1630, 2006.
- [15] MIPS. The functional catalogue (FunCat). 2007. <<http://mips.gsf.de/projects/funecat>>.
- [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.

Bioinformatics analysis identify novel OB fold protein coding genes in *C. elegans*

Daryanaz Dargahi, Dave Baillie, *Frederic Pio

Simon Fraser University, Molecular Biology & Biochemistry Department, 8888 University Dr, V5A1S6. Corresponding (contact) author (*): Frederic Pio: fpio@sfu.ca, co-author email: ddargahi@sfu.ca, Baillie@sfu.ca

Keywords: *C. elegans* genome annotation, comparative modeling, OB fold, Hidden Markov Model,

Abstract:

Background

The *C. elegans* genome has been extensively annotated by the wormbase consortium that uses state of the art bioinformatics pipelines, functional genomics and manual curation approaches, as a result the identification of novel genes *in silico* in this model organism is becoming more challenging and require novel approaches. The oligonucleotide-oligosaccharide binding fold is a highly divergent fold where proteins sequences of the family in spite of having the same fold share very little sequence identities (5-25%). Therefore, sequence based annotation evidences may not be sufficient to identify all the members of this highly divergent family. In *C. elegans* the number of OB fold proteins reported is remarkably low (n=46) compared to other evolutionary related eukaryotes such as yeast *S. Cerevisiae* (n=344) or fruit fly *D. melanogaster* (n=84). Genomics rearrangements during evolution may have occurred or differences in the level of annotation for this protein family may explain these discrepancies.

Methodology/Principal Findings

This study examines the possibility that novel OB fold coding genes exist in the worm. We developed a bioinformatics approach that uses the most sensitive sequence-sequence, sequence-profile and profile-profile similarity searches methods followed by OB-fold 3D-structure prediction as a filtering step to eliminate false positive candidate sequences. We have predicted 18 coding gene containing the OB-fold. Remarkably, most of their corresponding genes have not or partially been characterized in *C. elegans*.

Conclusions/Significance

Further study of the function of these novel candidates is critical to enhance our understanding of the biology of this family. This study raises the possibility that the annotation of highly divergent protein fold families can be improved in *C. elegans*. Similar strategies could be implemented for large scale analysis by the wormbase consortium when novel build and version of the genome sequence of *C. elegans* or other evolutionary related species are being released.

Introduction

Bioinformatics analysis of the complete genome sequence of *C. elegans* by the wormbase consortium initially revealed over 19000 coding genes [1]. When the genome of *C. elegans* closely related species *C. briggsae* was sequenced and comparative analysis was performed between both species 6% more coding genes were predicted (20261 coding genes) [2]. Since bioinformatics annotation pipeline from the wormbase consortium are evolving new protein-coding genes are constantly being predicted. As such, the latest version of the *C. elegans* genome sequence (WS228) predict (24610) coding genes [3] which may indicates that novel protein-coding genes remains to be identified considering that twice more genes have been predicted using gene prediction algorithm. Novel approaches can be developed to explore different search spaces that may reveal even more protein-coding genes.

Indeed, additional evidences suggest that more protein may exist in *C. elegans* in the case of old protein fold families that have evolved long time ago from divergent (or convergent) evolution [4]. Such protein family members are renowned to be difficult to identify by conventional sequence alignment software since they share very little sequence identity. The OB fold is one example [5]. The domain is a compact structural motif frequently used for nucleic acid recognition. It is composed of a five-stranded beta-sheet forming a closed beta-barrel. This barrel is capped by an alpha-helix located between the third and fourth strands. Structural comparison and analysis of all OB-fold/nucleic acid complexes solved to date confirms the low degree of sequence similarity among members of this family arisen from divergent evolution [6]. In addition, Loop connecting the secondary structures elements are highly variable in length making them difficult to compare at the sequence level. In *C. elegans* the number of predicted proteins containing Ob-fold is remarkably low compared to other related organisms by evolution. The number of OB fold proteins vary widely from human (256 OB fold proteins), mouse (246 OB fold proteins), yeast (*Saccharomyces cerevisiae*) (344 OB fold proteins) to fruit fly (*Drosophila melanogaster*) (84 OB fold proteins) and *C. elegans* (46 OB fold proteins at the time we started this project). Genomics rearrangements during evolution may have occurred or differences in the level of annotation for this protein family may explain these discrepancies.

The identification of distant related sequences or remote homologues from functional domain families has been extensively improved this last decade. Methods that can detect intermediate sequence to connect sequences sharing insignificant BLAST scores between each other have been implemented ([7,8]). Sequence-sequence and sequence-profile alignment algorithm BLAST [9] and PSI-BLAST [10] have been cited more than 60000 times. The sensitivity and alignment quality depend on the information that is used to compare proteins. The most sensitive methods use sequence-profiles or profile-profile alignments (Table 1. Sequence Discovery Module). They contain position-specific substitution scores that are computed from the frequencies of amino acids at each position of a multiple alignment of related sequences. Further improvement have been remarkable by the introduction of Hidden Markov Model [11] and profile hidden markov model [12] that can compute more accurately gap, insertion and deletion in the alignments compared to previous methods. Moreover, fold recognition methods that build a 3D structure model of a protein sequence from a sequence alignment have been very efficient in their ability to align correctly sequence/profile to profile of known structure (Table 1. Structure Discovery Module). Building model that are very similar structurally to the templates structure from these alignments can be used to validate a correct alignment especially if such alignment is between sequence that have very low sequence similarities. More recently, many bioinformatics studies suggest that consensus methods that pool together the results of different software that perform similar tasks perform better than isolated methods.

This study examines the possibility that novel OB fold coding genes exist in the worm. We developed a consensus approach that uses the most sensitive sequence-sequence, sequence-profile and profile-profile similarity searches methods followed by OB-fold 3D-structure

prediction as a filter to eliminate false positive candidate remote sequences. We have predicted 18 coding gene containing the OB-fold. Remarkably, most of their corresponding genes have not or partially been characterized in the worm. Few of them are essential genes since their knockout produces embryonic lethal phenotypes.

Results

From the 46 proteins used to generate the profile by MEME [13] and PSI-BLAST about 200 candidate proteins that may contain OB fold were identified by SeqDIM. Further validation by the StrucDIM package confirmed the OB fold prediction for only *brc-2* and *pot-1*. This finding was not far from our expectation, since many OB fold family members share less than 10% sequence similarity between each others, which is consistent with the high degree of sequence divergence of this family that occurred during evolution. Therefore even though very sensitive sequence alignment methods are used, detection of novel OB fold proteins remained difficult.

Since very divergent sequences that do not share significant sequences identity may have the same fold and considering the conserved structure of OB-fold, we used fold recognition methods of StrucDIM to investigate if more of OB fold proteins could be obtained directly. The underlying assumption is that if a correct model can be build by comparative modeling using a sequence alignment between a protein sequence of an OB fold of known structure with an OB fold candidate sequence then the sequence alignment is significant if the model is correct. It allow us to put some confidence in the pairwise alignment of sequences that share a level of sequence identity below the twilight zone (25% identity) since sequence alignment statistics cannot determine their significance at this level of identity. Effectively, un-correct alignments do not generate well folded homology models. Using this direct approach, 4300 sequences from a dataset of genes present in the germline of *C. elegans* [14] were submitted directly to fold recognition servers using StrucDIM. This dataset is expected to be enriched in genes involved in DNA processes including DNA repair and replication which mostly posses OB fold 3D-structure.

By this direct approach, we determined that 35 out of 46 of the known OB fold proteins in *C. elegans* were present and predicted in this dataset [14]. These results confirm that the dataset is enriched in OB fold sequences. It also shows that StrucDiM approach is valid and can be used to further identify novel Ob fold protein coding genes. Indeed, further analysis of these results revealed that we were correct since we could obtain 18 novel OB fold candidate proteins that have not been predicted previously to our knowledge (Table 2). However it should be noted that the Ob fold 3D-structure of the human homologue *pot-1* has been recently deposited in the Protein Data Bank (PDB accession number: 1XJV).

To further identify additional OB fold gene coding proteins we searched for orthologues and homologues of the identified candidates in both human and *C. elegans*. Using the protein family orthologues, and paralogues module in the comparative genomics toolbox of ENSEMBL database we were able to identify 3 additional candidates homologues of *pot-3* (*pot-2*, *mrt-1*, F48E8.6) and one homologues of F25B5.5 (Y92H12BL.2). We expected to see these proteins also have OB fold similar to their paralogues. In addition, we then used phyre2 a protein fold recognition server [15] to predict the structure of these proteins. As expected, all candidates were confirmed to contain OB fold. These 4 novel OB fold proteins had not been previously predicted and annotated in wormbase however for 2 of them (*mrt-1* and *pot-2*) we found one publication mentioning about these two genes as containing OB fold domain [16].

Discussion

One important question regarding this study is why the annotation of these genes has been missed from wormbase. The obvious lack of sequence similarity among members of this family is one possible explanation since it makes these proteins undetectable through sequence based

searches. This is consistent with our inability to identify novel OB fold protein coding genes using SeqDIM module. On the contrary, we showed that structural based methods are more robust at predicting OB fold proteins and these methods are generally not considered in genome annotation pipelines which may explain why many of these OB fold containing genes have not been annotated

Regarding the genes that have been identified it is remarkable that most of these genes have not been well studied (Table 3). However, a significant fraction of these genes seems to perform important function during development and are essential genes since RNAi phenotype (EXOS-3) as well as knockout when available shows embryonic lethality. It includes protein coding genes involved DNA replication and repair (F12F6.7, BRC-2) and growth rate and reproduction (EXOS-1, C05D11.10, F10E9.4) as well as the protection of telomere protein POT-3 involved in telomere maintenance. Other OB fold candidate proteins do not seem to be essential during development since they only show no or non-lethal phenotype. Those include gene coding proteins involved in nucleic acids and RNA binding (EXOS-2) a component of the exosome complex (with EXOS-1 and EXOS-3), DIS-3, ZK470.2, W08A12.2, T07C12.12, F25B5.5 as well as POT-1 involved in telomere maintenance. To annotate further the function of these genes we look at protein-protein interaction in the STRING [17] and BIOGRID [18] database, no interactions were found for most of them in BIOGRID database with the exception of EXOS-3, C05D11.10, POT-1, BRC-2 that interact with genes involved in cell division, nucleic-acid binding and RNA processing, DNA repair for BRC-2 and POT-1 involved in IGF signaling, lifespan extension and longevity.

We have shown that comparative modelling approaches are a powerful tool to identify novel protein coding genes with interesting and uncharacterized functions even in a genome and proteome of a model organism as extensively annotated as *C. elegans*

Material and methods:

Input sequences:

Protein sequences used in this study to identify novel OB fold proteins were obtained from the 46 OB fold known proteins in wormbase and an enriched data set of 4300 expressed genes in the germ line of *C. elegans* [14]. This data set should be enriched in novel genes containing OB fold since OB fold proteins are generally involved in many DNA transaction and DNA repair process such processes are highly active in *C. elegans* germline ().

Consensus Discovery Pipeline:

The pipeline has 3 modules (i) **Sequence based Discovery Module** (ii) **Structure based Discovery Module** and filtering (iii) **Functional Discovery Module**:

Sequence based Discovery Module:

From the 46 OB fold known proteins in *C. elegans* a position specific scoring matrix of OB fold motifs were built using PSI-BLAST [10] as well as a Hidden Markov Model using MEME [13] from sequences of the nr database. Each of the profiles were subsequently submitted to different database scanning software using sequence-profile based alignment methods against wormpep210 protein sequence database. For the profile-profile HHSenser [19] methods the database to scan for was made-up of sequence profiles of all the known protein families. For each method default threshold of significance were used to select for novel candidate OB fold protein sequences

Structural Discovery Module

The 4300 sequences from claycomb et al. as well as the 200 sequences OB fold candidates obtained from SeqDiM were submitted to the consensus fold recognition metaserver [20] to perform and confirm fold prediction. This method collects and score many different fold prediction results using the 3D jury consensus method from a protein sequence [21]. Model building for predicted OB fold motif in candidates were further performed by modeller [22] from metaserver alignment results and resubmitting candidate sequence to the 3D structure prediction server I-tasser [23]. Model quality and validation were further performed using TM-align [24]. A TM-score < 0.2 indicates that there is no similarity between two structures; a TM-score > 0.5 means the structures share the same fold.

Functional Discovery Module

To gain some insight into the function of the novel OB fold candidates discovered, protein-protein interaction databases, subcellular localization and gene ontology predictors were interrogated (Table 1. Function Discovery Module).

Author Contributions

References

1. C. elegans Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* 282(5396): 2012-2018.
2. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. (2003) The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol* 1(2): E45. 10.1371/journal.pbio.0000045.
3. Magrane M, Consortium U. (2011) UniProt knowledgebase: A hub of integrated protein data. *Database (Oxford)* 2011: bar009. 10.1093/database/bar009.
4. Murzin AG. (1998) How far divergent evolution goes in proteins. *Curr Opin Struct Biol* 8(3): 380-387.
5. Murzin AG. (1993) OB(oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J* 12(3): 861-867.
6. Theobald DL, Cervantes RB, Lundblad V, Wuttke DS. (2003) Homology among telomeric end-protection proteins. *Structure* 11(9): 1049-1050.
7. Li W, Pio F, Pawlowski K, Godzik A. (2000) Saturated BLAST: An automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics* 16(12): 1105-1110.
8. Soding J, Remmert M. (2011) Protein sequence comparison and fold recognition: Progress and good-practice benchmarking. *Curr Opin Struct Biol* 21(3): 404-411. 10.1016/j.sbi.2011.03.005.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol* 215(3): 403-410. 10.1016/S0022-2836(05)80360-2.
10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389-3402.
11. Eddy SR. (1996) Hidden markov models. *Curr Opin Struct Biol* 6(3): 361-365.
12. Eddy SR. (1998) Profile hidden markov models. *Bioinformatics* 14(9): 755-763.
13. Grundy WN, Bailey TL, Elkan CP, Baker ME. (1997) Meta-MEME: Motif-based hidden markov models of protein families. *Comput Appl Biosci* 13(4): 397-406.
14. Claycomb JM, Batista PJ, Pang KM, Gu W, Vasale JJ, et al. (2009) The argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation. *Cell* 139(1): 123-134. 10.1016/j.cell.2009.09.014.
15. Kelley LA, Sternberg MJ. (2009) Protein structure prediction on the web: A case study using the phyre server. *Nat Protoc* 4(3): 363-371. 10.1038/nprot.2009.2.

16. Meier B, Barber LJ, Liu Y, Shtessel L, Boulton SJ, et al. (2009) The MRT-1 nuclease is required for DNA crosslink repair and telomerase activity in vivo in *caenorhabditis elegans*. *EMBO J* 28(22): 3549-3563. 10.1038/emboj.2009.278.
17. Snel B, Lehmann G, Bork P, Huynen MA. (2000) STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 28(18): 3442-3444.
18. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue): D535-9. 10.1093/nar/gkj109.
19. Soding J, Remmert M, Biegert A, Lupas AN. (2006) HHsenser: Exhaustive transitive profile search using HMM-HMM comparison. *Nucleic Acids Res* 34(Web Server issue): W374-8. 10.1093/nar/gkl195.
20. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. (2001) Structure prediction meta server. *Bioinformatics* 17(8): 750-751.
21. Ginalski K, Elofsson A, Fischer D, Rychlewski L. (2003) 3D-jury: A simple approach to improve protein structure predictions. *Bioinformatics* 19(8): 1015-1018.
22. Fiser A, Sali A. (2003) Modeller: Generation and refinement of homology-based protein structure models. *Methods Enzymol* 374: 461-491. 10.1016/S0076-6879(03)74020-8.
23. Roy A, Kucukural A, Zhang Y. (2010) I-TASSER: A unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4): 725-738. 10.1038/nprot.2010.5.
24. Zhang Y, Skolnick J. (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33(7): 2302-2309. 10.1093/nar/gki524.
25. Soding J, Biegert A, Lupas AN. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33(Web Server issue): W244-8. 10.1093/nar/gki408.
26. Sadreyev RI, Tang M, Kim BH, Grishin NV. (2007) COMPASS server for remote homology inference. *Nucleic Acids Res* 35(Web Server issue): W653-8. 10.1093/nar/gkm293.
27. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303(5657): 540-543. 10.1126/science.1091403.
28. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, et al. (2007) WoLF PSORT: Protein localization predictor. *Nucleic Acids Res* 35(Web Server issue): W585-7. 10.1093/nar/gkm259.
29. Hawkins T, Luban S, Kihara D. (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci* 15(6): 1550-1556. 10.1110/ps.062153506.
30. Gallo CM, Munro E, Rasoloson D, Merritt C, Seydoux G. (2008) Processing bodies and germ granules are distinct RNA granules that interact in *C. elegans* embryos. *Dev Biol* 323(1): 76-87. 10.1016/j.ydbio.2008.07.008.
31. Chen D, Pan KZ, Palter JE, Kapahi P. (2007) Longevity determined by developmental arrest genes in *caenorhabditis elegans*. *Aging Cell* 6(4): 525-533. 10.1111/j.1474-9726.2007.00305.x.
32. van Haften G, Romeijn R, Pothof J, Koole W, Mullenders LH, et al. (2006) Identification of conserved pathways of DNA-damage response and radiation protection by genome-wide RNAi. *Curr Biol* 16(13): 1344-1350. 10.1016/j.cub.2006.05.047.
33. Arur S, Ohmachi M, Nayak S, Hayes M, Miranda A, et al. (2009) Multiple ERK substrates execute single biological processes in *caenorhabditis elegans* germ-line development. *Proc Natl Acad Sci U S A* 106(12): 4776-4781. 10.1073/pnas.0812285106.
34. Xue H, Xian B, Dong D, Xia K, Zhu S, et al. (2007) A modular network model of aging. *Mol Syst Biol* 3: 147. 10.1038/msb4100189.
35. Coghlan A, Wolfe KH. (2004) Origins of recently gained introns in *caenorhabditis*. *Proc Natl Acad Sci U S A* 101(31): 11362-11367. 10.1073/pnas.0308192101.
36. Andachi Y. (2008) A novel biochemical method to identify target genes of individual microRNAs: Identification of a new *caenorhabditis elegans* let-7 target. *RNA* 14(11): 2440-2451. 10.1261/rna.1139508.
37. Lowden MR, Meier B, Lee TW, Hall J, Ahmed S. (2008) End joining at *caenorhabditis elegans* telomeres. *Genetics* 180(2): 741-754. 10.1534/genetics.108.089920.

38. Raices M, Verdun RE, Compton SA, Haggblom CI, Griffith JD, et al. (2008) *C. elegans* telomeres contain G-strand and C-strand overhangs that are bound by distinct proteins. *Cell* 132(5): 745-757. 10.1016/j.cell.2007.12.039.
39. Lemmens BB, Tijsterman M. (2011) DNA double-strand break repair in *caenorhabditis elegans*. *Chromosoma* 120(1): 1-21. 10.1007/s00412-010-0296-3.
40. Youds JL, Barber LJ, Boulton SJ. (2009) *C. elegans*: A model of fanconi anemia and ICL repair. *Mutat Res* 668(1-2): 103-116. 10.1016/j.mrfmmm.2008.11.007.
41. Ko E, Lee J, Lee H. (2008) Essential role of *brc-2* in chromosome integrity of germ cells in *C. elegans*. *Mol Cells* 26(6): 590-594.
42. Kruisselbrink E, Guryev V, Brouwer K, Pontier DB, Cuppen E, et al. (2008) Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCD1-defective *C. elegans*. *Curr Biol* 18(12): 900-905. 10.1016/j.cub.2008.05.013.
43. Youds JL, Barber LJ, Ward JD, Collis SJ, O'Neil NJ, et al. (2008) *DOG-1* is the *caenorhabditis elegans* BRIP1/FANCD1 homologue and functions in interstrand cross-link repair. *Mol Cell Biol* 28(5): 1470-1479. 10.1128/MCB.01641-07.
44. Goodyer W, Kaitna S, Couteau F, Ward JD, Boulton SJ, et al. (2008) HTP-3 links DSB formation with homolog pairing and crossing over during *C. elegans* meiosis. *Dev Cell* 14(2): 263-274. 10.1016/j.devcel.2007.11.016.
45. Pispas J, Palmes S, Holmberg CI, Jantti J. (2008) *C. elegans* *dss-1* is functionally conserved and required for oogenesis and larval growth. *BMC Dev Biol* 8: 51. 10.1186/1471-213X-8-51.
46. Min J, Park PG, Ko E, Choi E, Lee H. (2007) Identification of Rad51 regulation by BRCA2 using *caenorhabditis elegans* BRCA2 and bimolecular fluorescence complementation analysis. *Biochem Biophys Res Commun* 362(4): 958-964. 10.1016/j.bbrc.2007.08.083.
47. Ward JD, Barber LJ, Petalcorin MI, Yanowitz J, Boulton SJ. (2007) Replication blocking lesions present a unique substrate for homologous recombination. *EMBO J* 26(14): 3384-3396. 10.1038/sj.emboj.7601766.
48. Petalcorin MI, Galkin VE, Yu X, Egelman EH, Boulton SJ. (2007) Stabilization of RAD-51-DNA filaments via an interaction domain in *caenorhabditis elegans* BRCA2. *Proc Natl Acad Sci U S A* 104(20): 8299-8304. 10.1073/pnas.0702805104.
49. Petalcorin MI, Sandall J, Wigley DB, Boulton SJ. (2006) CeBRC-2 stimulates D-loop formation by RAD-51 and promotes DNA single-strand annealing. *J Mol Biol* 361(2): 231-242. 10.1016/j.jmb.2006.06.020.
50. Garcia-Muse T, Boulton SJ. (2005) Distinct modes of ATR activation after replication stress and DNA double-strand breaks in *caenorhabditis elegans*. *EMBO J* 24(24): 4345-4355. 10.1038/sj.emboj.7600896.
51. Martin JS, Winkelmann N, Petalcorin MI, McIlwraith MJ, Boulton SJ. (2005) RAD-51-dependent and -independent roles of a *caenorhabditis elegans* BRCA2-related protein during DNA double-strand break repair. *Mol Cell Biol* 25(8): 3127-3139. 10.1128/MCB.25.8.3127-3139.2005.
52. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O. (2008) *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat Methods* 5(10): 865-867. 10.1038/nmeth.1249.
53. Yang W, Hekimi S. (2010) Two modes of mitochondrial dysfunction lead independently to lifespan extension in *caenorhabditis elegans*. *Aging Cell* 9(3): 433-447. 10.1111/j.1474-9726.2010.00571.x.
54. Boerckel J, Walker D, Ahmed S. (2007) The *caenorhabditis elegans* Rad17 homolog HPR-17 is required for telomere replication. *Genetics* 176(1): 703-709. 10.1534/genetics.106.070201.
55. Harris J, Lowden M, Clejan I, Tzoneva M, Thomas JH, et al. (2006) Mutator phenotype of *caenorhabditis elegans* DNA damage checkpoint mutants. *Genetics* 174(2): 601-616. 10.1534/genetics.106.058701.

Figures

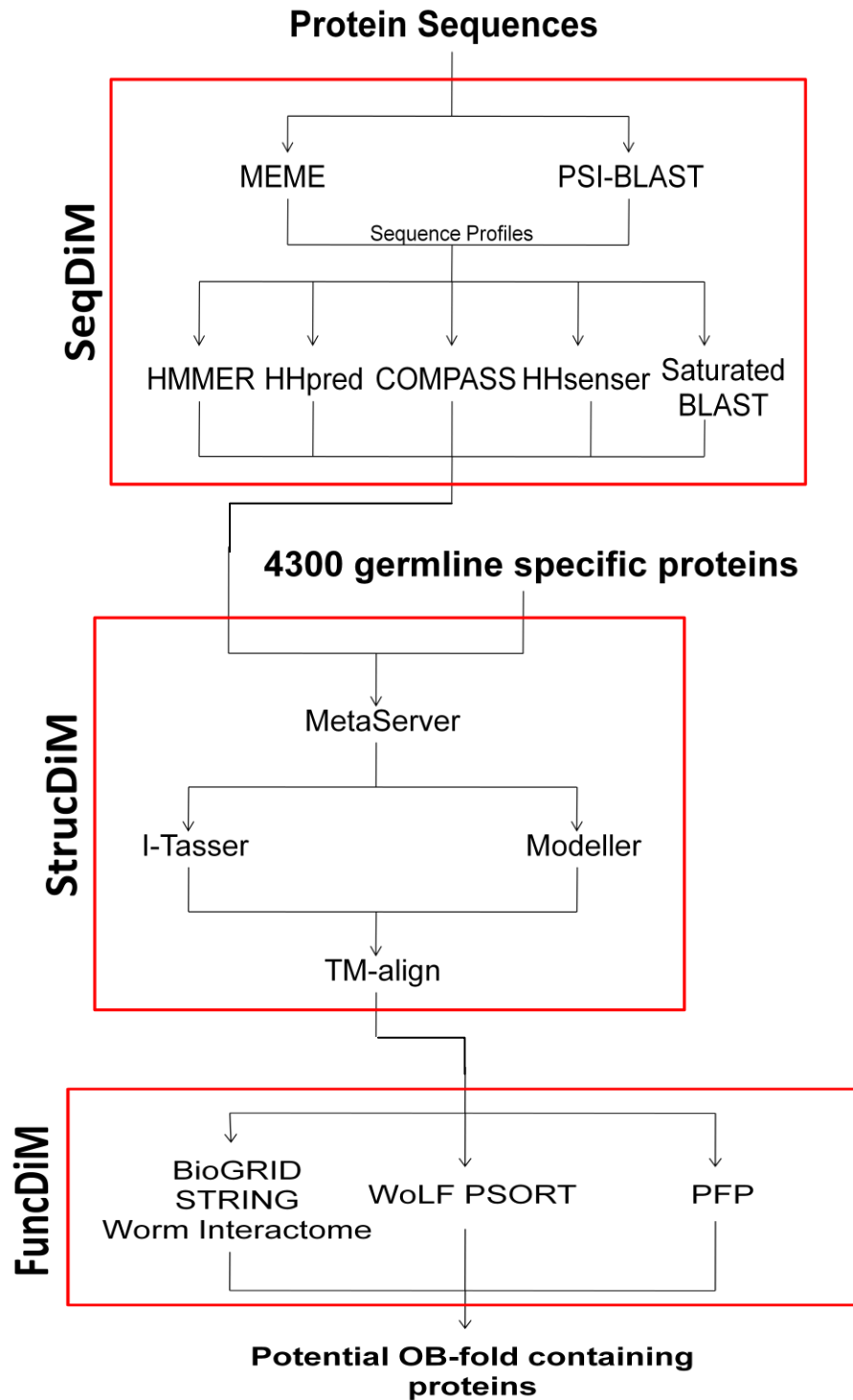


Figure 1. Discovery Pipeline of novel OB fold protein coding genes. It contains 3 Discovery Modules. SeqDiM: Sequence alignment Discovery Module; StrucDiM:3D Structure prediction Discovery Module; and a Functional prediction Discovery Module FuncDiM.

Tables

Table 1: Tools used in this study.

Tools	Description	Reference
Sequence Discovery Module		
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool	<i>Altschul et al. 1990</i> [10]
MEME	Motif based Hidden Markov Model of protein families	<i>Grundy et al. 1997</i> [13]
HMMER	Bio-sequence analysis tool using profile hidden Markov Models	<i>Eddy, 1998</i> [12]
HHpred	Homology detection & structure prediction tool by HMM-HMM comparison	<i>Soding et al. 2005</i> [25]
COMPASS	Alignment tool of multiple protein sequence profiles	<i>Sadreyev et al. 2007</i> [26]
HHsenser	Exhaustive intermediate profile search tool using HMM-HMM comparison	<i>Soding et al. 2006</i> [19]
Saturated-BLAST	Automated toolbox that implement the multiple intermediate sequence search method	<i>Li et al. 2000</i> [7]
Structure Discovery Module		
MetaServer	A Server that submit and collect fold recognition results from different methods and makes 3D-prediction using a consensus approach called 3D-jury.	<i>Bujnicki et al. 2001</i> [20]
I-Tasser	Protein 3D-structure prediction server that uses threading methods	<i>Roy et al. 2010</i> [23]
Modeller	Protein 3D-structure modeling tool from target-template sequence alignment based on satisfaction of spatial restraints	<i>Fiser et al. 2003</i> [22]
TM-Align	Protein 3D-structure alignment algorithm that compute the TM-Score	<i>Zhang et al. 2005</i> [24]
Functional Discovery Module		
BioGrid	Database of Protein and Genetic Interactions	<i>Stark et al. 2006</i> [18]
STRING	Database of Functional protein association networks	<i>Snel et al. 2000</i> [17]
Worm Interactome	A high quality yeast two-hybrid protein-protein interactions database of <i>C. elegans</i>	<i>Li et al. 2004</i> [27]
WoLF PSORT	Protein sub-cellular localization predictor	<i>Horton et al. 2007</i> [28]
Kihara PFP	Protein function predictor	<i>Hawkins et al. 2006</i> [29]

Table 2: Model Quality of novel OB fold protein coding genes.

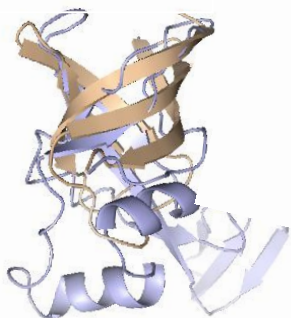
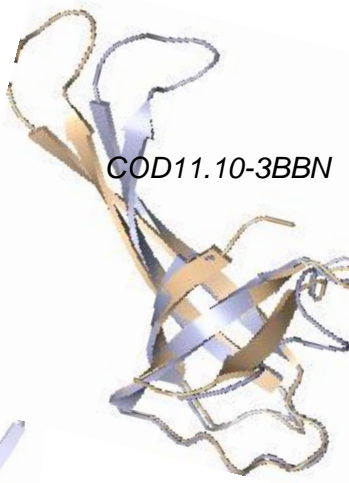
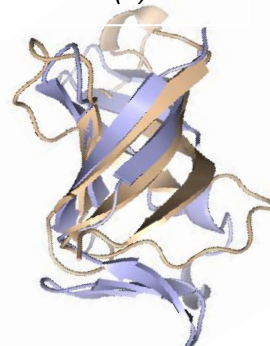
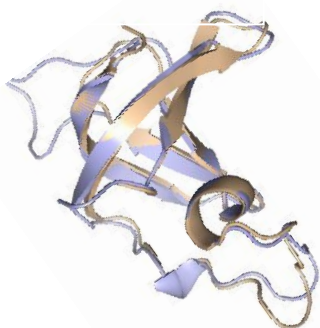
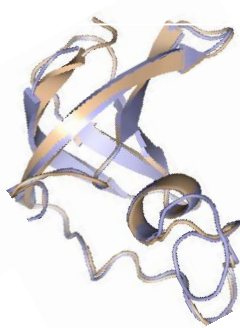
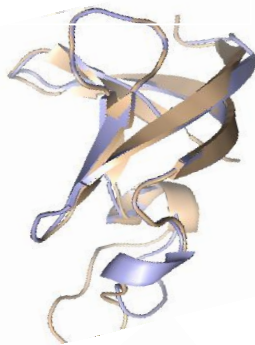
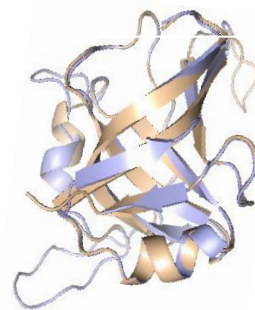
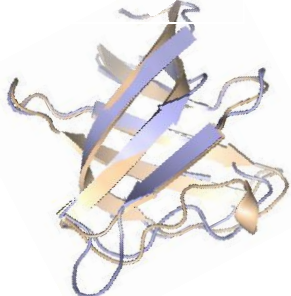
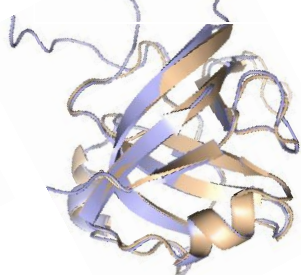
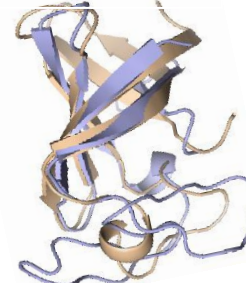
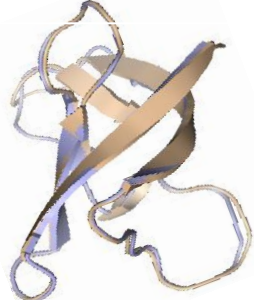
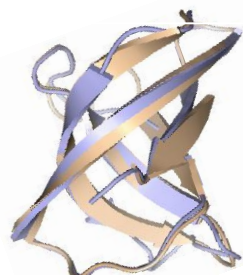
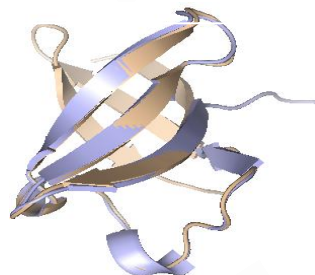
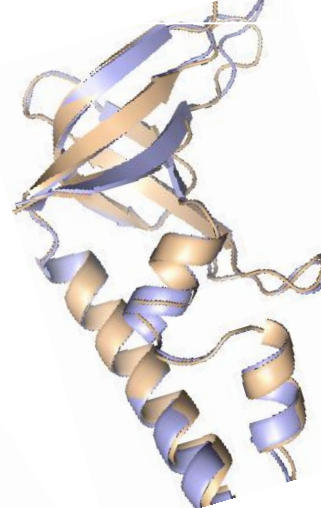
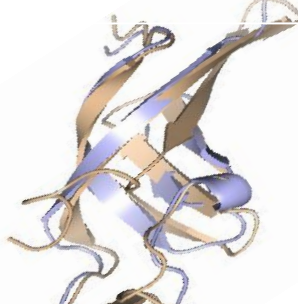
OB fold Candidates target	Template	RMSD	TM-score	Equivalent C _α superimposed
F12F6.7	3E0J	0.9	0.91618	104/110
F25B5.5	2QGQ	1.08	0.79684	57/64
exos-2	2Z0S	0.39	0.91855	80/86
exos-3	2Z0S	1.33	0.93357	66/66
exos-1	2Z0S	0.97	0.83856	76/85
dis-3 (First OB fold)	2IX1	2.15	0.77503	81/92
dis-3 (Second OB fold)	1UEB	3.66	0.51393	76/98
ZK470.2	3BBN	1.22	0.90075	43/45
C05D11.10	1HZA	1.88	0.8186	77/82
W08A12.2	2C35	1.27	0.91183	58/59
F10E9.4	1XJV	1.98	0.81487	61/61
Pot-1	1L1O	1.11	0.89915	128/135
brc-2	1XJV	3.43	0.43998	74/115
Pot-3	3MXN	1	0.83903	115/133
T07C12.12	1XJV	1.49	0.90455	132/139
Pot-2	3KJO	0.4	0.86529	110/126
mrt-1	2QGQ	1.03	0.83313	115/135
Y92H12BL.2	1YZ6	0.62	0.89597	56/60
F48E8.6	3E0J	2.27	0.64349	66/81

Table 3: **Functional analysis of Novel OB folds protein coding genes.**

* refers to predicted functions. Homologues and paralogues referred to human

OB folds	WB ID	Biblio	RNAi Phenotype	Knockout	Function	homologues	Paralogs
F12F6.7	WBGene0008722	NA	Embryonic lethal	ok2252	DNA replication, DNA binding, DNA-directed DNA polymerase activity	POLD2 polymerase (DNA directed), delta 2, regulatory subunit 50kDa	NA
F25B5.5	WBGene00017776	NA	NA	NA	RNA modification, iron-sulfur cluster binding, 4 iron, 4 sulfur cluster binding, catalytic activity	CDK5RAP1 CDK5 regulatory subunit associated protein 1	Y92H12BL.2
exos-2	WBGene00022232	[30]	late larval arrest	NA	*nucleic acid binding, RNA binding,	EXOSC2 <u>exosome component 2</u>	NA
exos-3	WBGene00010325	[30-32]	Embryonic lethal	NA	growth,nematode larval development,receptor-mediated endocytosis	EXOSC3 <u>exosome component 3</u>	NA
exos-1	WBGene00012966	[30]	Embryonic lethal, lethal	ok807	positive regulation of growth rate, reproduction	EXOSC1 <u>exosome component 1</u>	NA
dis-3	WBGene00001001	[33-35]	Slow growth, sick, sterile progeny	ok357	RNA binding, ribonuclease activity, sequence-specific DNA binding, reproduction	DIS3 mitotic control homolog (S. cerevisiae)	F48E8.6
ZK470.2	WBGene00022745	[36]	NA	ok5876	*single-stranded telomeric DNA binding, ion binding, monosaccharide metabolism	NA	NA
C05D11.10	WBGene00015487	NA	Embryonic lethal, lethal, slow growth	ok5298	growth, nematode larval development, positive regulation of growth rate, reproduction	NA	NA
W08A12.2	WBGene00021079	NA	NA	NA	*purine nucleotide binding, adenyly nucleotide binding, cellular macromolecule metabolism	NA	NA
F10E9.4	WBGene00017356	NA	Slow growth, larval lethal	NA	growth, nematode larval development, positive regulation of growth rate, reproduction	NA	NA
Pot-1	WBGene00015105	[16,37,38]	organism development variant, telomere homeostasis variant	NA	*cAMP-dependent protein kinase activity, transition metal ion binding, ion binding	Pot1 <u>Protection Of Telomeres 1</u>	NA

brc-2	WBGene00020316	[39-51]	Embryonic lethal, lethal, embryonic arrest	ok1629	strand invasion, double-strand break repair, reproduction, single-stranded DNA and protein binding	Brca1 <u>Breast Cancer type 1</u> susceptibility protein	NA
Pot-3	WBGene00007065	[16,38]	lethal	ok1530	*cation binding, adenyl nucleotide binding, heterocycle metabolism	Pot1 <u>Protection Of Telomeres 1</u>	pot-2, mrt-1
T07C12.12	WBGene00011576	[52]	Embryonic lethal	NA	*adenyl nucleotide binding, rRNA (adenine) methyltransferase activity, purine nucleotide binding	RMI1, RecQ <u>mediated genome instability 1</u>	NA
Pot-2	WBGene00010195	[16,38]	NA	NA	*cation binding, adenyl nucleotide binding, heterocycle metabolism	NA	pot-3, mrt-1
MRT-1	WBGene00045237	[16,38,53-55]	Sterile, lethal	ok758	Nuclear excision repair, telomere maintenance via telomerase, reproduction, Single stranded DNA binding	NA	pot-2, pot-3
Y92H12BL.2	WBGene00022363	NA	NA	NA	Iron-sulfur cluster binding	CDKAL1, CDK5 regulatory subunit <u>associated protein 1-like 1</u>	F25B 5.5
F48E8.6	WBGene00018612	NA	NA	NA	RNA binding, ribonuclease activity	DIS3L2, DIS3 mitotic control homolog (<i>S. cerevisiae</i>)- <u>like 2</u>	dis-3

BRC-2-1L10*COD11.10-3BBN**DIS-3(1)-2IX1**DIS-3(2)-2IX1**EXOS-1-2Z0S**EXOS-2-2Z0S**EXOS-1-2Z0S**MRT-1-3KJO**POT1-1XJV**POT3-1XJV**F25B5.5-2QGQ**F48E8.6-1YZ6**W08A12.2-1HZA**Y92H12BL.1-2QGQ**ZK470.2-1UEB**T07C12.12-3MXN**F12F6.7-3EZJ**F10E9.4-2C35*

Search in the Tumor Liberated Protein (TLP) for Specific Peptides of Non Small Cell Lung (NSCL) Cancer

Giulio Tarro, MD, PhD

Department of Biology, Center for Biotechnology, Sbarro institute for Cancer Research and Molecular Medicine, Temple University, Philadelphia, PA, USA.
Committee on Biotechnologies and VirusSphere, World Academy of Biomedical Technologies, UNESCO, Paris, France.

Foundation T. & L. de Beaumont Bonelli for Cancer Research, Naples, Italy
Correspondence to: Prof. Dr. Giulio Tarro, Via Posillipo 286, 80123 Naples, Italy
e-mail: gitarro@tin.it giuliotarro@gmail.com

Abstract - Rabbit polyclonal immune serum against TLP was produced by immunizing rabbits with the RTNKEASIC (the 1st) and NQRNRD (the 2nd) synthetic peptides, at the Rockland Immunochemicals Inc, PA. Protein extraction was performed from three cell lines of lung carcinoma including A549 cells, accordingly to previously studies. Cell lysates were loaded into two polyacrilamide gels and then were transferred to the nitrocellulose membranes. One nitrocellulose was hybridized with the anti-TLP serum and the other one with the pre-immunization serum. Preliminary results showed two major intense bands with a molecular weigh of 50 and 100kDa, which should correspond to the monomeric and dimeric form of TLP, respectively. In conducting a competition assay (PCA) to verify the specificity of the 50 and 100 kDa bands for TLP, the antibody anti-TLP was pre-incubated with the 1st and the 2nd peptides, before its hybridization with the nitrocellulose. In fact, a reduced intensity of the 100 kDa band following the PCA assay was observed, suggesting its specificity for the antibody anti-TLP. Currently, the Rockland Immunochemicals Inc is improving the signal specificity by purifying the antiserum on chromatographic columns, through the agarose matrix and the 1st and the 2nd peptides. Moreover, next purpose will be to immunoprecipitate TLP from the cell lysate and to load it on the SDS-PAGE gel. Then, the protein band of interest will be excised from the stained gel and the peptides will be extracted from the gel slice and the aminoacid sequence will be analyzed.

Keywords: TLP, NSCL, CRC, Immunotherapy, Vaccine

1 Introduction

Long years of research were required for boosting the immune system to fight cancer [1]; [2]. In the 1890s, mixtures of dead bacteria were injected by William B. Coley into cancer patients to stimulate the immune system. According to Paul Ehrlich (1909) the immune system may suppress tumor development. In the 1960s, both in animals and men neoplastic cell antigens stimulated the onset of specific humoral and cellular antibodies [3]. In 1972 Immunogenicity of a soluble transplantation antigen from adenovirus 12 - induced tumor cells was demonstrated in

inbred hamsters (PD-4) [4]. In 1975 there was the discovery of Monoclonal Antibodies, highly specific immunological tools, and in 1980 mass-production of interferon, the immune-stimulating molecule, was obtained after inserting its coding gene into bacteria. In 1986 Interferon is approved by the Food and Drug Administration (FDA) for the treatment of hairy cell leukemia. In 1997 the FDA okays the first monoclonal antibody (MA) treatment against cancer (for non-Hodgkin's lymphoma), and in 1998 the FDA approves the MA Herceptin for the treatment of metastatic breast cancer. Basic cellular immune response to cancer [5]: 2002 – National Cancer Institute researchers prove that two kinds of immune cell – CD4+ T cells and CD8+ T cells-are required for the treatment against cancer. The CD4+ cell releases cytokine molecules that help to activate the CD8+ cells prompting them to attack other cells with the same antigen. Therapeutic Vaccine Strategies [6]; [7]: Tumor cells are removed from a patient and treated biochemically or irradiated. Then the extracts of the dead cancer cells are reinjected, boosting the immune system to attack the tumor cell. Tumor liberated protein (TLP) boosts the immune system's cancer responsive capabilities, 1983 [8]. TLP may have the potential to greatly improve the cure rate and/or serve as a lung cancer vaccine, 1991 [9]; [10]. Detection of lower levels of TLP/antiTLP may be of clinical relevance, 1992 [11]: TLP as candidate marker for the early detection of NSCL cancer. More on therapeutic Vaccine Strategies: Tumor – associated antigens resulting from protein bits, or from synthesized peptides specific for the cancer tissue, can be used successfully as vaccine to mount a vigorous antitumor attack: Development of a vaccine approach for therapeutical and preventive application [12]. The dendritic cell is an immune cell that presents specific antigens taken from a tumor cell to two other immune cells, the CD4+ and CD8+ cells. The dendritic cells of a cancer patient are removed and loaded with antigens from the tumor. The dendritic cells grow outside the body and then are reinjected, triggering a powerful response by the T cells [13]; [14]. The FAA approves the first therapeutic cancer vaccine for advanced prostate cancer (Provenge 2010).

Previously, we identified a -100 kDa protein, which is part of a protein complex named tumor liberated proteins (TLP), as a promising blood marker for early

diagnosis of lung cancer [12]; [15]. In particular, this protein proved to have high specificity and sensitivity for stage I patients with NSCL. TLP might also represent a predictive marker of cell transformation since it is expressed in interstitial lung fibrosis. Moreover, TLP showed a specific immunogenic activity, suggesting its possible use as an anticancer vaccine. Indeed, it is able to induce delayed hypersensitivity reactions and to promote blastogenesis in cultured lymphocytes from patients presensitized with TLP.

Research is ongoing to obtain the complete sequence of TLP, by proteomics approaches, in order to achieve adequate antigen preparations that might be used to generate assays for early diagnosis and, possibly, a specific anticancer vaccine [16]

2 Results

According to the partial sequencing of TLP, two peptides were synthesized: TLP peptide 1: Ac-RTNKEASI-Ahx-C-amide TLP peptide 2: Ac-Ahx-C-amide-NQRNRD A mixing of the two peptides was administered to two rabbits in order to obtain a serum for subsequent analysis. Therefore different sera samples were taken at various dates. The capability of sera to recognize TLP was analyzed by Western blotting using protein extracts of lung cancer cell lines (A549, H23, H82, H187) and control lines (MET -SA, NL-20 and primary line of fibroblasts). The signal obtained by anti-TLP antibodies was found to be not very specific.

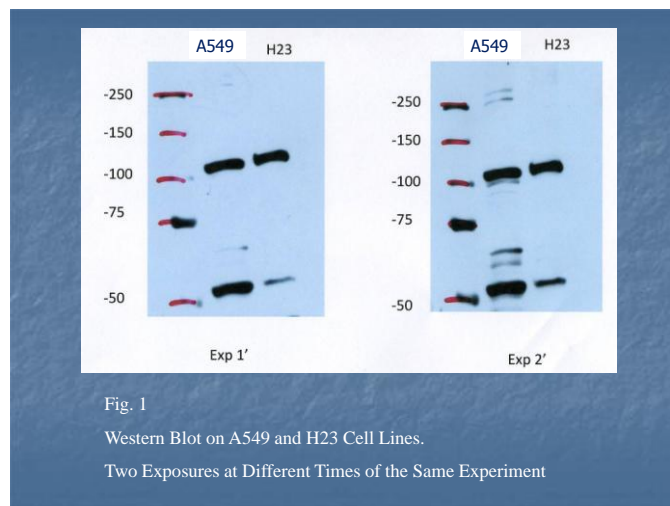
In order to improve the specificity of the anti-TLP antiserum a Peptide Competition Assay was carried on. In this assay, the antibody is preincubated with the peptides before its use in the immunoblotting.

The immunoblotting experiment is conducted in duplicate, one with the antibody preincubated with the peptide and the other one with the control antibody. The results show a better signal quality and on the basis of these data, a request has been made to the company responsible for the production of the sera to purify the antibodies on a series of resins conjugated with the peptides TLP1 and TLP2.

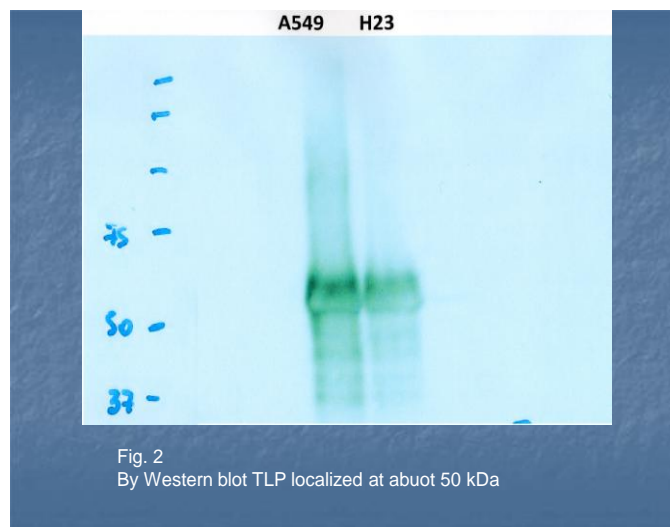
The serum obtained after purification was found to be more specific, in particular a sample specifically recognized the band of 100 kDa and 50 kDa protein, presumably corresponding to the TLP. However in numerous subsequent analysis the data has not been confirmed. For this reason the company has been requested a new specimen of purified anti-TLP serum.

In parallel several immune precipitation assays were carried out using cell extracts of A549 and H23 lines in order to obtain a precipitate containing only the TLP protein (Fig 1). This would allow complete sequencing of the protein TLP and would also exclude the possibility

that TLP and Corin are the same protein. Corin shows high homology with TLP and is present in various isoforms in the lung.



From the first analysis of the immunoprecipitation followed by Western blotting TLP (Fig. 2) and corin seem to localize at the same height (around 50kDa) and are recognized by the same antibodies.



We are currently trying to get enough staff of immunoprecipitated TLP in order to make the protein sequence and at the same time we would like to immunoprecipitate fragments of the two proteins (TLP and Corin). If the fragments from cutting with thrombin proved to be the same the data would support the hypothesis that TLP and Corin are the same protein.

At the same time we are arranging to get a plasmid that allows us to transfect and over-express human Corin with the purpose to assess by Western blotting (with anti-TLP and anti-Corin antibodies) whether the two proteins are actually the same protein or are different proteins [17].

3 Conclusions

TLP is a tumor-associated antigen and a 100 kDa protein overexpressed in lung tumors and other epithelial adenocarcinomas [12]. It is immunogenic in humans as shown by serum antibodies [18].

Since TLP is a fragment of a protein identified in extracts of human NSCL cancer [19]; [20] and colorectal cancers (CRC) [21]; [22] and its sequences stimulate cytotoxic immunoresponse in humans and animal models, it is possible to design potential active and passive immunotherapies for NSCL cancer and CRCs based on TLP epitopes and humanized antibodies [23]; [24].

Therefore, TLP is a platform technology that can be used for: - a cancer diagnostic test to measure TLP levels in serum [12]; [15]; - a cancer therapy monitoring test - might measure changes in TLP levels in response to therapy [11]; [15]; - a cancer therapy - fragments of TLP can be used to stimulate immune response to attack existing tumors [10]; [25]; - a cancer vaccine: at-risk populations could be inoculated with TLP fragments to stimulate immune response to undetected or newly developing tumors [26]; [27].

We can use sequence information to express proteins, and then screen against phage antibody libraries for "pull down" for single chains of antibodies and test antibody against cell lines, colon and lung tissue microarrays.

Finally, the ability of the immune system to recognize TLP, thus enabling development of a vaccine approach for therapeutic application, represents a main target of this field of research.

[2] DW. Weiss. Tumor antigenicity and approaches to tumor immunotherapy. An outline In: current topics in microbiology and immunology. Berlin, Heidelberg, New York: Springer Verlag, 1980.

[3] MD. Prager et al. Immunity induction by multiple methods, including soluble membrane fractions to a mouse lymphoma. *J Natl Cancer Inst* 51: 1607, 1973.

[4] A. Hollinshed et al. Immunogenicity of a soluble transplantation antigen from adenovirus 12 - induced tumor cells demonstrated in inbred hamsters (PD-4). *Can. J. Microbiol* 18:1365-1369, 1972.

[5] OJ. Finn. Molecular origins of cancer: Cancer immunology. *N Engl J Med* 358: 2704-2715, 2008.

[6] M. Vergati et al. Strategies for cancer vaccine development. *J of Biomedicine and Biotechnology*. 2010

[7] S. Perez et al. A new era an anticancer peptide vaccines. *Cancer* 116: 2071-2080, 2010.

[8] G. Tarro, A. Pederzini, G. Flaminio, S. Maturo. Human tumor antigens inducing in vivo delayed hypersensitivity and in vitro mitogenic activity. *Oncology* 40: 248-254. 1983.

[10] G. Tarro. Present and future of cancer immunotherapy: A. Sagripanti, C. Gagliardi, A. Carpi. G. Tarro. , editors. *Progress in Medicine and Surgery, Proc. Nat. Meeting, San Romano (Pisa) 13 April 1991, 181-186 ETS, Publisher, Pisa. 1991.*

[11] G. Tarro, C. Esposito. Progress and new hope in the fight against cancer: Novel developments in early detection of lung cancer. *Int Med* 10: 7-11. 2002.

[12] G. Tarro. Tumor liberated protein from lung cancer and perspectives for immunotherapy. *J Cell Physiol* 221:26-30. 2009.

[13] G. Murphy et al. Phase I clinical trial: T-cell therapy for prostate cancer using autologous dendritic cells pulsed with HLA-A0201 – specific peptides from prostate-specific membrane antigen. *Prostate* 29: 371-380, 1996.

[14] FO. Nestle et al. Vaccination of melanoma patients with peptide-02 tumor lysate-pulsed dendritic cells. *Nat Med* 4: 328-332, 1998.

[15] G. Tarro, A. Perna, C. Esposito. Early diagnosis of lung cancer by detection of tumor liberated protein. *J Cell Physiol*. 203: 1-5. 2005.

The author declares no conflict of interests.

4 References

[1] BE. Rennik. Cancer immunotherapy: facts and fancy. *J Clin* 29: 362-365, 1979.

SESSION

COMPARATIVE SEQUENCE, GENOME ANALYSIS, GENOME ASSEMBLY, AND GENOME SCALE COMPUTATIONAL METHODS

Chair(s)

TBA

On Identifying Metabolic Functions of Noncoding RNAs in *S. cerevisiae*

Eric Struminger Max H. Garzon Sungchul Ji
The University of Illinois, Urbana *The University of Memphis* *Rutgers University*
 strumin1@illinois.edu mgarzon@memphis.edu sji@rci.rutgers.edu

Abstract

We produce putative biological functions of over 150 non-coding RNAs in *S.c.*, out of over 2,800 unknowns RNAs, together with an analysis that provides confidence levels, obtained using two major computational intelligence techniques, multilayer perceptrons (MLPs) and Self-Organizing Maps (SOMs). The identifications fall in two groups, depending on the level of confidence with which the function is being assessed (high and low). In the remaining group of RNAs (over 2,700), “hard core” RNAs remain elusive to identification. The first two groups of putative categories may be representing a new ontology worthy of further research and validation by the biological community, given other successes in the application of MLPs and SOMs as less researcher-biased classification tools. Although analyses of microarray data are plentiful by other techniques, a novel contribution of this paper is that the analyses has been carried out in the researcher-independent ontology implicit in the inherent properties of neural network, which are based solely on the given data.

Key Words: *S. cerevisiae*, biological function, noncoding RNA, microarray data analysis, neural networks.

1. Introduction

The human genome project of the 1990s marked a critical transition in the study of biological organisms and has transformed theory and practice of experimental biology. While sequencing has been in itself a challenge, enormous progress has brought genome sequencing to the verge of a commodity that can be had for well under \$1,000 in the near future [1].

As already anticipated by many, this progress has brought to the front the second and more important phase of the post-genome project era, i.e. the elucidation of the molecular mechanisms underlying

the genotype-phenotype coupling. Despite the enormous amount of data generated by genome sequencing, they pale in comparison by the extraordinary amount of analytic and computational resources required to do the bioinformatics of assembling an accurate picture of the complex molecular interactions among genes, RNAs and proteins in living cells that sustain life. A primary problem in this program is the identification of the metabolic functions of long non-coding RNAs (ncRNA) usually defined as non-polyadenylated RNAs with greater than 200 nucleotides [8,9,10]. Intense research over the past decade or so has demonstrated that many noncoding RNAs participate in regulating cell functions including RNA splicing, RNA editing, transcription factor transport, translation, and transcript degradation [9].

In this paper, we focus on identifying the metabolic functions of non-coding RNAs (ncRNAs) in one of the biologist’s favorite organisms, *Saccharomyces cerevisiae* (*S.c.* hereafter), or baker’s yeast [11]. This organism was sequenced in 1996 (the first eukaryotic genome that was fully sequenced, annotated, and made publicly available) and shown to consist of over 6,000 genes [6][11]. Of these, about 3,000 genes code for proteins with known metabolic functions, but the remaining genes code for RNAs that do not encode proteins and hence their metabolic functions are unknown. The RNAs encoded by such genes are referred to as non-coding RNAs [9]. Using two independent techniques described in Section 3, we have identified the possible metabolic functions of over 170 of these ncRNAs with greater than 90% confidence and propose possible molecular mechanisms underlying their suggested functions in Section 4. Finally, some discussion of the credibility of this assessment, its interpretation and general biological significance is presented in Section 5.

2. The measurement of the transcriptome

In this Section we describe the tools used and the data

utilized to train and enable neural network to make the predictions subject of this paper.

DNA microarrays have been used as a means to ascertain the possible functions of non-protein coding RNAs through analyses of their transcript levels (TL). In the study reported in [3], the *S.c.* strain BQS252 was grown overnight at 28 °C in YPD medium (2% glucose, 2% peptone, 1% yeast extract) to exponential growth phase ($OD_{600} = 0.5$) [3]. Cells were recovered by centrifugation, resuspended in YPGal medium (2% galactose, 2% peptone, 1% yeast extract), and allowed to grow in YPGal medium for 14-15 hours. Cell samples were taken at 0, 5, 120, 360, 450 and 850 minutes after the glucose-galactose shift. The 850 minute sampling time corresponded to the exponential growth phase in the YPGal medium. The TLs were measured with DNA arrays as described in [3]. The total amount of poly(A) mRNA per cell was measured and used to normalize the microarray signals. The original data contained readings of 5,914 mRNAs (ORFs) which were assigned to 531 metabolic pathways (henceforth referred to as “categories”) as illustrated in Fig. 1. Therefore, there were 2,817 remaining RNAs (and their transcripts designated as ncRNAs, or noncoding RNAs) whose metabolic functions remain unknown. Each RNA in the data is identified by a name, a category (perhaps “Unknown”), and mRNA expression level readings taken at the six different time points.

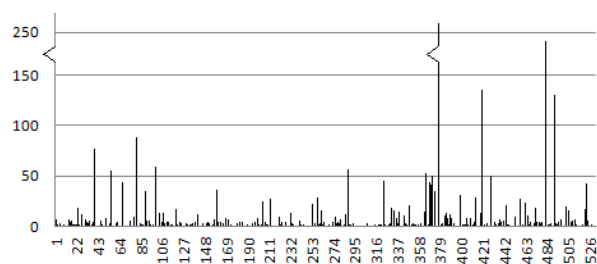


Figure 1. Histogram of the genes assigned to the 531 biological functions (categories) in *S.c.*

3. Genomics with Multi-layer Perceptrons

To determine the biological function of the ncRNA (i.e., RNAs with unknown functions), several methods exist in the literature that would allow an “educated extrapolation” based on a computational analysis of their microarray expression profiles (also called RNA trajectories, or RNA traces). Techniques vary from statistical approaches, to neural networks, to evolutionary algorithms, to ad-hoc approaches such as chaos theory. Neural networks appear most appropriate for this task because of their proven generalization ability, based solely on a large data corpus, as is the

case here. Fig. 1 shows the distribution of the frequency of these genes across the 531 categories. Most common among these are protein synthesis, transcription, transport, cytoskeleton, DNA replication, mRNA splicing, cell wall genesis, protein degradation, glycosylation, and signaling.

Neural networks can be obtained through the use of learning algorithms that “discover” patterns in the known data and enable them to extrapolate answers to unknown data. We used two types of neural networks, multi-layer perceptrons (MLPs, for which supervised learning algorithms such as the well-known backpropagation are available [7, Chap 4].) Another approach with self-organizing maps (SOMs, for which unsupervised learning algorithm are available) is described in Section 4. Both were trained on the data consisting of the six-feature vectors describing the expression profile of a given RNA molecule. The reader is referred to any textbook on neural nets ([7], for example) for further background details about these types of neural networks.

Input RNAs must be encoded as so-called *features* (i.e., numerical vectors) to train a neural network. Given the mRNA expression profile data available, the easiest way was to use the 6-feature mRNA expression level for each RNA as a 6D input, and the categories (biological functions) an integer 1-531 as the output.

To fit the model of a multi-layer perceptron, the known data is usually partitioned into a training set and a testing set. Through trial and error, it was determined that selecting 33% of the mRNAs with known categories for training and the remaining 67% for testing gave the best results. In order to preserve the proportion of RNAs in the various categories, a random selection was made from data in each category in these proportions. To avoid unintended patterns in the data (due to alphabetic presentation by names, for example) the exemplars were presented at random in the learning phase. The RNAs with unknown categories were stripped down to only their 6D-feature vector so they could be used as an input after testing to determine results. In order to improve training, categories with less than 30 mRNAs in them were excluded, as they would probably lead to memorization of the inputs by the MLP and thus poor generalization (more below.) Although that reduced the number of RNAs to attempt identification to 1,696 in 25 categories, the levels of confidence for the predictions increases substantially by avoiding poor generalization.

3.1 Training Phase

There were several possible approaches to training a network to predict ncRNA function (as a category). The ideal result is a single network able to correctly

classify all RNAs in S.c. . This approach failed for such a large number of RNAs and functions despite many attempts in strategy for coding, for architecture selection and for training. However, the alternate approach of training an individual network for each category separately, was very successful for all categories. That means that 531 networks were trained on the same data (creating separate exemplar sets by changing the desired answer to “1” for those RNAs in the most frequent category “376” and “0” for all other RNAs, for example). Therefore a neural network was created by backpropagation for the purpose of determining whether an RNA (with known or unknown function) is a member of a given category or not. Since all the networks received the same input, these 531 networks can be in fact assembled to produce a single MLP with 6 inputs and 531 outputs, each feature value in the output coarsely coding for the categories as a 531D-Boolean vector) whose 1 values may produce appropriate categories for every mRNA.

To create a specific neural network for a given category, the standard training procedure for MLPs was followed: a neural network was first created with randomly assigned weights with mean 0 and small std (standard deviation.) In addition, multiple architectures were used in which only the number of nodes in the hidden layer(s) were changed. Neurons in these layers were assigned sigmoid functions as transfer functions (such as the inverse trigonometric tangent, $arctan$) for the input and all hidden layers and a pure linear function for the output neuron/node, producing a continuous value in the interval [0,1] for an output to other neurons, or as output of the neural network.

In backpropagation training, there is usually an optimal number of epochs (i.e., repeated presentations of the training data) that gives high values for both. Many choices for the number of epochs to train with were used until each architecture's optimal number of epochs was determined. This was achieved by choosing the number of epochs where the training and testing percentages were the highest but there was no large drop off from training percentage to testing percentage, i.e., there was little evidence of “memorizing” answers. If the testing percentage is high (> 90%), it was considered a success because the probability is greater than 90% that the predicted category for each RNA is the correct category.

3.2 Testing Phase

Once optimal training and generalization rates were obtained, the networks were then put to use in the testing phase. The ncRNA in the testing set put aside were, naturally, coded in the same form (as a 6D vector

array of mRNA expression levels) and given to the network for a specific mRNA as an input in order to produce a putative category in which the network would put each. The output determines an answer to membership in a given category by assigning a certain threshold (0.5, consistent with the data) in order to predict whether the RNA belongs in a given category (≥ 0.5) or not (< 0.5). The generalization performance (i.e., testing accuracy) is based on the percentage of RNAs the neural network predicts correctly for a given category.

Once the training was complete and a satisfactory network was obtained for each mRNA from various architectures using a MLP, we selected the top three architectures that worked effectively on the training and testing data for most mRNAs, as can be seen in Table 1 and Fig. 2. These rates are considered very good to excellent for the typical performance of MLPs in this type of problem. The mean square error (MSE) over all exemplars in Fig. 2 was used along with the training and testing accuracy percentages to determine how well the neural network is predicting the categories among the known data, and so build some confidence interval for the prediction phases, as shown in Fig. 3. The Mean Squared Error (MSE) is the usual average error between predicted values (0 or 1) and actual network outputs for a given RNA. Therefore, the MSE is a measure of the overall quality of the predictions by the neural network.

Table 1. Average Training and Testing Accuracies for three best MLP classification of ncRNA in S.c.

Architecture	Avg. Train %	Avg. Test %
[6 4 1]	99.84	99.82
[6 6 3 1]	99.83	99.83
[6 12 9 1]	99.84	99.80

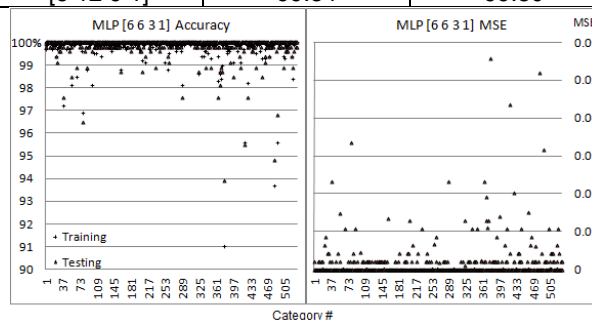


Figure 2. Overall performance of the MLPs on the 3,082 known mRNAs, given by the accuracy and MSE in the training and testing phases. Consistently with the training data, a prediction is considered accurate if the MLP produces a response at or over 0.5 for a target 1, whereas it is in error if the value is under 0.5 for a target 0, Low MSEs indicate confident predictions overall.

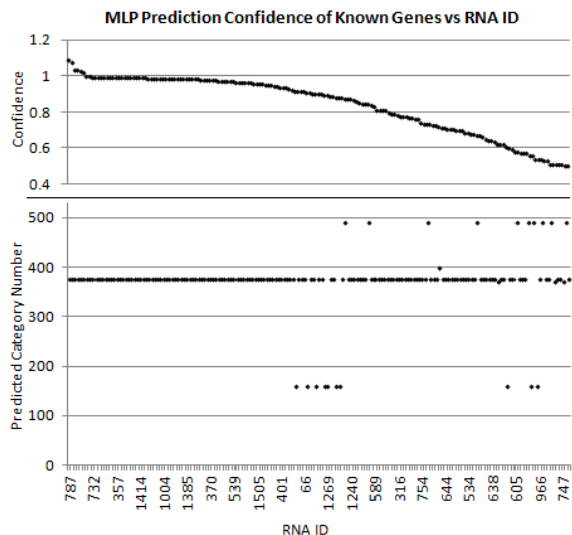


Figure 3. Predicted biological function of ncRNAs for the testing phase by three MLP architectures.

3.3 Prediction Phase

Once optimal architectures with high training and generalization rates were obtained on the known data, a composite neural network was used to predict categories based on the 6D-vector of the ncRNA with unknown categories as opposed to mRNA with known categories as in testing. The results are displayed in Fig. 3. The results from the three architectures are not equal, as expected from the results in the testing phase, or even within the same architecture for a particular mRNA, so some more detail is required.

The predictions can be classified into three basic groups. The first group consists of 153 ncRNAs that are being claimed to be in a unique category (i.e., biological process.) Fig. 4 (top) shows them, as sorted by the level of confidence (as defined above) with which the network makes the prediction; Fig. 4 (bottom) shows the corresponding categories.

The second group consists of 391 mRNAs that are being claimed by more than one category. This may appear contrary to the data, in which every mRNA gets assigned a unique category. Upon reflection, however, it makes sense biologically because an mRNA may be involved in several biological processes. We interpret RNAs in this second group as being so, on the evidence presented by the corpus of data.

The third group consists of the remaining 1,596 mRNAs, so-called “orphans” because they were not assigned to any given category by the MLP. This conclusion can be interpreted in two different ways.

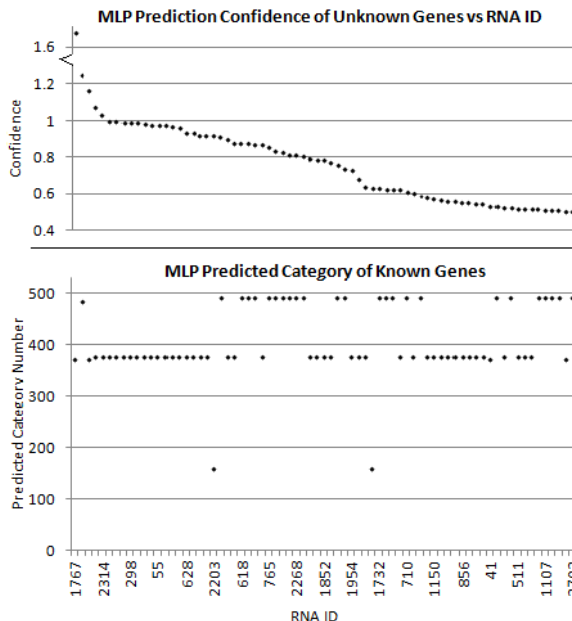


Figure 4. Predicted biological function of ncRNAs for the prediction phase by three feedforward MLP architectures that tested as shown in Fig. 3.

One possibility is that the MLPs are not “smart” enough to tell in which category they are. The alternative possibility is that there are yet unknown biological processes that are not present in the original data, so that the MLP is actually discovering hitherto unknown processes at play in *S.c.*, or that these mRNAs do not share many features with the known ones to allow MLPs to ascertain one category.

Table 2. Classification of biological function in *S.c.* by three best performing MLP architectures.

mRNA Group Prediction\	Known	Unknown
1 (Unique Cat)	171	74
2 (Various Cats)	3	3
3 (Orphans)	2,908	2,740
Totals (5,899)	3,082	2,817

4. Genomics with Self-Organizing Maps (SOMs)

A second set of predictions was produced by categorizing the ncRNAs (i.e., RNAs with unknown functions) using another neural network technique, the so-called Self Organizing Map (SOM) [7, Chap 9]. SOMs afford an unsupervised learning algorithm. While supervised learning imposes a specific assignment of biological function by requiring a label (“teacher”) on each input for training, SOMs takes the

inputs with no labeling and thus produces a classification with no *a priori* assumptions about how the inputs should be clustered together, based solely on any patterns of similarity that might be identified in the course of training. Therefore they have the potential to identify objective classifying criteria that may be complementary (or even conflicting) with MLPs or even individual researchers' ontologies, but which, on the other hand, might suggest some more objective criteria on the transcriptomics of *S.c.* for further analyses.

SOM training uses the same input as the multi-layer perceptron (a 6D-feature vector) but produces a so-called *topological map* of "locations" or "nodes" on a metric space (such as a common geometric plane or 3D space) whose relative proximity (or distance) is arranged to capture the relative similarities (or differences, respectively) among the clusters represented by the various locations/nodes. Therefore, the features that are mapped in the prediction phase to the same node are close enough to be considered to belong to the same category.

4.1 Training Phase

Various topologies were tried to develop a SOM of the given data, including 1D and 2D architectures of various sizes (18x18, 24x24, 30x30, and 40x40). Eventually, as before, we selected the top performing couple of architectures, namely 40x40 locations/nodes. The 6D-feature vectors of the entire set of 91 RNAs representing approximately the top 50%+ of the mRNAs with known categories with 30+ mRNAs were taken from the original data and used as the input to the SOM for training. The training consisted in presenting an input vector to the SOMs, identifying the location with maximum output on that input (called "the winner"), strengthening the connections to it so that next time the same will happen again, and weakening the connections to other nodes proportionally to their topological distance (here in the plane) from the winner (hence the name "winner-takes-all" used to describe this type of network.) The SOM maps were trained for 15,000 epochs (3,000 in the ordering phase and the rest in the converging phase [7, Chap. 7].)

Once trained, the SOM will do a "forward" pass and produce a classification into a *unique* location (the winner for that input) for any given node. RNAs mapped to the same node can be regarded as belonging to the same category, which then has to be identified using prior knowledge about the data.

4.2 Labeling Phase

Once the SOM is trained, it is necessary to figure out the meaning of the classification being made by it, before proceeding to the prediction phase. That requires inspection of the results in light of preliminary prior experience with the data in order to figure out what input patterns each node may be capturing.

Ideally, every location/node in the SOM should be regarded as defining a single category (here, a biological process), although RNAs in a category could be mapped to several locations, which together would represent that category. In particular, if a location gets only RNAs from a single category, it is clear it should represent that category. Locations capturing an overwhelming number of mRNAs (over 80%) from a single category were also considered to be "uniquely" labeled by that category. The full category itself is thus represented by all such locations. There were 932 such locations and they turn out to represent about 244 unique categories shown in Fig. 5, a hit rate of over 94%. The top categories labeling locations with high confidence are shown by the distinct labels in Table 3.

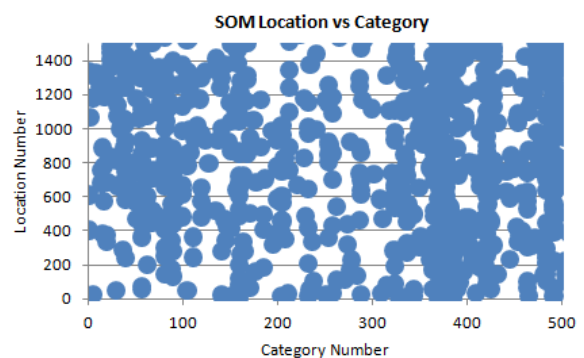


Figure 5. Results of classification by a 40x40 SOM for mRNAs in 91 biological functions capturing all unknown mRNAs, trained over 15,000 epochs. Shown are the 932 uniquely labeled locations (as described in the text) representing 244 (out of 259, over 94%) categories (biological functions) over the 1,600 locations. The radii of the circles are proportional to the number of mRNAs in a category being mapped to each location.

The remaining locations get RNAs from more than one category, and therefore the label (the category they represent) is not obvious. They might represent mRNAs involved in several higher-level biological functions. More sophisticated analyses or additional data may be required to produce putative single categories for these mRNAs.

4.3 Prediction Phase

Once the SOM has been labeled, we proceeded to obtain its classification for the unknown ncRNAs by presenting them to the network. This time, the unknown ncRNAs can be placed in two groups, those for which a highly confident prediction is made, and those for which the prediction is unclear. Some of the results are shown in Table 3. Once again, further analyses or additional data may be required to produce putative categories for the latter mRNAs.

Table 3. Summary of predicted unique identification with high confidence of most important biological function for 74 mRNAs in *S.c.* by MLPs and/or SOMS architectures described above. The prediction for boldfaced mRNAs are matched by both types of architectures and therefore can be assigned a very high level of confidence. Other putative predictions are made by the MLP (second column) or SOM (third column) architecture only and so are made with less confidence (not all of them are shown for lack of space.)

Gene	MLP	SOM
Glycolysis	YDR134C YAL037C-B	YDR053W YLR190W YJL161W YFR017C YAL037C-B YCR013C YKL164C YHLO34W-A
Protein Synthesis	YMR116C YLR076C YLR150W YEL034C-A YMR093W YCLO46W YMR049C YMR290C YHR214W YGR285C YPL126W YLR123C YER156C YAL036C YBR025C YNL255C YPL142C YGR211W YEL040W YPL093W YPL226W YGR050C YOR146W YLR413W YOR277C YGR160W YKL056C YGL102C YLLO44W YER006W YBL077W YGR090W YLR339C YJR071W YPR044C YDR417C YLR179C YOR091W YOR309C	YOR354C YDR169C YCR045C YDR116C YHR070W YPL093W YCLO46W YGL102C YKR079C YNLO10W YMR290C YER156C YHR049W YOR359W YDR526C YDR029W YMR141C YCLO49C YPL067C YKR046C YGL068W YOR084W YBR235W YIL157C YEL034C-A YLR339C YPL142C YLO036W YLR236C YJR070C YOR091W YGR090W YLR150W YGR285C YHR140W YLR123C YEL033W YNL255C YLO077C YCLO20W
Transport	YDR525W-A YGL056C YDR222W YPL054W YOR382W YAL034C YLR282C YLR327C YHLO21C YMR082C YOL152W YER067W YLR089C YMR206W YMR291W YGR039W YNR002C YOR383C YJL217W YNL208W YBR212W YILO57C YGR243W YGL157W YAL060W	YNL191W YDR219C YGR236C YILO64W YHR149C YBR089W YJL217W YBR212W YBR053C YJL119C YDLO60W YAL022C YJR146W YMR069W YLR168C YPR023C YKL053W YMR082C YHR199C YGL056C YMR291W YPR154W YNL177C YBL100C YNR002CYELO57C YHR076W YLO031C YMR247C YJL213W YDR505C YNL278W
Protein Folding	YOR120W YLR110C YOR121C	YMR194C-A YKL100C YLO41W YER080W YDR442W YER079C-A YDR344C YPL113C YDR033W YGR136W YJR044C
Transcription	YIL131C	YGL111W YGR128C YBR267W YGR123C YNL162W-A YLR217W YGR168C YGL052W

5. Discussion and Conclusions

We have used two major computational intelligence techniques, MultiLayer perceptrons (MLPs) and Self-Organizing Maps (SOMs), to identify putative biological functions of over 170 non-coding RNAs in *S.c.*, together with an analysis that provides confidence levels about the identifications being made, in three groups. In the first, 174 of the unknown RNAs are given putative biological functions by one of the two networks. In the second group, 26 of the remaining RNAs are given a putative biological function that is

less certain but still worthy of consideration given that none has been hitherto been suspected. In the remaining group of RNAs, “hard core” RNAs remain elusive to classification. The first two groups of putative categories may now be subjected to further validation by the biological community, and may be representing a new ontology worthy of further research, given other successes in the application of MLPs and SOMs [7].

Some of the results presented here can be ported to other cell systems. The theory of grand unification [4] holds that information about a shared gene and associated proteins contributes to our understanding of all the diverse organisms that share it, so that knowledge of such roles illuminate and provide strong inference of its role in other organisms. For example, about 12% of the worm genes (~18,000 genes) encode proteins whose biological roles could be inferred from their similarity to their putative orthologues in *S.c.* (or about 27% of the *S.c.* genes) [6]. Further, most of these proteins have been found to have a role in the ‘core biological processes’ common to all eukaryotic cells, such as DNA replication, transcription and metabolism [4]. It would not be surprising if the same is true of genes with noncoding RNAs.

Finally, a word of caution is in order in assessing the results presented in this paper in the proper context. It is important to keep in mind (i) that a given category refers to either a gene (which is a DNA molecule) or its transcript (which is an RNA molecule), depending on the context, and (ii) that the intracellular level of an RNA molecule at any given time (referred to as the transcript level, TL) is determined by the balance of two factors – the transcription rate (TR) and transcript degradation rate (TD), which can be algebraically represented by the following equation:

$$dTL/dt = TR - TD \tag{1}$$

where dTL/dt indicates the rate of change in TL with time [2]. Garcia-Martinez et al. [3] measured both genome-wide TL and TR simultaneously in *S. cerevisiae* following glucose-galactose shift. Some examples of their data are plotted in Figure 6. As evident in Fig. 6 (top), the average behaviors of the glycolytic and oxidative phosphorylation, RNA trajectories (also called TL kinetics, TL traces, or gene expression profiles) reflect the metabolic functions of the genes coding for the RNA molecules in a goal-directed manner. The glycolytic RNAs decrease, since they are no longer needed due to the removal of glucose, while the oxphos RNAs are required to metabolize ethanol left over from the previous glucose metabolism and the new nutrient galactose [2, 5].

One of the most surprising findings of Garcia-Martinez et al. [3] is that the rate of change in TL, i.e., dTL/dt , can vary independently of TR. This observation is supported by the fact that, although TR can change in the same direction for the glycolytic and oxphos pathways, their TL kinetics can be opposite (see the TL and TR traces between 5 and 360 minutes in Fig. 6, top and bottom.) This finding cannot be explained unless we take into account the transcript degradation rate (TD), in agreement with Eq. (1). Thus, an important conclusion one can draw from Eq. (1) is that “It is impossible to infer the genes responsible for metabolic functions solely based on analyzing gene expression profiles.” [9]. However, theoretical considerations strongly indicate that it should be possible to infer the metabolic functions of unknown RNAs based on the similarity of their TL traces with those of known RNAs [5]. Although large-scale prediction [8] have been used before in analyses of microarray data, a novel contribution of this paper is that the analyses has been carried out in a researcher-independent ontology implicit in the inherent properties of neural networks, which are based solely on the given data.

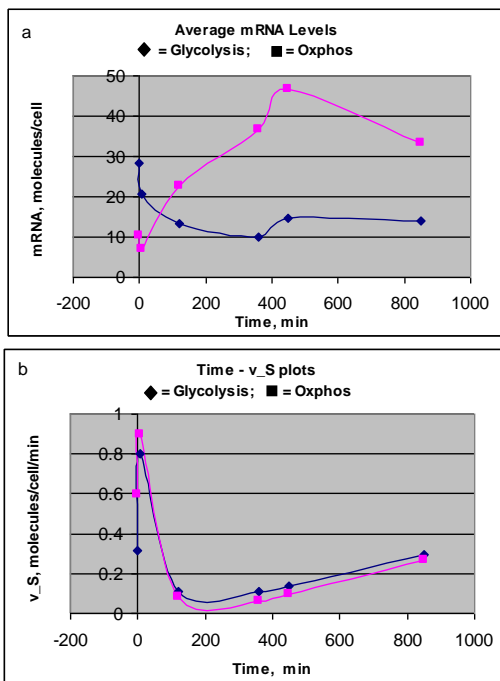


Figure 6. The average time courses of the transcript levels (TL) and rates (TR) of 14 each of the glycolytic and respiratory (also called oxidative phosphorylation, or oxphos) genes, as described in [2].

6. References

- [1] <http://singularityhub.com/2008/12/30/whole-genome-sequencing-to-cost-only-1000-by-end-of-2009/> (accessed July, 2011).
- [2] S. Ji, A. Chaovalitwongse, N. Fefferman, W. Yoo, W. & J.E. Perez-Ortin, “Mechanism-based Clustering of Genome-wide mRNA Levels: Roles of Transcription and Transcript-Degradation Rates”. In: *Clustering Challenges in Biological Networks*, S. Butenko, A. Chaovalitwongse, and P. Pardalos, (eds.), World Scientific Publishing Co., Singapore, pp. 237-255.
- [3] J. Garcia-Martinez, A. Aranda, and J.E. Perez-Ortin, “Genomic Run-On Evaluates Transcription Rates for all Yeast Genes and Identifies Gene Regulatory Mechanisms”, *Mol Cell* **15** (2004):3303-313, (at <http://scsie.uv.es/chipsdna/chipsdnae.html#datos>).
- [4] The Gene Ontology Consortium, “Gene Ontology: tool for the unification of biology”, *Nat Genet* 2000 May; 25(1), 25-29.
- [5] S. Ji, *Molecular Theory of the Living Cell: Concepts, Molecular Mechanisms and Biomedical Applications*, Springer, New York, 2012.
- [6] A. Goffeau, et al. “Life with 6000 genes”. *Science*. 1996;274:546
- [7] S. Haykin. *Neural Networks and Learning Machines*, 3rd ed. Pearson, Prentice-Hall, New York, 2009.
- [8] Q. Liao, C. Liu, X. Yuan et al., Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Research*, 2011, **1-15**
doi:10.1093/nar/gkq1348.
- [9] J.S. Mattick (2004). RNA regulation: a new genetics? *Nature Reviews Genetics* **5**, 316-323.
- [10] T.R. Mercer, M.E. Dinger, J.S. Mattick, “Long non-coding RNAs: insights into functions”, *Nature Reviews Genetics* **10**: 155-159.
- [11] *Saccharomyces cerevisiae* Genome Snapshot/ Overview at <http://www.yeastgenome.org/cache/genomeSnapshot.html>

De novo identification of “heterotigs” towards accurate and in-phase assembly of complex plant genomes

Jared C. Price¹, Joshua A. Udall², Paul M. Bodily¹, Judson A. Ward³, Michael C. Schatz⁴, Justin T. Page², James D. Jensen¹, Quinn O. Snell¹, and Mark J. Clement¹

¹Computer Science Department, Brigham Young University, Provo, UT, USA

²Plant and Wildlife Sciences Department, Brigham Young University, Provo, UT, USA

³Department of Horticulture, Cornell University, NYSAES, Geneva, New York, USA

⁴Simons Center for Quantitative Biology, CSHL, Cold Spring Harbor, New York, USA

Abstract—*Accurate and in-phase de novo assembly of highly polymorphic diploid and polyploid plant genomes remains a critical yet unsolved problem. “Out-of-the-box” assemblies on such data can produce numerous small contigs, at lower than expected coverage, which are hypothesized to represent sequences that are not uniformly present on all copies of a homologous set of chromosomes. Such “heterotigs” are not routinely identified in current assembly algorithms and could be used for haplotype phasing and other assembly improvements for such genomes. We introduce an algorithm which attempts to robustly identify heterotigs present in the assembly of a highly polymorphic diploid organism. The algorithm presented is for use with the 454 platform and for diploid assembly, but is readily adaptable to other sequencing platforms and to polyploid assembly.*

Keywords: heterozygous, genome, assembly, plant, heterotig, raspberry

1. Introduction

1.1 Background

Genome assembly is a relatively young field, but one which has been the subject of intense research. Motivated by a desire to reconstruct the human genome as rapidly as possible, the Whole Genome Shotgun strategy for genome assembly was introduced [1]. In this approach, genome structure inference is left entirely to software which takes as input a huge number of short DNA sequences (“reads”) sampled from the entire genome. Although this approach was initially met with skepticism, a seminal paper provided the necessary proof of concept [2] and, due to its simplicity and cost-effectiveness, this approach has dominated genome projects since.

There are two primary classes of algorithms that are applied today to the Whole Genome Shotgun assembly problem. The first approach is referred to as the “overlap-layout-consensus” approach and the second approach is based on De Bruijn graphs. See [3] for a comparison of the two. We

will focus on the overlap-layout-consensus approach, but the ideas regarding identification of heterotigs are applicable to both.

Overlap-layout-consensus assemblers often construct a data structure known as a “contig graph”. A contig is simply a contiguous sequence of nucleotides inferred, via alignment of the input reads, to be present in the target genome. For various reasons, but primarily because of repetitive sequence, these contigs can essentially never be extended to full chromosome length in reasonably complex genomes, using current technologies. For this reason, the contig graph must represent not only the contigs themselves but also all of the possible adjacency relationships between contigs that are supported by the alignments. A common approach to representing this information, and the approach used in the 454 software, is to let the vertices of the graph represent contigs and the edges represent adjacency relationships between contigs. Because contigs have polarity (a 5’ and a 3’ end) the edges do not directly connect contigs, per se, rather, they connect specific ends of contigs. For example, an edge may indicate that the 5’ end of contig 25 is adjacent to the 3’ end of contig 1.

Critical to the upcoming discussion is a clear understanding of why assembly algorithms tend to collapse repetitive sequence into a single contig and the effect this has on the contig graph. Consider the case where a sequence of nucleotides (longer than the read length) occurs in the genome 3 times. Reads which are sampled from entirely within this repetitive sequence will align to each other with near perfect identity and will likely be collapsed into a single contig (in the absence of paired-end reads which align to unique sequences bordering the repeat). We will assume for demonstrative purposes that the sequences adjacent to each of the 3 copies are themselves unique. In the contig graph, the 5’ end of the repeat contig will be adjacent to 3 different contigs, as will the 3’ end (see Figure 1).

Notice that in Figure 1, in order to extend the contig that is currently represented as a collapsed 3-copy repeat any farther than the repeat sequence itself, you must accurately select a particular pair of contigs (one adjacent to the 5’

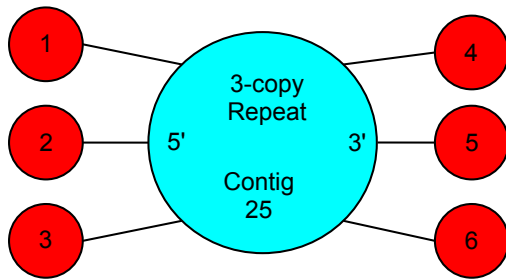


Fig. 1: A 3-copy repeat (collapsed in the assembly into a single contig) in a contig graph. Each circle is a vertex in the contig graph representing a particular contig in the assembly. Solid lines between contigs (more exactly between specific ends of those contigs) suggest that the contigs are adjacent to one another in the genome. Contigs 1-6 are single-copy contigs, each of which is adjacent to one of the 3 copies of the repeat. Each instance of the repeat is surrounded by a pair of contigs (one from the set $\{1, 2, 3\}$ and one from the set $\{4, 5, 6\}$). The pair of contigs surrounding a particular instance of the repeat constitute the “context” of that instance. When an assembly algorithm is unable to accurately determine the context around a particular copy of the repeat, contig extension must end at the repeat boundary. Worse, if the algorithm extends a contig through the repeat, but with the incorrect context, the resulting contig will contain sequence from 2 different locations in the genome.

end of the repeat and the other adjacent to the 3' end) with which to extend the contig. If you are not careful you might select contigs to use for the extension that are adjacent to different copies of the repeat in the actual genome, thereby constructing a contig that doesn't actually exist in the genome and whose 5' and 3' ends are in different locations in the genome. For this reason, repeats longer than the read length produce fragmentation of the contig graph and consequently smaller contigs in the assembly. The correct “context” for each copy of the repeat must be constructed carefully, usually using paired-end reads at a known distance and orientation with respect to each other.

1.2 Motivation

Highly polymorphic diploid and polyploid plant genomes have proven to be particularly difficult to assemble. Plants tolerate hybridization and polyploidization much more readily than most organisms that have been assembled by the Whole Genome Shotgun approach. These data present different challenges to assembly algorithms than those presented by highly homozygous diploid or monoploid organisms, for which traditional genome assembly algorithms are primarily designed. Notable examples of recent plant genome assembly projects include the small *Fragaria vesca* genome [4], a

relatively heterozygous grapevine variety [5] and the large and ultra-repetitive maize genome [6].

Rubus idaeus cultivar ‘Heritage’ is an important commercial variety of raspberry which holds both biological and economic interest. Heritage is resistant to many of the most common raspberry diseases and has two raspberry subspecies in its recent pedigree, namely, *Rubus idaeus ssp. strigosus* and *Rubus idaeus ssp. vulgatus*. Such a scenario is not unique to Heritage, and is very common in raspberry breeding. Furthermore, hybridization, in general, is relatively common among plants.

Despite being very similar in appearance, amenable to hybridization, and prominent in the pedigrees of many commercial varieties of raspberry, these two subspecies have historically been geographically isolated with *strigosus* being a North American variety, and *vulgatus* a Eurasian variety. Furthermore, despite both varieties often being labeled as subspecies of *Rubus idaeus* taxonomists currently favor classifying these organisms as two different species, namely *Rubus strigosus* and *Rubus idaeus*.

Until recently, and to a great extent even today, the genomes of diploid and polyploid organisms have been assembled and presented in a monoploid form. Such an approach minimizes sequencing cost (greater depth is often required by algorithms that attempt to perform true diploid or polyploid assembly) and increases algorithmic simplicity for such tasks as genome assembly, mapping reads to a reference, and viewing a genome in a genome browser. Despite these advantages, such an approach also has distinct disadvantages. For example, diploid assemblies can provide a more accurate depiction of sequence diversity within a pair of homologous chromosomes than simple mapping back to a monoploid reference can provide. This information can then be used to improve numerous downstream analyses.

Genome assemblers that provide only a monoploid representation of a diploid or polyploid organism often contain algorithms that obscure sequence diversity or, worse, produce sequence not actually present in the target genome. For example, sequence diversity can be hidden when an algorithm deals with polymorphic regions by simply selecting one of the possible paths and ignoring all other possibilities. In the context of a highly heterozygous genome, the monoploid representation of the assembly can often “jump” between different members of a homologous set of chromosomes. Worse, an assembler may deal with polymorphic regions by producing a single contig that is a composite of the polymorphic paths in the contig graph, thereby producing sequence that isn't actually present on any chromosome.

With the advent of next-generation sequencing technologies, the field of genome assembly is aggressively pursuing more accurate and comprehensive representations. The Broad Institute's ALLPATHS-LG [7] is a notable example which represents the genome as the assembler actually sees it, that is to say, as a graph, thereby maintaining important

information about sequence diversity that may otherwise have been lost. Another fascinating approach, published very recently, applies colored de Bruijn graphs to the genome assembly problem in an attempt to assemble multiple eukaryotic genomes simultaneously [8] and to handle polymorphism in a more disciplined way.

We introduce an algorithm for identifying contigs present in an assembly which represent sequences that are not uniformly present on all members of a homologous set of chromosomes. We have chosen to call such contigs “heterotigs”, and their counterparts, which are present on all members of the set, “homotigs”. The algorithm presented here leverages coverage statistics, adjacency patterns between contigs in a contig graph, and paired end reads to identify heterotigs present in the assembly of a highly heterozygous diploid organism, and has been designed for use with the 454 sequencing platform, but the concepts are readily adaptable to polyploid assembly and to other sequencing platforms. Robust identification of heterotigs enables differential treatment of such sequences within an assembly algorithm and presents opportunities for producing more accurate and more complete assemblies of highly polymorphic species.

2. Heterotig Identification

We are now in a position to more formally define the problem with which this paper is primarily concerned. Let R represent a whole-genome set of sequencing reads from a highly polymorphic diploid species. Let C represent the set of contigs produced by an assembly of R , parameterized so as to separate “heterotigs” as cleanly as possible. Let E represent the set of edges in the contig graph. Let M represent the set of all meta-data available about the assembly, for example, alignment depths for each contig, contig lengths, etc. Let H represent the set of contigs whose sequence is found on only one copy of a homologous pair of chromosomes. Given C , E , and M is it possible to determine H to within an acceptable degree of accuracy? We will use a whole-genome sequencing data set from the highly heterozygous diploid organism *Rubus idaeus* ‘Heritage’ throughout this section as an example data set.

2.1 Inference Based on Coverage Statistics

The first question that arises in the context of identifying heterotigs is whether the depth of the read alignment from which a particular contig is constructed can be reliably used to infer the number of times the nucleotide sequence that contig represents is likely to appear in the target genome.

Consider the idealized case where read sampling from the genome is truly random and there are no other sources of coverage bias, for example from PCR artifacts or cloning bias. This idealized scenario is never realized in practice but is instructive for the real-world case which we will shortly turn to. Consider further that the organism being

sequenced is diploid and expected to have very high rates of polymorphism. At every base in a particular contig there is a multiple alignment depth. Take the average of these depths across all bases in the contig and record this value as the “contig alignment depth”.

What might the probability density function of contig alignment depths in a highly polymorphic diploid assembly look like? Let’s say for illustrative purposes that we have sequenced the genome to 60x coverage, which is now routinely done with the advent of next generation sequencing. For a diploid organism, genome coverage is typically calculated in terms of the haploid genome size (total number of bases / haploid genome size), so this number is equivalent to the coverage we should expect for a single-copy homotig. We expect single-copy homotigs to be numerous and therefore expect a mode at approximately 60 in the probability density function. By this same logic, if heterotigs are indeed present in the assembly in significant amounts a mode should also be present at about half that coverage (30x). We expect there to be some breadth to the distribution around each peak and so high coverage will likely be necessary to determine if the modes are indeed present. Some of the density will be at much higher coverage (high-copy-number repeats) but we probably have no reason to expect that a particular copy number is more prevalent than another for high-copy-number repeats, so we expect no significant modes above our single-copy homotig mode.

Let’s now turn our attention to a real-world case. A recent whole-genome shotgun assembly project collected high-coverage sequence data from *Rubus idaeus* cultivar ‘Heritage’. The sequence was assembled using the 454 assembler and the resulting contigs were queried for their contig alignment depths (see Figure 2).

Close examination of Figure 2 illuminates several interesting properties of the contigs from this assembly. First, and most obviously, modes corresponding to our theoretical peaks (one peak composed primarily of single-copy heterotigs and another peak composed primarily of single-copy homotigs) are clearly discernible across contigs of all lengths. If these peaks represent what we have hypothesized, the homotig-mode to heterotig-mode ratio should be very near 2, as is indeed the case, with the value ranging between 2.05 and 2.17 for the sets of contigs examined. Could there be another explanation besides the heterotig-homotig hypothesis we have presented for the strongly bimodal distribution? If so, the alternate hypothesis must account for why the lower mode (lower in terms of the coverage value, not necessarily peak height) is at nearly exactly half the coverage of the higher mode.

More encouraging (for the purposes of heterotig identification) than the mere presence of the peaks is the observation that for many of the sets of contigs examined the density between the peaks is very low, suggesting that, at least for this data set, coverage can be used to make inference on copy

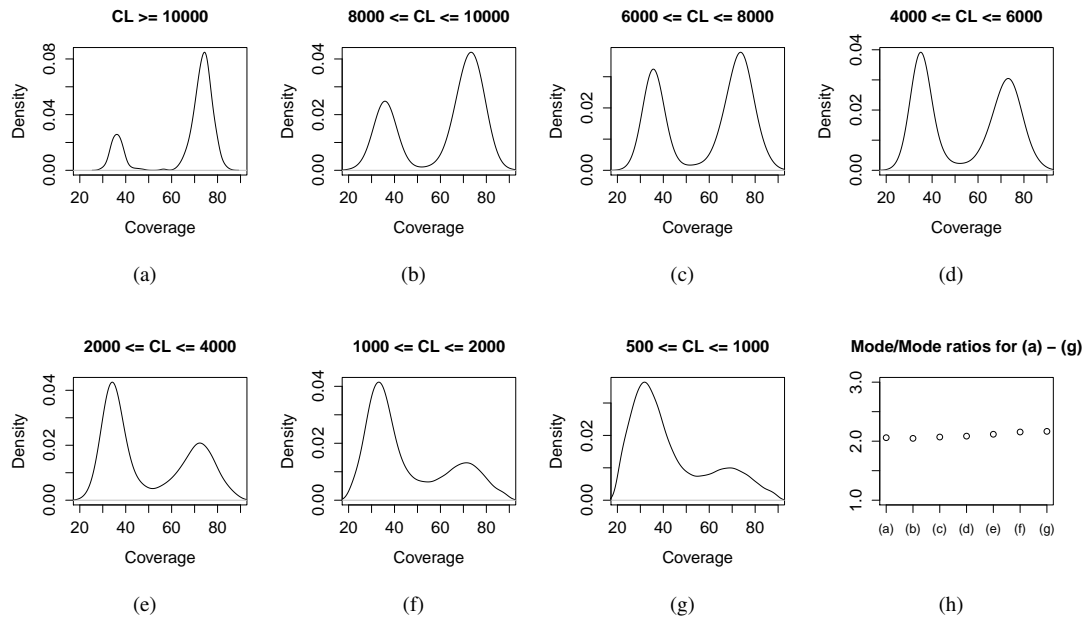


Fig. 2: (a)-(g) Probability density functions (PDFs) of contig alignment depth calculated from the set of contigs produced in an assembly of *Rubus idaeus* ‘Heritage’. Contig alignment depth is defined as the mean of the single-position alignment depths calculated at each position in the contig. Each PDF analyzes contigs within a particular length class (CL = Contig Length). Contigs with contig alignment depths outside of the interval [20, 90] are excluded. The largest contigs are predominantly at “homotig” coverage while the smaller contigs are predominantly at “heterotig” coverage. (h) “Homotig” peak mode over “heterotig” peak mode ratios for (a)-(g). The minimum value was 2.05 and the maximum value was 2.17

number. The bimodal nature of the distribution is consistent across contigs of all sizes. In contrast, the relative density under each peak differs dramatically for contigs in different length classes. The longest contigs are predominantly single-copy homotigs while the shorter contigs are predominantly single-copy heterotigs. Furthermore, as the contig length gets smaller the density between the peaks increases, although never enough to make the peaks difficult to see.

2.2 Inference Based on Contig Graph Structure

If our hypothesis from the previous section is accurate, namely, that the bimodal PDFs in the previous section suggest an extremely heterozygous diploid genome where many of the contigs are present on only a single chromosome (as opposed to both chromosomes of a homologous pair), then it is safe to assume that many of the heterotigs will be broken at boundaries where they are adjacent to single-copy homotigs. Consider a chromosome *A* and its homologous pair *B*. Now consider a single-copy homotig *C* that is present on both *A* and *B*. On chromosome *A*, *C* is adjacent to a single-copy heterotig *D*. On chromosome *B*, by the definition of heterotig, *C* must be adjacent to some sequence other than *D*, and consequently, the extension of contig *C* must be broken to account for these 2 different adjacencies. Recalling that assemblers must break contigs whenever there

is a repeat longer than the read length (see Figure 1), notice that in the context of such extreme heterozygosity, single-copy homotigs can behave similarly to 2-copy repeats, having one context in one homologous chromosome and another context in the other, providing one explanation for the extremely bimodal PDFs presented in the previous section.

Assuming this explanation is correct, such data are not likely to assemble well using traditional assembly algorithms. First, the assembly is likely to be extremely fragmented, with thousands, if not hundreds of thousands, of small contigs. There will be many more “ambiguous” adjacency relationships between contigs than would be seen in either homozygous diploid or monoploid assemblies. Furthermore, the extent to which homotig order is consistent in the two members of a homologous pair is critical to the tractability of an algorithmic solution. If the order of single-copy homotigs is strictly consistent the problem is greatly simplified. Under that scenario, only a few different signature patterns should occur in the contig graph, for example, it is probably safe to assume that under such a condition the graph should contain numerous “bubbles”, locations where a single-copy homotig bifurcates to two single-copy heterotigs which both immediately converge to a second single-copy homotig.

Length Class	Percentage	Length Class	Percentage
length \geq 10000	78 %	length \geq 10000	2 %
8000 \leq length \leq 10000	76 %	8000 \leq length \leq 10000	2 %
6000 \leq length \leq 8000	78 %	6000 \leq length \leq 8000	6 %
(4000 \leq length \leq 6000)	82 %	4000 \leq length \leq 6000	8 %
2000 \leq length \leq 4000	86 %	2000 \leq length \leq 4000	18 %
1000 \leq length \leq 2000	89 %	1000 \leq length \leq 2000	31 %
500 \leq length \leq 1000	90 %	500 \leq length \leq 1000	41 %

(a)

(b)

Fig. 3: (a) All contigs of alignment depth between 25 and 40 in various length classes were marked as “candidate heterotigs”. The percentages given indicate the percentage of candidate heterotigs connected on either the 5’ or 3’ end to at least one contig end which participated in exactly 2 edges (suggestive of a homotig-heterotig boundary possibly being the cause for contig breakage). (b) The same as (a) except the percentage now reflects the percentage of candidate heterotigs that were found in “perfect bubbles”. See Figure 4 for the precise way in which we have defined the term “perfect bubble”.

If this scenario predominates, assembling two homologous chromosomes exhibiting extremely high heterozygosity would, to a considerable extent, reduce to the problem of identifying heterotigs, and subsequently treating heterotig-to-heterotig paired-end data differently than homotig-to-homotig paired-end data. Homotig-to-homotig paired-end data would help lay out the structure shared between the two members of the pair and heterotig-to-heterotig paired-end data could help keep one chromosome separate from the other, to as great a degree as possible, when building contigs. Notice that the higher the rate of heterozygosity in this scenario the better because it gives you more heterotig anchors for keeping each chromosome “in phase”.

What about the case where the order and orientation of the homotigs differs somewhat between homologs? This would mean that in addition to assembly “bimodality” in the sense of having significant populations of both heterotigs and homotigs, there would also be assembly bimodality in the relationships between homotigs (a certain set of relationships prevailing on one homolog, and another set of relationships prevailing on the other). For example, on one chromosome, a pair of homotigs may occur at one distance and orientation with respect to each other, yet on the homolog, the same pair of homotigs may occur at a different distance and/or orientation. Such a scenario would obviously pose tremendous difficulties to traditional genome assembly algorithms. How do you correctly estimate the singular distance between two homotigs using paired end data when there are, in fact, two distances? How do you layout a genome when there are, in fact, two different layouts? The problems posed in this scenario would require the assembler itself to also be “bimodal”, that is to say, it would have to deal differentially with each homolog. The multiple “modes” could be represented using multiple graphs or by having multiple passes through the same graph. In either case, the assembler would need robust and accurate identification of heterotigs throughout the process.

The current manuscript does not attempt to perform a comprehensive analysis of the contig graph patterns observed in the assembly of *Rubus idaeus* ‘Heritage’, however, Figure 3 provides a sense of what the contig graph looks like internally. In particular it examines what the contig graph looks like immediately around “candidate heterotigs” (contigs that appear to be heterotigs based on coverage alone). Notice that for contigs in every length class examined, large majorities of the candidate heterotigs are connected either on their 5’ or 3’ end to a contig end that participates in exactly two edges, providing a measure of supporting evidence for a homotig-heterotig boundary (a particular end of a single-copy homotig, which is adjacent to a heterotig in one homolog, would likely be adjacent to exactly one other sequence in the other homolog, thereby participating in exactly 2 edges). Furthermore, only a relatively small percentage of heterotigs are found in “perfect bubble” patterns in the contig graph, suggesting that algorithms which rely on simple graph patterns to identify heterotigs may significantly underestimate the true sequence diversity. It is also interesting that, as the average length of a set of candidate heterotigs decreases, the percentage of those candidate heterotigs found in perfect bubbles increases (see Figure 3).

3. Algorithm

Definitions:

A = Alignment depth of a contig

B_{hc} = Boolean, true if $H_{min} \leq A \leq H_{max}$

B_{per} = Boolean, true if H_{cand} is in a perfect bubble

C = A contig (a node from G_{454})

C_e = A contig end (5’ or 3’)

C_{num} = The total number of contigs in the assembly

G_{454} = a 454ContigGraph.txt file (from Newbler)

H = The true set of heterotigs

$H_c = \{C : C \in H\}$ (with high confidence)

H_{cand} = Any C where B_{hc} holds

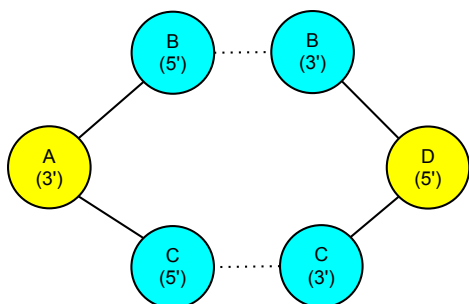


Fig. 4: Graphical depiction of a perfect bubble in a contig graph. Edges connecting contig ends are denoted with solid lines. A, B, C, and D identify contigs. (5') and (3') each identify a particular end of a contig. We say the structure is a perfect bubble when the following hold: (1) A, B, C, and D are 4 distinct contigs. (2) All 4 ends of the heterotigs (B and C) participate in exactly one edge each. (3) The ends of A and D that are connected to the heterotigs participate in exactly 2 edges each.

H_{max} = Maximum A for a heterotig candidate
 H_{min} = Minimum A for a heterotig candidate
 L = Length of a contig
 P = The set of all “paired-end flows” reported in G_{454}

Domain: $\{x : x = G_{454}\}$

Range: $\{y : y = H_c\}$

```

function IDENTIFYHETEROTIGS( $H_{min}, H_{max}$ )
  Add to  $H_c$  all  $H_{cand}$  such that  $B_{per}$  holds
  for all  $H_{cand}$  with  $L \geq 2000$  do
    if  $H_{cand}$  connects to bifurcating  $C_e$  then
      Add  $H_{cand}$  to  $H_c$ 
    end if
  end for
  while  $H_c$  grows with each iteration do
    for all  $H_{cand}$  do
      if  $P$  links  $H_{cand}$  with  $e \in H_c$  then
        Add  $H_{cand}$  to  $H_c$ 
      end if
    end for
  end while
return  $H_c$ 
end function

```

4. Discussion

We have presented evidence that complex plant genomes, particularly highly heterozygous organisms arising through hybridization or polyploidy, present unique and difficult challenges to the Whole Genome Shotgun assembly problem that are not encountered in either monoploid or homozygous

genome assembly.

When heterozygosity rates are sufficiently high, and coverage sufficiently deep, it is possible to perform de novo identification of “heterotigs” (sequences not uniformly present on all copies of a homologous set of chromosomes) via inference on a combination of coverage statistics, contig graph patterns, and paired end reads (when available). These heterotigs can then serve as guideposts in the assembly process to improve assembly quality and completeness, as well as to minimize how often the assembled scaffolds and contigs “jump” from sequence in one homolog to sequence in the other.

We have also given preliminary evidence suggesting that algorithms that identify heterotigs via very simple graph patterns, such as the perfect bubbles analyzed in section 2.2, are likely to underestimate true sequence diversity in highly heterozygous species. Furthermore, we have suggested several ways in which more robust identification of heterotigs could lead to more accurate and complete assemblies for such data. This scenario necessitates a more rigorous treatment of “heterotigs” which we begin laying the foundation for here.

We believe that robust identification of, and intelligent treatment of, such sequences could dramatically improve the state of the art with regards to the genome assembly of highly polymorphic diploid and polyploid species.

References

- [1] J. L. Weber and E. W. Myers, “Human whole-genome shotgun sequencing,” *Genome Research*, vol. 7, no. 5, pp. 401–409, 1997. [Online]. Available: <http://genome.cshlp.org/content/7/5/401.short>
- [2] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, “A whole-genome assembly of *Drosophila*,” *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000. [Online]. Available: <http://www.sciencemag.org/content/287/5461/2196.abstract>
- [3] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan, “Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph,” *Briefings in Functional Genomics*, 2011. [Online]. Available: <http://bfg.oxfordjournals.org/content/early/2011/12/18/bfgfp.elr035.abstract>
- [4] V. Shulaev, D. Sargent, R. Crowhurst, T. Mockler, O. Folkerts, A. Delcher, P. Jaiswal, K. Mockaitis, A. Liston, S. Mane, *et al.*, “The genome of woodland strawberry (*Fragaria vesca*),” *Nature genetics*, vol. 43, no. 2, pp. 109–116, 2010.
- [5] R. Velasco, A. Zharkikh, M. Troggio, D. Cartwright, A. Cestaro, D. Pruss, M. Pindo, L. FitzGerald, S. Vezzulli, J. Reid, *et al.*, “A high quality draft consensus sequence of the genome of a heterozygous grapevine variety,” *PLoS One*, vol. 2, no. 12, p. e1326, 2007.
- [6] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges,

- E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddelloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson, "The b73 maize genome: Complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, 2009. [Online]. Available: <http://www.sciencemag.org/content/326/5956/1112.abstract>
- [7] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proceedings of the National Academy of Sciences*, vol. 108, no. 4, pp. 1513–1518, 2011. [Online]. Available: <http://www.pnas.org/content/108/4/1513.abstract>
- [8] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean, "De novo assembly and genotyping of variants using colored de bruijn graphs," *Nature Genetics*, 2012.

On Using Information-Theoretic Quantities in Characterization Dissimilarity of DNA Strings

Faruil Mohd-Zaid¹, Xiaoping Shen² and Katheryn A. Farris³

April 25, 2012

Abstract. To discern similarity and differences in partial DNA strings based on dissimilarity (distance/difference) among the various SNPs, one of the challenge aspects is to select felicitous metrics or measurements. Some of information theoretic quantities are often employed in practice. Unfortunately, certain information-theoretic variables for example, information distance and mutual information, may not yield consistent results for decision-making. In this paper, we investigate the consistency of information theoretic quantities. Experiments are designed to show that the selection of measures and metrics in information-theoretic based analysis is crucial for decision-making. Future possible research directions are discussed.

Keywords: Distance metric, information theory, DNA, SNPs

1. Problem Specification.

SNPs (single nucleotide polymorphisms) are DNA sequence variation that occurs when a single nucleotide in the genome differs. SNP arrays are a type of DNA microarray that detect SNP occurrences and act as samples of DNA strings that can be extracted from microchips (hardware) and other devices that come in contact with the DNA of living organisms. These SNP arrays do not represent a complete DNA string, which, e.g. for a human, would consist of about 3.2×10^9 base pairs of the human chromosome. A typical SNP arrays would represent a fragment of this string with a length of, perhaps, up to 500,000 base pairs. Each base pair of the human DNA may be in one of four states (A, C, T, or G). The goal is to correctly identify genetic sequences of different individuals to help classify chromosomal regions where genetic variants are shared. For crops and animals, the study of SNPs is important in fertilization and breeding. For human DNA, the extracted SNPs may define how people contract diseases and respond to certain treatments, drugs, vaccines, chemicals, pathogens and other agents.

SNPs may be great enablers in developing personalized medicine, but there is some controversy of how this information may be possibly abused. Some basic definitions are appropriate in this work:

Definition 1: Phenotype – Is a measure of a trait/skill of an individual.

Definition 2: Genotype – The information carried by the genes.

Definition 3: Homozygous – The chromosomes are identical in every state.

Definition 4: Heterozygous – There exists a SNP between two chromosomes.

It is noted in Definition 4, that a SNP is not a weighted difference, in the sense that no distinction has been made between the states, e.g. A and G as being further apart from A and C. Future work may weigh different pair combinations as having distances between SNPs that are predicated on which base pairs are involved. For example, in Figure (1) three DNA strings are shown which are constructed, for simplicity, from a hypothesized 8 base pair fragment of DNA. For simplicity, the notation will be used that A=1, C=2, T=3, and G = 4, for the cells (alleles) although they are categorical variables. It is seen that DNA₁ and DNA₂ differ from each other by only one base pair. However, DNA₁ and DNA₃ differ by four base pairs. In some similarity sense using a distance/difference metric then DNA₁ is closer to DNA₂ and DNA₁ is further apart from DNA₃. This paper will investigate how to characterize the distance/difference of the various SNPs to discern similarity and differences in partial DNA strings. The use of information-theoretic variables will be employed to study the use of a measure of distance of SNPs via mutual information as well as alternative means.

Since the investigation of the SNPs will clearly depend on the appropriate measure of distance/difference between candidate DNAs, the use of classical information theoretic variables will be employed. Figure (2) displays an information theory channel [1-4]. In Figure (2) – Basic elements of an Information Channel from Shannon [1]. Figure (3) is a Venn diagram of the key information-theoretic measures involving two random variables X and

In memory of Dr. Daniel R. Repperger. BIOCAMP'12 - The 2012 International Conference on Bioinformatics & Computational Biology, Las Vegas, USA.

¹- Contact author. 711 HPW, AFRL, WPAFB, Ohio 45433-7022.

Email: Faruil.Mohd-Zaid@wpafb.af.mil

²- Dept. of Math., Ohio University, Athens, Ohio 45701. Email:

shenx@ohio.edu

³- 711 HPW, AFRL, WPAFB, Ohio 45433-7022.

Email: Katheryn.Farris@gmail.com

Y (Cover and Thomas [2], Sheridan and Ferrell [3], Repperger, et al. [4]).

DNA ₁	1	2	3	4	3	2	1	2
DNA ₂	1	2	1	4	3	2	1	2
DNA ₃	4	2	2	4	3	2	2	1

$$\text{distance}(\text{DNA}_1\text{-DNA}_3) > \text{distance}(\text{DNA}_1\text{-DNA}_2)$$

Figure (1) – Three DNA strings with Different Relative Distances.

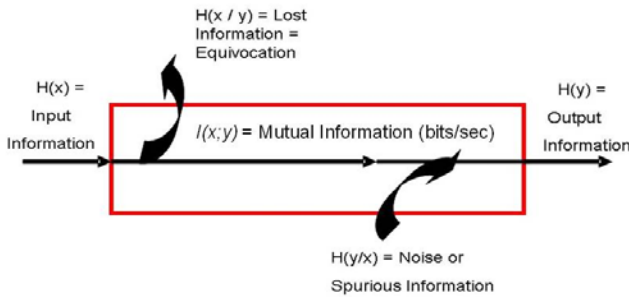
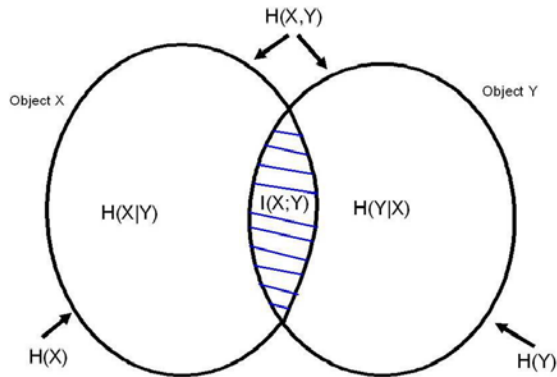


Fig. (2) – Basic elements of an Information Channel from Shannon [1]



$$H(x,y) = I(x;y) + D_R(x;y) = I(x;y) + H(x|y) + H(y|x)$$

The Information Variables in a Venn Diagram

Figure (3) – A Venn Diagram of the Key Variables

In Figure (3) the five information-theoretic quantities that describe the types of uncertainties (entropies) between the input and output elements of the information channel in Figure (2) are portrayed. Three of these five variables can be shown to be independent.

From Figures (2,3), the five basic entities of an information channel can be expressed as follows:

- $H(x)$ = The input uncertainty to the channel (1)
- $H(y)$ = The output uncertainty of the channel. (2)
- $H(x/y)$ = Equivocation lost to the environment. (3)
- $H(y/x)$ = Spurious uncertainty from the environment (4)
- $I(x;y)$ = Mutual information transmitted (5)

More specifically, equations (1-5) can be better described by letting $p(\cdot)$ represent the probability of an event. For an information channel with input symbol set in Figure (2), $x \in X$, of size n , and received symbols $y \in Y$ at the output set of size q (q may not equal n), the following entropy ($H(\cdot)$) relationships can be defined:

$$H(x) = \sum_{i=1}^n p(x_i) \log_2(1/p(x_i)) \quad (6)$$

$$H(y) = \sum_{j=1}^q p(y_j) \log_2(1/p(y_j)) \quad (7)$$

$$H(x,y) = \sum_{i,j}^{n,q} p(x_i, y_j) \log_2(1/p(x_i, y_j)) \quad (8)$$

$$H(x/y) = \sum_{i,j}^{n,q} p(x_i, y_j) \log_2(1/p(x_i | y_j)) \quad (9)$$

$$\text{and } H(y/x) = \sum_{i,j}^{n,q} p(x_i, y_j) \log_2(1/p(y_j | x_i)) \quad (10)$$

The important relationships that pertain to the modeling of the SNPs are dependent on the key variables (1-5). From Figures (2,3) and the basic definitions (6-10), the following relationship can be shown to be true (Cover & Thomas [2]):

$$I(x;y) = H(x) + H(y) - H(x,y) \quad (11)$$

where the mutual information $I(x;y)$ also satisfies:

$$I(x;y) \geq 0 \quad (12)$$

Finally, another important variable that will be used in the sequel is the relative information distance $D_R(x;y)$:

$$D_R(x;y) = H(x/y) + H(y/x) = H(x) + H(y) - 2I(x;y) \quad (13)$$

where $D_R(x;y)$ also has a positivity property, as in equation (12):

$$D_R(x;y) \geq 0 \quad (14)$$

There are advantages the variable D_R provides over $I(x;y)$ which are known in the literature (Cover & Thomas, [2], [5], and [6]) and restated here:

Property 1: $D_R(x;y)$ is a metric; however, $I(x;y)$ is only a measure. Please see the appendix and a counter example where $I(x;y)$ fails as a metric by not satisfying the triangular inequality.

A second property can be stated as follows:

Property 2: The relative information distance metric $D_R(x;y)$ is the complement of $I(x;y)$, i.e.

$$D_R(x;y) = \bar{I}(x;y) \text{ or } I(x;y) = \bar{D}_R(x;y) \quad (15)$$

Appendix A demonstrates this second property.

2. Methods and Technical Solutions

Contingency tables (Sheridan and Ferrell [3], Repperger, et al., [4], and Kullback [7]) will be used to formulate the SNP similarity and difference problem to utilize information-

theoretic quantities in examining the distance/difference between DNA strings.

Using the DNA strings in Figure (1), Contingency Table 1 is constructed which compares DNA₁ versus DNA₂ in terms of similarity and differences. Contingency Table 2 then compares DNA₁ versus DNA₃, and finally Contingency Table 3 compares DNA₂ versus DNA₃.

Contingency Table 1 – DNA₁ versus DNA₂

		DNA ₂ →			
		A=1	C=2	T=3	G=4
DNA ₁ ↓	A = 1	2	-	-	-
	C = 2	-	3	-	-
	T = 3	1	-	1	-
	G = 4	-	-	-	1
		3/8	3/8	1/8	1/8

Contingency Table 2 – DNA₁ versus DNA₃

		DNA ₃ →			
		A=1	C=2	T=3	G=4
DNA ₁ ↓	A = 1	-	1	-	1
	C = 2	1	2	-	-
	T = 3	-	1	1	-
	G = 4	-	-	-	1
		1/8	4/8	1/8	2/8

Contingency Table 3 – DNA₂ versus DNA₃

		DNA ₃ →			
		A=1	C=2	T=3	G=4
DNA ₂ ↓	A = 1	-	2	-	1
	C = 2	1	2	-	-
	T = 3	-	-	1	-
	G = 4	-	-	-	1
		3/8	3/8	1/8	1/8

Next, a normalized matrix is calculated based on the total number of responses in each table. The normalized matrices are summarized below for Contingency Tables 1-3.

Table 1 – Normalized

		DNA ₂ →				
		A=1	C=2	T=3	G=4	H(x) ↓
DNA ₁ ↓	A = 1	2/8	0	0	0	2/8
	C = 2	0	3/8	0	0	3/8
	T = 3	1/8	0	1/8	0	2/8
	G = 4	0	0	0	1/8	1/8
		3/8	3/8	1/8	1/8	H(y) →

Table 2 – Normalized

		DNA ₃ →				
		A=1	C=2	T=3	G=4	H(x) ↓
DNA ₁ ↓	A = 1	0	1/8	0	1/8	2/8
	C = 2	1/8	2/8	0	0	3/8
	T = 3	0	1/8	1/8	0	2/8
	G = 4	0	0	0	1/8	1/8
		1/8	4/8	1/8	2/8	H(y) →

Table 3 – Normalized

		DNA ₃ →				
		A=1	C=2	T=3	G=4	H(x) ↓
DNA ₂ ↓	A = 1	0	2/8	0	1/8	3/8
	C = 2	1/8	2/8	0	0	3/8
	T = 3	0	0	1/8	0	1/8
	G = 4	0	0	0	1/8	1/8
		1/8	4/8	1/8	2/8	H(y) →

To calculate the requisite entropies, the following procedures are then employed:

Step 1: Calculate $H(x)$ across the rows and then summing down the column on the right side of the normalized matrix (cf. Table 1-Normalized).

Step 2: Calculate $H(y)$ down the columns and then summing across the row on the bottom of the normalized matrix (cf. Table 1-Normalized).

Step 3: Calculate $H(x,y)$ for all cells in the normalized matrix. Then

$$I(x;y) = H(x) + H(y) - H(x,y) \tag{16}$$

and

$$D_R(x;y) = H(x/y) + H(y/x) = H(x) + H(y) - 2I(x;y). \tag{17}$$

The calculations proceed as follows for Table 1, for example:

$$H(x) = -2 * (2/8) \log_2(2/8) - (3/8) \log_2(3/8) - (1/8) \log_2(1/8) = 1.9056 \text{ bits} \tag{18}$$

$$H(y) = -2* (3/8) \log_2(3/8) - 2* (1/8) \log_2(1/8) = 1.8113 \text{ bits} \tag{19}$$

$$H(x,y) = - (3/8) \log_2(3/8) - (2/8) \log_2(2/8) - (3) * (1/8) \log_2(1/8) = 2.1556 \text{ bits} \tag{20}$$

$$I(x;y) = H(x) + H(y) - H(x,y) = 1.5613 \text{ bits} \tag{21}$$

$$D_R = H(x) + H(y) - 2I(x;y) = 0.594 \text{ bits} \tag{22}$$

Finally, it is noted in Figure (1) that in a distance/difference sense, it is expected that:

$$\text{dist. (DNA}_1\text{- DNA}_3\text{)} > \text{dist. (DNA}_1\text{- DNA}_2\text{)} \quad (23)$$

Table 4 summarizes these results. It is seen that dissimilarities between DNAs are generally associated with large D_R values, small $I(x;y)$ values, and larger Hamming distance values. The Hamming distance (independent of position) is defined as the percent of cells that differ in a dyadic comparison and is the gold standard in discerning differences between computer words.

Table 4 – Distances between the SNPs in Figure (1)

Distance Variable	DNA ₁ -DNA ₂	DNA ₂ -DNA ₃	DNA ₁ -DNA ₃
Hamming	0.125	0.50	0.50
$I(x;y)$	1.5613	1.3113	0.9056
$D_R(x;y)$	0.5944	1.1887	1.8444

Typically a reduction in the value of D_R would be accompanied by an increase in $I(x;y)$. For the two random variable case (as shown in the appendix), it can be demonstrated that D_R and $I(x;y)$ are complements of each other (i.e. $\bar{D}_R = I(x;y)$ and $\bar{I}(x;y) = D_R$). The results of Table 4 are consistent. As the Hamming distance increases (column 2 in row 2) when compared to either column 3 or column 4, then $I(x;y)$ decreases and D_R increases, as expected. Two counter examples are now presented.

3. Empirical Evaluation

The first counter example is illustrated with Venn diagrams in Appendix A which shows that $I(x;y)$ is not consistent in discerning distance/differences between DNAs because it does not satisfy the triangular inequality.

Case 1 Counter Example with Venn Diagrams

Please see appendix A for an example using Venn diagrams and set theory. This presentation is based on geometric arguments. It is shown that $I(x;y)$ violates the triangular inequality thus does not satisfy the property of being a norm. The second counter example deals with SNPs.

Case 2 Counter Example with SNPs

To generalize the counter example, analogous to the Venn diagrams in appendix A to DNA identification, the following three DNA strings are constructed:

DNA ₁	1	2	3	4	3	2	1	3	2	4
DNA ₂	1	3	2	1	2	4	1	4	3	1
DNA ₃	1	2	2	4	3	2	3	4	2	4

Figure (4) – Counter Example in terms of SNPs

To show that difficulties may occur by using $I(x;y)$ as well as D_R to characterize distance/difference between DNAs, the three normalized matrices resulting from the contingency tables are displayed for the counter example DNAs in Figure (4). Using similar notation, as before, Table 5 portrays

Table 5 – Normalized

DNA₂ →

DNA ₁ ↓	2/10	0	0	0	H(x) ↓
	0	0	2/10	1/10	2/10
	0	2/10	0	1/10	3/10
	2/10	0	0	0	3/10
	4/10	2/10	2/10	2/10	2/10
H(y) →					

Table 6 – Normalized

DNA₃ →

DNA ₁ ↓	1/10	0	1/10	2/10	H(x) ↓
	0	1/10	1/10	0	4/10
	0	2/10	0	0	2/10
	0	1/10	0	1/10	2/10
	1/10	4/10	2/10	3/10	
H(y) →					

Table 7 – Normalized

DNA₃ →

DNA ₂ ↓	1/10	0	1/10	0	H(x) ↓
	0	3/10	0	0	2/10
	0	1/10	1/10	1/10	3/10
	0	0	0	2/10	3/10
	1/10	4/10	2/10	3/10	
H(y) →					

DNA₁ versus DNA₂, Table 6 shows DNA₂ versus DNA₃, and Table 7 illustrates DNA₁ versus DNA₃. The calculations from Tables 5-7 are summarized in Table 8. Also enclosed in this table is the calculation from the Hamming distance, which has been a traditional measure of distance between computer words [8,9].

Table 8 – Results of the Calculation of the SNPs in Figure 4

Information-theoretic Variable (bits)	DNA ₁ vs. DNA ₂ Table 5	DNA ₁ vs. DNA ₃ Table 7	DNA ₂ vs. DNA ₃ Table 6
$H(x)$	1.971	1.9710	1.9219
$H(y)$	1.9219	1.8464	1.8464
$H(x,y)$	2.5219	2.6464	2.9219
$I(x;y)$	1.371	1.1710	.8464
$D_R(x;y)$	1.151	1.4755	2.0755
Hamming Distance	.80	.30	.70

Note the following conclusions are reached from Table 8:

- (1) The Hamming Distance (gold standard) is a well accepted metric and will be used as a baseline (ground truth) to evaluate the information-theoretic variables investigated herein.
- (2) In Table 8, bottom row, comparing column 3 to column 4 (results of Table 7 versus Table 6), as the Hamming distance increased (from .3 in column 3 to 0.7 in column 4) then $I(x;y)$, decreased from 1.171 to 0.8464 which was expected. Also, D_R increased, accordingly.
- (3). However, when comparing column 3 to column 2 (results of Table 5 versus Table 7), when the Hamming distance increased from 0.3 to 0.8, the $I(x;y)$ should have decreased, but it increased from 1.171 to 1.371, which is inconsistent. This same inconsistency occurred with the variable D_R . Thus the information variables differ in their determination of distance/difference between DNAs and are not consistent with the ordering provided by the Hamming metric.

Next a discussion is presented on the transitive property of key variables and related to the measures and metrics discussed so far involving decision making, in general.

4. Transitive Property of Measures/Metrics

From Logic: Definition: A dyadic relation R is said to be transitive in a set S if whenever $a R b$ and $b R c$ imply $a R c$. For example, the relation “is greater than or equal” satisfies the transitive property for scalar numbers.

The structure of transitivity is the mainspring of deductive reasoning. An argument is said to be deductive when the truth of the conclusion is purported to follow necessarily. Deductive reasoning is one of the two basic forms of valid reasoning. While inductive reasoning argues from the particular to the general, deductive reasoning argues from the general to a specific instance. The basic idea is that if something is true of a class of things in general, this truth applies to all legitimate members of that class. The key, then, is to be able to properly identify members of the class. Miss-classifying (or miss-categorizing) will result in invalid conclusions and affecting decision making, adversely.

One of the most common and useful forms of deductive reasoning is the syllogism. The syllogism is a specific form of argument that has three easy steps, for example

1. Every X has the characteristic Y. This thing is X.
2. Therefore, this thing has characteristic Y.

Also, the transitive property makes elimination possible; if $a R b$ and $b R c$, we can eliminate b and assert $a R c$.

Finally, as applied to decision making, if a decision is made that the distance/difference between two DNAs is greater for one pair as compared to another pair, then the data may be mined out if the goal was to find highly correlated DNA pairs. Using $I(x;y)$ may lead to an error by mining out more correlated pairs of DNAs. If a distance metric such as the Hamming distance (as discussed in this paper) were employed, then the conclusion would not suffer from that error. As mentioned previously, the weakness of the Hamming distance is that it is a relative measure, not an absolute measure (the position of where the SNPs are lost).

5. Significance and Impact

Decision making based on closeness as measured by distance/difference between candidate DNAs is critically important if DNA analysis is used to make accurate determinations in data. Problems of consistency are seen when selecting D_R and mutual information ($I(x;y)$), being widely used in the literature. The property that D_R is a metric and $I(x;y)$ is only a measure, demands that proper decision making should be predicated on at least a good measurement tool (D_R in lieu of $I(x;y)$). Apparently D_R satisfying the triangular inequality still does not guarantee consistency in the decision making, as shown earlier, on the decision regarding simple binary choice of a string of DNA being more or less similar.

6. Future Work and New Research Directives

As mentioned previously, the classification of the similarity and differences between sample DNAs and the causality mapping between the SNP's scripts with the phenotype traits is a wide open area of research. A discussion on some of the fundamental problems in this area and possible solutions are now conducted. First some basic history is presented.

The human genome project has its early roots in the 1940's when the Department of Energy made an effort to develop new energy resources and still understand the potential health and environmental risks associated with these resources. In 2001, two publications [10, 11] described the initial sequencing and analysis of the human genome. By 2003, the sequencing was completed, two years earlier than anticipated. The generalizations are now far reaching. The DNA in each human cell is packaged into 46 chromosomes

arranged into 23 pairs. Each chromosome contains many genes (approximately 25,000 for the human genome), which are the basic physical and functional units of heredity. Genes are specific sequences of bases that encode instructions on how to make proteins. It is from the action of the proteins that the phenotype traits emerge.

Previously discussed, SNPs are variations in the DNA that may be extracted as SNP arrays by microchips or by other processes. The question arises if the sample is representative of that portion of the DNA string being relevant to the phenotype trait of interest? This is better understood from some other properties that reside within the human DNA sequencing: (1) Only about 2% of the genome actually encodes the instruction for the synthesis of proteins, (2) The human genome sequence is almost (99.9%) exactly the same in all people, and (3) particular gene sequences in animals have been associated with numerous diseases and disorders, including breast cancer, muscle disease, deafness, and blindness. For example, in a mouse, [12] cancer susceptibility can be related to new gene-mapping resources and specific genes can be indentified that concur with mice that contract the disease.

The tumor classification problem is of high interest in the field of bioinformatics [13]. The design of the candidate chips to extract the fragment DNA (SNPs) is a problem of considerable concern. Such systems are far from perfect and the environment can exert an undue influence in the process. The environment can mutate certain genes, thus producing a gene with a higher vulnerability to disease. For example, exposure to smoking is known to mutate a gene and thus produce cells that may start developing cancer. Thus if only one difference occurs in a base pair, this is still very important to capture since it may greatly influence a phenotype trait.

The future problems that may be studied in this area can be investigated in an algorithmic way on how certain SNPs signatures may result in a phenotype trait. The trait could be a “good attribute” like resistance to disease, increased strength, size, and other qualities. Alternatively, the modified SNP signature may also be a “bad attribute” including susceptibility to viruses, diseases, etc. For simplicity of discussion, the presumption will be made that the phenotype trait will exist in only two states, e.g.

Phenotype trait 1: No disease outcome (being resistant to a specific disease).

Phenotype trait 2: Being vulnerable to a specific disease.

Assume four DNA samples are taken from four individuals that equally fell into one of the two states above. Table 9 would classify the four DNA samples:

Table 9 – Four DNA samples obtained

Individual Number	No Disease State	Disease State
1	DNA ₁	
2		DNA ₂
3	DNA ₃	
4		DNA ₄

Recall that only 2% of the DNA is related to producing proteins that will affect the phenotype outcomes, then to develop the similarities and differences between the sample DNAs in Table 9, the following six steps should be conducted:

Step 1: Remove all common alleles (this includes the 98% of the DNA not associated with the protein production). Let the symbol Ω represent those common cells (alleles) that **are not** related to differences between the DNAs. In a set theory description, it represents the intersection of all the sample DNAs, i.e.

$$\Omega = \text{DNA}_1 \cap \text{DNA}_2 \cap \text{DNA}_3 \cap \text{DNA}_4 \quad (24)$$

Then let the underlined notation characterize that part of each DNA_i different from the common intersection of all SNPs, i.e.

$$\underline{\text{SNP}}_1 = \text{DNA}_1 - \Omega \quad (25)$$

$$\underline{\text{SNP}}_2 = \text{DNA}_2 - \Omega \quad (26)$$

$$\underline{\text{SNP}}_3 = \text{DNA}_3 - \Omega \quad (27)$$

$$\underline{\text{SNP}}_4 = \text{DNA}_4 - \Omega \quad (28)$$

Next, from Table 9, take those common SNP values for the diseased State:

Step 2: $\text{SNP}_D = \underline{\text{SNP}}_2 \cap \underline{\text{SNP}}_4 \quad (29)$

To characterize those common SNP values for the non diseased state:

Step 3: $\text{SNP}_{ND} = \underline{\text{SNP}}_1 \cap \underline{\text{SNP}}_3 \quad (30)$

Step 4: Now check if the disease and non diseased SNP portions are mutually exclusive: Is it true that:

$$\text{SNP}_D \cap \text{SNP}_{ND} = \phi \quad (31)$$

where ϕ is an empty set? If (31) is not true, then recalculate steps 1-3 until the result in equation (31) is satisfied.

Step 5: Now repeat steps 1-4 for more than two individuals.

Step 6: With a sufficient data base built up on the two classes {SND_D} and {SND_{ND}} predictions can then be made for individuals **outside the data used to develop the two classes**. This will test the efficacy of this method.

References

[1] Shannon, C. E. (1949). Communications in the presence of noise, *Proceed. of the IRE*, **37**, 10-22.
 [2] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley & Sons, Inc.
 [3] Sheridan, T. B. and Ferrell, W. R. (1981). *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*, The MIT press, Cambridge, Mass.
 [4] Repperger, D. W., Roberts, R. G., Lyons, J. B., and Ewing, R. L., “Optimization of an Air Logistics Systems via a Genetic Algorithm Model,” To appear in *International Journal of Logistics Research*, 2011.
 [5] J. P. Crutchfield, “Information and Its Metric,” in *Nonlinear Structures in Physical Systems-Pattern Formatio Chaos and Waves*, L. Lam and H. C. Morris, Eds., L. Lam and H. C. Morris, Eds., Springer-Verlag, NY (1990), 119-130.

[6] C. H. Bennett, P. Gacs, M. Li, P. M. B. Vitanyi, and W. H. Zurek, "Information Distance," *IEEE Transactions on Information Theory*, **44**(4), July, 1998, pp. 1407- 1423.

[7] S. Kullback, *Information Theory and Statistics*, New York, Dover, 1968.

[8] W. Zhao, E. Serpedin, and E. R. Dougherty, "Inferring Connectivity of Genetic Regularity Networks Using Information-Theoretic Criteria," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **5**(2), April-June, 2008, pp. 262-274.

[9] A. G. O'yachkov and D. C. Torney, "On Similarity Codes," *IEEE Transactions on Information Theory*, **vol. 46**, no. 4, July, 2000, pp. 1558-1564.

[10] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, and J. Baldwin, "Initial Sequencing and Analysis of the Human Genome," *Nature*, **409**, 2001, pp. 860-921.

[11] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, and G. G. Sutton, "The Sequence of The Human Genome," *Science*, **291**, 2001, pp. 1304-1351.

[12] P. Demant, "Cancer Susceptibility in the Mouse: Genetics, Biology, and Implications for Human Cancer," *Nature Reviews/Genetics*, **vol. 4**, September, 2003, pp. 721-735.

[13] R. Desper, J. Khan, A. A. Schaffer, "Tumor Classification Using Phylogenetic Methods on Expression Data," *Journal of Theoretical Biology*, **228**, 2004, pp. 477-496.

[14] M. Li, X. Chen, L. Xin, M. Bin, and P. M. B. Vitanyi, "The Similarity Metric," *IEEE Transactions on Information Theory*, **50**,(12), Dec 2004, pp. 3250-3264.

Appendix A – Counter Example 1 – With Venn Diagrams

Since geometric proofs using Venn diagrams are not technically permissible (Eves, [26]), we show as a test of relationships properties 1 and 2. In this appendix, it will be stated (Cover and Thomas, [2]) that D_R is a metric and satisfies the follow following four relationships for a metric $\rho(x,y)$:

- (M-1) $\rho(x,y) > 0$ if $x \neq y$. (positivity) (A.1)
- (M-2) $\rho(x,y) = \rho(y,x)$ (similarity) (A.2)
- (M-3) $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$ (triangular inequality) (A.3)
- (M-4) $\rho(x,y) = 0$ if and only if $x = y$ (A.4)

However, it is shown by a testing example below that $I(x;y)$ violates equation (A.3), i.e.

$$I(x;z) > I(x;y) + I(y;z) \tag{A.5}$$

for three random variables X, Y, and Z.

Part A – A Constructed Example to Show That Equation (A.3) is Violated:

Figure (A-1a) is presented to define areas $A_1, A_2,$ and A_3 consistent with Figure (3):

$$H(x/y) = A_1 \tag{A.6}$$

$$H(y/x) = A_3 \tag{A.7}$$

$$I(x;y) = A_2 \tag{A.8}$$

Figure (A-1b) now generalizes this concept to three random variables X, Y, Z. In terms of the seven areas (A_1 - A_7) displayed, the following relationships become generalizations of Figure (A-1a) into Figure (A-1b):

$$I(x;y) = A_2 \tag{A.9}$$

$$H(x/y) = A_1. \tag{A.10}$$

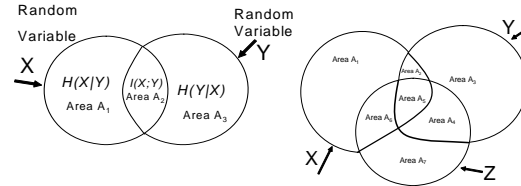


Figure (A-1a) Two Random Variables X and Y (left)

Figure (A-1b) – Three Random Variables X, Y, and Z

$$H(y/x) = A_3, \tag{A.11}$$

$$H(x/y) = A_1 + A_6, \quad I(x;y) = A_2 + A_5 \tag{A.12}$$

$$H(y/x) = A_3 + A_4, \quad I(y;x) = A_5 + A_2 \tag{A.13}$$

$$H(z/x) = A_4 + A_7, \quad I(z;x) = A_5 + A_6 \tag{A.14}$$

$$H(x/z) = A_1 + A_2, \quad I(x;z) = A_6 + A_5 \tag{A.15}$$

$$H(y/z) = A_2 + A_3, \quad I(y;z) = A_5 + A_4 \tag{A.16}$$

$$H(z/y) = A_6 + A_7, \quad I(z;y) = A_4 + A_5 \tag{A.17}$$

the left and the random variable Y to the right until:

$$A_6 > A_2 + A_4 + A_5. \tag{A.18}$$

But: $A_6 = I(x;z) - A_5 \tag{A.19}$

$$A_2 = I(x;y) - A_5 \tag{A.20}$$

$$A_4 = I(y;z) - A_5 \tag{A.21}$$

Hence from (A.18):

$$I(x;z) - A_5 > I(x;y) - A_5 + I(y;z) - A_5 + A_5 \tag{A.22}$$

by construction, and

$$I(x;z) > I(x;y) + I(y;z) \tag{A.23}$$

Thus it is demonstrated that equation (A.5) is satisfied and condition (A.3) is violated.

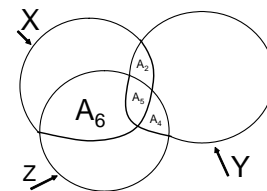


Figure A-2 – Counter example to show $(A_6 > A_2 + A_4 + A_5)$

Property 2: D_R and $I(x;y)$ are complements of each other.

It is intriguing that D_R satisfies the property of a metric but property 2 states that its complement $I(x;y)$ does not. With reference to Figure (A-1a) we wish to demonstrate that D_R and $I(x;y)$ are complements. By definition:

$$I(x;y) = H(x) + H(y) - H(x,y) \tag{A.24}$$

$$D_R(x;y) = H(x/y) + H(y/x) \tag{A.25}$$

Let S represent the entire space in Figures (3) and (A-1a).

Then let e be an element of S and $S = A_1 \cup A_2 \cup A_3$ where \cup indicates the union of sets. Note A_1, A_2 and A_3 are disjoint sets in Figure (A-1a). From (A.11-A.12) and Figures (5) and (A-1a) it follows that all the elements of $I(x;y)$ are in A_2 and all the elements of $D_R(x;y)$ are in $A_1 \cup A_3$. For notational simplicity denote the complement of a set A as A' , then (Eves, [26]) two cases now exist: (1) if $e \in A_1 \cup A_3$, then $e \notin A_2$, thus

$$(A_1 \cup A_3)' = A_2 \tag{A.26}$$

or (2) if $e \in A_2$, then $e \notin A_1 \cup A_3$, thus

$$(A_2)' = (A_1 \cup A_3) \tag{A.27}$$

It then follows that if e is an element of $(A_1 \cup A_3)$ then e is an element of $(A_2)'$, and if e is an element of (A_2) , and if e is an element of $(A_1 \cup A_3)'$, whence (A_2) and $(A_1 \cup A_3)$ are complements.

Computational Criteria for the Disablement of Human GAPDH Pseudogenes

C.S. Theisen¹, K.A. Seidler¹, and N.W. Seidler¹

¹Department of Biochemistry, Kansas City University of Medicine and Biosciences, Kansas City, Missouri, USA

Abstract – *Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), the glycolytic enzyme, exists as an asymmetric homotetramer and is apparently a product of a single somatic gene. This protein is considered a moonlighting protein meaning that it exhibits functions typically ascribed to other proteins. The functions include cellular processes that involve gene expression and, intriguingly, those that are associated with membrane-binding. There are more than 60 human pseudogenes for GAPDH. These pseudogenes, by definition, are considered non-functional, although multiple transcripts for GAPDH have been reported, suggesting that one or more of these pseudogenes are active. To assess whether there is selective pressure to preserve these sequences, we developed criteria to identify whether or not disablement of these pseudogenes occurred. The criteria involve analysis of the structural features of the $\alpha 1$ -helix of real and pseudo GAPDH proteins. This region is required for membrane-association, a conserved property that may be lost in pseudogenes.*

Keywords: glyceraldehyde 3-phosphate dehydrogenase, pseudogenes, retrotransposition, membrane-associated proteins, Parkinson's disease.

1 Introduction

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) plays a vital role in glycolysis, an energy-generating pathway in all human cells. Net energy is not generated until stage two, where glyceraldehyde 3-phosphate is converted to pyruvate. The first reaction of this stage of glycolysis is catalyzed by GAPDH. The substrates are D-glyceraldehyde 3-phosphate, inorganic phosphate and NAD^+ , and the products are 1,3-bisphospho D-glycerate and NADH. The reaction is an oxidative phosphorylation and involves a covalent intermediate between the substrate D-glyceraldehyde 3-phosphate and the active site cysteine residue [6]. In addition to this well-known function of GAPDH, this protein participates in many other non-glycolytic functions including membrane fusion activity [7]. Additionally, it has been shown

to be nitrosylated, translocated to the nucleus and participate in apoptotic signaling that has been linked to Parkinson's disease. GAPDH, which is an abundant cellular protein, contains only one functional gene, which is on chromosome 12. Curiously, however, the human genome contains over 60 pseudogenes dispersed throughout the genome [8]. Other mammals have also been shown to carry an unusually high number of GAPDH pseudogenes. Pseudogenes refer to non-functional genes that are related in some way to a functional gene in the genome. The pseudogenes may arise through duplication or retrotransposition [3]. Attendant with duplication are mutations that end up disabling the gene making it non-functional. These genes represent complete copies, but are not transcribed due to a disabling mutation and are called non-processed pseudogenes. Retrotransposition, on the other hand, involves the reverse transcription of an mRNA transcript and then the re-integration of the subsequent cDNA back into the genome. This type is called processed pseudogenes. In the absence of any selective pressure, one would expect that the pseudogenes would accumulate disabling mutations and that the pseudogenes would exhibit a decay that is at a faster rate than the parent gene [9]. Liu and coworkers [8] indicated that GAPDH pseudogenes are preferentially spared the disabling mutations that are typically seen with other pseudogenes. We were curious about the disablement of the GAPDH pseudogenes and in particular the interpretation of disablement. We were interested in examining the sequence associated with the first helix (designated as, $\alpha 1$ -helix) that is found in the NAD^+ -binding domain, which is located from residues 12 to 23 (i.e. human numbering includes initial methionine). In addition to the entire NAD^+ -binding domain exhibiting conserved homology through evolution, the N-terminal end of the NAD^+ -binding region is particularly conserved [10, 11]. This $\alpha 1$ -helix is thought to be imbedded in target membranes horizontal to the plane of the plasma membrane due to the amphipathic nature of this helix [4].

Interestingly, diverse pathogenic microorganisms have evolved to utilize GAPDH's multi-functionality in unique ways. *Listeria monocytogenes*, for example, is an intracellular

parasite that disrupts normal host cell phagocytosis by mono-ADP-ribosylation (i.e. inactivation) of a host Rab protein [4]. Others have shown that GAPDH interacts with various target membranes [12, 13, 14, 15].

We wanted to explore whether this region exhibits significant disablement, presumably due to mutations following retrotransposition, or, alternately, occurring prior to and as a selection pressure for promoting retrotransposition of the GAPDH transcript. The criteria for disablement would include a significant loss of functional conformational status. This is assessed by examining the structural properties of this initial helical region of the NAD⁺-binding domain.

2 Materials and methods

Materials. We utilized public accessible databases. The literature contains numerous articles on the multi-functionality of GAPDH. In addition to the GAPDH pseudogene literature, we consulted articles pertaining to GAPDH's ubiquitous properties of reversibly binding to membranes. Furthermore, one website is solely focused on pseudogenes (<http://pseudogenes.org/glycolysis>) and maintained by the Gerstein Lab at Yale University. Another pseudogene database is NCBI (<http://ncbi.nlm.nih.gov/gene>).

Computation of the Central Longitudinal Plane. The helix is assumed to exhibit a rigid cylinder. We set out to determine the central longitudinal plane of this cylinder that delineates the amphipathic parts with one half of the cylinder representing hydrophobic tendency and the other half hydrophilic. The standard helical wheel, which is a transverse image of the cylinder, was used to compute the central longitudinal plane. The central longitudinal plane was assigned to zero degree angle, dividing the alpha-helical wheel into equal halves, based on previous assignment [4] and placing the Gly-11 (i.e. *Listeria monocytogenes* GAPDH) at 350°. Each of the 12 residues that make up the helix was assigned their position (i.e. degree angle) around the wheel accordingly in a clockwise fashion that circumscribes 360°.

The most preferred position of a hydrophilic (or, hydrophobic) residue was assumed to be 90° relative to the central longitudinal plane, creating a theoretical amphipathic configuration for each residue (Figure 1). A counterclockwise displacement was designated as positive and a clockwise displacement as negative.

This assumption allowed us to determine an assignment for the central longitudinal plane based on the summation of effects by all of the residues around the helical wheel. We computed the degree of displacement of the central longitudinal plane

that would allow for optimal positioning of the residue.

The contribution of this displacement by a single residue in the helix was determined mathematically by first assigning a polarity index using the Carugo's hydrophobicity scale [5] and then determining the fractional coefficient, or fC (i.e. the *relative* polarity index of the residue) for each residue regardless of their position on the wheel as indicated by equation (1).

$$fC = PI_n / \langle PI \rangle \quad (1)$$

Where, PI_n is the polar index of the amino acid residue, $n = 1, 2, 3, \dots$, and the value $\langle PI \rangle$ represents the total sum of the PI values for all of the amino acid residues.

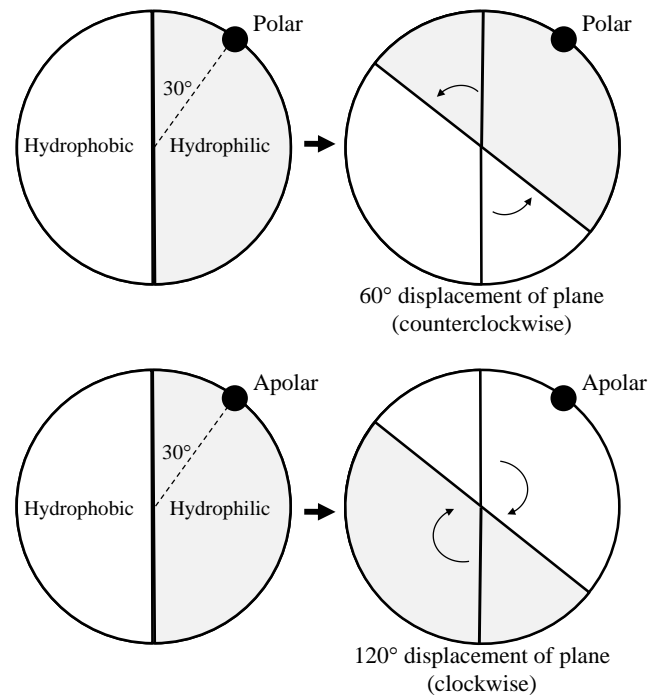


Figure 1: Procedure for determining degree of displacement of virtual plane bisecting the helix. An initial arbitrary plane is established delineating the hydrophobic and hydrophilic halves of the alpha-helical wheel. Considering a polar residue that is 30° off center (upper left), it would prefer displacement of the plane by 60° counterclockwise (upper right). Considering now an apolar residue that is 30° off center (lower left), it would prefer displacement of the plane by 120° clockwise (lower right).

The fC represents the magnitude of each amino acid residue's contribution to the overall hydrophobicity. Next, the effect of each of the residues on the degree of rotation was determined

by calculating the fractional effect, which is dependent on the residue's location and its polar/apolar property (Table 1).

For example, if an arginine residue was 30° from the default central plane, it would prefer a rotation of the helix 60° (and therefore the central plane is shifted 60°) to reach the assumed 90° optimal orientation. Each residue has a fractional effect on the determination of the ultimate angle of the central longitudinal plane (\angle CLP). The contribution of each residue on the position of the central longitudinal plane was determined by multiplying the fractional coefficient (equation 1) by the relative degree of displacement (Table 1).

$$fE_n = fC_n \cdot \angle D_n \quad (2)$$

$$\angle \text{CLP} = \sum (fE_n)_i \quad (3)$$

Where fE_n is the fractional effect of each of the residues, $n = 1,2,3,\dots$, determined by the product of fC_n (i.e. fraction coefficient) and its corresponding angular displacement, $\angle D_n$, and where \angle CLP is the sum of the fractional effect of each residue, $i = 1-12$. The final helical rotation, or \angle CLP, created a bisecting plane for each of the $\alpha 1$ -helicies in the various GAPDH sequences that were studied.

Table 1: Description of the sequential order of amino acid residues, their respective start angular positions and their designated displacements required to meet the optimal amphipathic bisecting plane as described above.

RESIDUE	START ANGLE	ANGULAR DISPLACEMENT (\angle D)	
		POLAR RESIDUE	APOLAR RESIDUE
1	350	+100	-80
2	90	0	+/-180
3	190	-100	+80
4	290	+160	-20
5	30	+60	+120
6	130	-40	+140
7	230	-140	+40
8	330	+120	-60
9	70	+20	-160
10	170	-80	+100
11	270	+/-180	0
12	10	+80	-100

Analysis of Deviation of the Helix from Horizontality. Once the central longitudinal plane was determined for a given helix, then the deviation from horizontality was determined

(Figure 2). We developed a parameter, designated Pw (for, **Positional Weight**), that is the product of the polar index (or, Carugo's hydrophobicity scale [5]) and the distance (or, d), in angstrom, perpendicular from the central longitudinal plane. These values (i.e. Pw) were then plotted as a function of the linear position from zero to 16.5 angstrom along the longitudinal axis of the helix. This plot generated a linear regression, which may represent a significant deviation from horizontality.

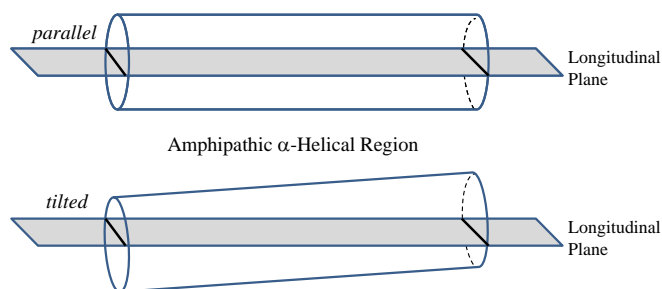


Figure 2: Procedure for determining degree of tilt.

The calculated central longitudinal plane bisected the helix into hydrophobic and hydrophilic halves. We calculated the vertical distance of each residue from the central plane using $\sin \theta = \text{distance}/\text{helical radius}$. We determined the angle θ based on the position of the central plane and the helical radius of 6 angstrom. We included the vertical distance of those residues that are positioned in its incompatible environment (i.e. polar residue in the apolar half) and was designated a negative integer. We subtracted this distance from 6 angstrom (i.e. the preferred position of the residue in its ideal surroundings). This net distance was multiplied by the PI_n (or, polar index).

For example, an arginine residue that is positioned in the apolar half gives a calculated vertical distance of 2 angstrom from the central longitudinal plane. In order for this residue to reach its preferred position, it would have to traverse these 2 angstrom as well as the 6 angstrom that represents the radius of the helix, suggesting that the distance is proportional to the force of repulsion. Therefore, this 2 angstrom distance was added to the 6 angstrom distance from the preferred position and the central plane for a total of eight angstrom. This net distance was multiplied by the residue's Polar Index value, giving a parameter that we designated as Positional Weight. A plot was made showing Positional Weight as a function of longitudinal distance along the helix.

The regression plot was made in SigmaPlot 11.0, which allowed for regression analysis, providing an r -squared value that was assessed statistically for deviation from horizontality. A significant correlation of Positional Weight and distance

along the longitudinal axis of the helix would suggest that the helix is significantly tilted and therefore was used as a criterion for disablement.

3 Results

We observed that the calculated central longitudinal plane, determined by our mathematical procedure for the *Listeria monocytogenes* GAPDH, was not that divergent from the designated bisecting plane as presented by Alvarez-Dominguez and coworkers [4]. It is presumed that those authors chose that plane by visual inspection. It appears that even by visual inspection the chosen bisecting plane is indisputable. The bisecting plane in their article exhibits the initial glycine residue (i.e. Gly-11 of *Listeria monocytogenes* GAPDH numbering includes the start methionine) positioned 10° below the hydrophobic-hydrophilic bisecting plane into the hydrophobic half of the amphipathic helix - the half that is presumably imbedded in the target membrane. We observed a calculated central longitudinal plane displaced 24.5° from this arbitrary start plane (Table 2).

Table 2: The calculated values for the central longitudinal plane for various gene products (or, in the case of pseudogenes, putative gene products)

GENE	CENTRAL LONGITUDINAL PLANE
Human GAPDH	37.2
<i>L. monocytogenes</i> GAPDH	24.5
GAPDHP44	unstable plane
GAPDHP23	unstable plane
GAPDHP71	44.1
GAPDHP62	38.6
GAPDHP34	86.5
GAPDHP2	34.4
GAPDHP19	25.3
GAPDHP58	65.3
GAPDHP41	unstable plane

Interestingly, the gene products for the human GAPDH and that of the pathogenic microorganism, *Listeria monocytogenes*, are not that different from one another in terms of the calculated central longitudinal planes. The required displacement for the wild-type human GAPDH from the arbitrary start plane is 37.2° , a mere 12.7° difference from that of *Listeria monocytogenes* GAPDH. The putative gene products of the pseudogenes varied from 25.3° to 86.5° . Four of the nine pseudogenes tested were within 12° of the wild-type human somatic GAPDH gene product (i.e. GAPDH-P71, -P62, -P2 and -P19). This minor difference would be - in the opinion of the present authors - not significant enough to consider them disabled (they still exhibit a defined central longitudinal plane). Two of the nine pseudogenes differed by over 24° in their required displacement. Even this greater requirement of displacement, these pseudogenes may not be disabled (again, for the same rationale that they exhibited a defined central longitudinal plane). However, there were three of the nine pseudogenes (i.e. GAPDH-P44, -P23 and -P41) that displayed a predicted unstable plane, as evidenced by at least one amino acid substitution that required a ± 180 displacement for its optimal location, likely creating a tendency of the predicted helix to wobble relative to the amphipathic properties and its half-submersion in the lipid bilayer. By this criterion, we designate that these pseudogenes (i.e. GAPDH-P44, -P23 and -P41) exhibit significant disablement.

The predicted center longitudinal plane for *Listeria monocytogenes* GAPDH is shown in Figure 3. One can see by visual inspection that the computed line gives a reasonable approximation of the longitudinal separation of the amphipathic helix.

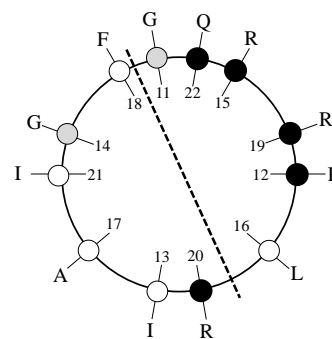


Figure 3: Helical wheel for the $\alpha 1$ -helix of *Listeria monocytogenes* GAPDH. The angular arrangement of amino acid residues 11 to 22 are given. The dotted line represents the calculated center longitudinal plane. Black-filled circles indicate strongly polar residues. Unfilled circles represent apolar residues. Gray-filled circles are slightly polar residues.

The human GAPDH exhibits a bisecting line (Figure 4) that is slightly more displaced than that seen with the *Listeria monocytogenes* GAPDH.

Despite the considerable sequence difference in pseudogene GAPDHP62, which is located on the q arm of chromosome 8, the calculated central longitudinal plane (Figure 5) was almost identical to that of the functional human GAPDH.

The assessment of the deviation of the helix from horizontality showed that the *Listeria monocytogenes* GAPDH was almost perfectly horizontal (Figure 6). Since the Polar Index values used in these calculations included their sign, the Positional Weight values were either positive or negative, based on hydrophobicity (i.e. positive equates to hydrophobic forces).

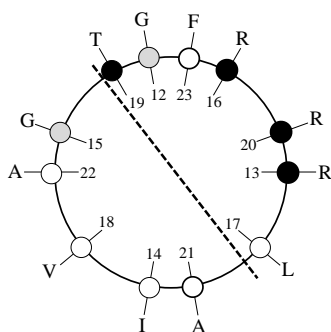


Figure 4: Helical wheel for the $\alpha 1$ -helix of human GAPDH. The angular arrangement of amino acid residues 12 to 23 are given. The dotted line represents the calculated center longitudinal plane. Black-filled circles, strongly polar residues; unfilled circles, apolar residues; gray-filled circles, slightly polar residues.

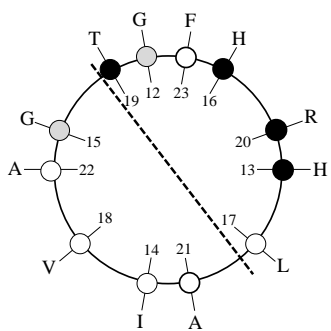


Figure 5: Helical wheel for the $\alpha 1$ -helix of human pseudogene GAPDHP62. The angular arrangement of amino acid residues 12 to 23 are given. The dotted line represents the calculated center longitudinal plane. Black-filled circles, strongly polar residues;

unfilled circles, apolar residues; gray-filled circles, slightly polar residues.

The relationship in the graph, shown in Figure 6, exhibits a slight angle of displacement from a horizontal orientation, but this deviation was not considered statistically significant.

The assessment of the human GAPDH exhibited an opposite pitch (Figure 7), but this deviation from the horizontal was also not statistically significant as evaluated by a Pearson r using 95% confidence limits.

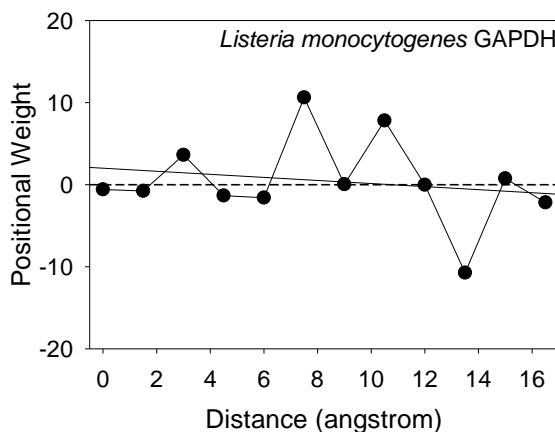


Figure 6: Assessment of horizontality of the $\alpha 1$ -helix of *Listeria monocytogenes* GAPDH. The Positional Weights of each of the 12 residues were plotted over the longitudinal distance of the helix. The dotted line represents a reference point of horizontality. The solid line is the calculated regression, representing the asymmetric contribution of the repulsive forces of the amino acid residues.

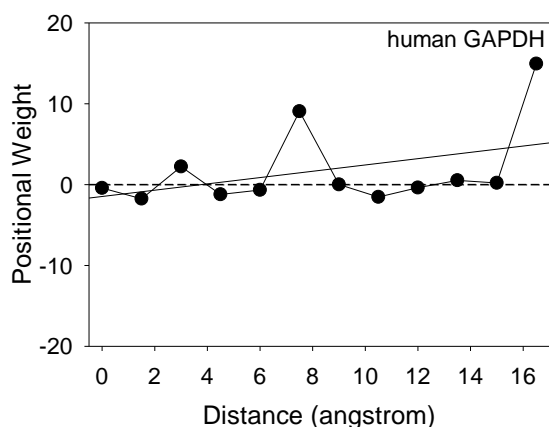


Figure 7: Assessment of horizontality of the $\alpha 1$ -helix of human GAPDH.

of human GAPDH. The Positional Weights of each of the 12 residues were plotted over the longitudinal distance of the helix. The dotted line, a reference point of horizontality. The solid line is the calculated regression line.

Upon evaluation of the six human pseudogenes (i.e. GAPDH-P71, -P62, -P34, -P2, -P19 and -P58) that passed the first criterion of disablement in that they all showed a defined central longitudinal plane, we observed that all of them also passed the second criterion of disablement. Each of these pseudogenes was assessed by the same procedures used for *Listeria monocytogenes* GAPDH (Figure 6) and human GAPDH (Figure 7). While there was a visual displacement from the horizontal reference line, none of the regression lines exhibited a significant deviation from horizontality as determined by a Pearson r analysis. As a representative example, GAPDHP62 analysis is shown in Figure 8.

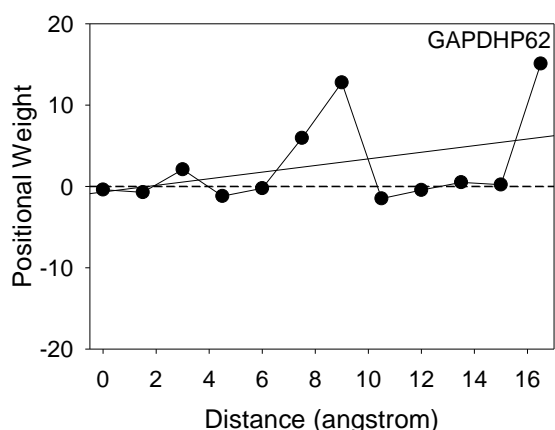


Figure 8: Assessment of horizontality of the α 1-helix of GAPDHP62. The Positional Weights of each of the 12 residues were plotted over the longitudinal distance of the helix. Dotted line, a reference point of horizontality; solid line, calculated regression line.

4 Discussion

GAPDH is a highly conserved glycolytic housekeeping enzyme that exists as an asymmetric homotetramer. It is thought to be derived from a single somatic gene. This protein exhibits moonlighting characteristic meaning that it partakes in multiple cellular functions. The cellular processes, with which GAPDH is involved, include gene expression and apoptotic signaling [7]. GAPDH's pro-apoptotic function is associated with dopaminergic cell death and may contribute to Parkinson's disease. We propose that expressed pseudogenes (i.e. GAPDHP44 is on the negative strand of an intron of a

protein phosphatase gene) may play a role as a anti-sense oligonucleotide. This would modulate the active levels of GAPDH mRNA, representing a potentially important function in cell survival.

Intriguingly, many of the moonlighting functions of GAPDH require that the protein reversibly associates with biomembranes. The ability to bind to biomembranes appears to be a highly conserved property [4, 10]. Overall GAPDH exhibits conformational malleability [1, 2, 16], though there are certain regions that appear highly conserved [10].

There are more than 60 human pseudogenes for GAPDH. These pseudogenes, are considered non-functional, although multiple transcripts for GAPDH have been reported [17], suggesting that one or more of these pseudogenes may be actively transcribed. We think that there is selective pressure to preserve these sequences. To examine the efficacy of this hypothesis, we developed criteria to identify whether or not disablement of these pseudogenes occurs. We define disablement in terms of protein structural changes that would affect functionality. In lieu of the ability to examine pseudogene products we examined the nucleotide sequences, converted them to putative amino acid sequences and developed criteria to assess if these sequences were considered disabled.

The criteria for determining disablement involve analysis of the structural features of the α 1-helix of real and *in silico*-converted GAPDH proteins. This particular region is required for membrane-association [4, 10], which we know is a conserved property. The loss of structural integrity at this region of the protein would significantly alter GAPDH's intrinsic properties, including but not limited to membrane-association and NAD^+ -binding. These properties are not easily observable by just inspection of the nucleotide sequences. Mutations may or may not affect the biophysical properties of the resulting protein that may be a product of expressed pseudogenes.

The first criterion involved computation of the central longitudinal plane of α 1-helix. We applied a novel method of analysis to mathematically determine the plane that bisects the helix into hydrophobic and hydrophilic halves that would be partially immersed in the lipid bilayer. The results of this analysis on α 1-helices from two functional GAPDH proteins (i.e. *Listeria monocytogenes* and human) indicate that the calculated central longitudinal plane is not much different that that drawn by visual inspection. We think that a mathematic approach avoids visual bias in assigning the amphipathic division. The assigned angle (i.e. relative to the default starting point) for *Listeria monocytogenes* GAPDH central plane was 24.5° . For the human GAPDH, it is 37.2° . We examined nine

human GAPDH pseudogenes and three of them exhibited unstable central planes, suggesting that the gene product, if transcribed and translated would be severely dysfunctional. We conclude that these three (i.e., GAPDH-P44, -P23, and -P41) pseudogenes contain mutations that render them disabled. The other six pseudogenes exhibited single values for assignment of the central longitudinal plane with four of the pseudogenes falling within 12° of the functional human GAPDH (Table 2).

The next criterion involved looking at the deviation from horizontality due to the longitudinal asymmetric distribution of the residues along the helix. The *Listeria monocytogenes* GAPDH exhibited almost perfect horizontality. The human GAPDH was also not significantly different from a horizontal orientation. The six pseudogenes that passed the first criterion also passed the next criterion in that there was no indication of a significant deviation from horizontality.

5 Conclusion

We think that the criteria developed by this study provide a useful assessment of the functionality of amphipathic helices. The *Listeria monocytogenes* appears to display the most ideal horizontal helix among those examined, including the functional human GAPDH. Interestingly, *Listeria monocytogenes* utilizes GAPDH, and in particular this α 1-helical region, as part of the phagocytotic strategy of virulence, indicating that this sequence appears most ideal for membrane association.

A limitation to these criteria include the chemical nature of the residues that a substituted in the amphipathic helix. While several human GAPDH pseudogenes showed reasonable defined central longitudinal planes and not deviation from horizontality, the amino acid substitutions in the putative pseudogene products may greatly alter their functional. For example, GAPDHP62 and GAPDHP34 exhibited histidine and cysteine substitutions for conserved arginines. These chemical differences may be significant. Conversely, all of these residues (i.e. arginines, histidines and cysteines) have been associated with mono-ADP-ribosylation, which is a catalytic property of *Listeria monocytogenes* GAPDH [4] and likely human GAPDH. This catalytic function is dependent on the α 1-helical region.

A curious observation in the study was noted upon inspection of the results from assessment of horizontality of the *Listeria monocytogenes* GAPDH (Figure 6). The eighth residue (i.e. Phe-18) evokes a repulsive force towards the hydrophobic environment and the tenth residue (i.e. Arg-20) elicits a repulsive force towards the hydrophilic milieu. Both of these residues are close to the bisecting plane, dividing aqueous and lipid compartments. We proposed that these complementary

repulsive forces contribute to the intrinsic mobility of the helix within a lipid bilayer.

6 References

- [1] Pattin AE, Ochs S, Theisen CS, Fibuch EE, Seidler NW. Isoflurane's effect on interfacial dynamics in GAPDH influences methylglyoxal reactivity. *Arch Biochem Biophys* 2010;498(1):7-12.
- [2] Swearingin TA, Fibuch EE, Seidler NW. Sevoflurane modulates the activity of glyceraldehyde 3-phosphate dehydrogenase. *J Enzyme Inhib Med Chem* 2006;21(5):575-9.
- [3] Harrison PM, Gerstein M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol.* 2002;318(5):1155-74
- [4] Alvarez-Dominguez C, Madrazo-Toca F, Fernandez-Prieto L, Vandekerckhove J, Pareja E, Tobes R, Gomez-Lopez MT, Del Cerro-Vadillo E, Fresno M, Leyva-Cobián F, Carrasco-Marín E. Characterization of a *Listeria monocytogenes* protein interfering with Rab5a. *Traffic.* 2008;9(3):325-37.
- [5] Carugo O. Prediction of polypeptide fragments exposed to the solvent. *In Silico Biol.* 2003;3:417-428.
- [6] Harris J, and Waters M (1975) Glyceraldehyde-3-phosphate. In Boyer PD (ed) *The Enzymes*, vol 13. Academic Press, Orlando
- [7] Seidler NW. GAPDH: Biological Properties and Diversity. *Advances in Experimental Medicine and Biology*, vol. 965. Springer. 2012
- [8] Liu YJ, Zheng D, Balasubramanian S et al (2009) Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotranspositional activity. *BMC Genomics* 10:480
- [9] Garcia-Meunier P, Etienne-Julan M, Fort P et al (1993) Concerted evolution in the GAPDH family of retrotransposed pseudogenes. *Mamm Genome* 4:695-703
- [10] Pancholi V, Fischetti VA (1992) A major surface protein on group A streptococci is a glyceraldehyde-3-phosphate-dehydrogenase with multiple binding activity. *J Exp Med* 176:415-426
- [11] Bottoms CA, Smith PE, Tanner JJ (2002) A structurally conserved water molecule in Rossmann dinucleotide-binding domains. *Protein Sci* 11:2125-2137
- [12] Glaser PE, Gross RW (1995) Rapid plasmethyleneamine-selective fusion of membrane bilayers catalyzed by an isoform of glyceraldehyde-3-phosphate dehydrogenase: discrimination between glycolytic and fusogenic roles of individual isoforms. *Biochemistry* 34:12193-12203
- [13] Nakagawa T, Hirano Y, Inomata A et al (2003)

Participation of a fusogenic protein, glyceraldehyde-3-phosphate dehydrogenase, in nuclear membrane assembly. *J Biol Chem* 278:20395-20404

[14] Kaneda M, Takeuchi K, Inoue K et al (1997) Localization of the phosphatidylserine-binding site of glyceraldehyde-3-phosphate dehydrogenase responsible for membrane fusion. *J Biochem* 122:1233-1240

[15] Pierce GN, Philipson KD (1985) Binding of glycolytic enzymes to cardiac sarcolemmal and sarcoplasmic reticular membranes. *J Biol Chem* 260:6862-6870

[16] 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.

[17] Arcari P, Martinelli R, Salvatore F. The complete sequence of a full length cDNA for human liver glyceraldehyde-3-phosphate dehydrogenase: evidence for multiple mRNA species. *Nucleic Acids Res.* 1984;12(23):9179-89.

FPGA Based Accelerator for Bioinformatics Haplotype Inference Application

N. Harb¹, M. A. R. Saghir², Z. Dawy³, and C. Valderrama¹

¹Electronics and Microelectronics Department, University of Mons, Mons, Belgium

²Electrical and Computer Engineering Department, Texas A&M University at Qatar, Doha, Qatar

³Electrical and Computer Engineering Department, American University of Beirut, Beirut, Lebanon

Abstract—¹ *Hardware accelerators have been used to accelerate various bioinformatics applications without altering their accuracy. These accelerators are used to speed up sophisticated algorithms where powerful computational techniques are used to analyse, simulate and estimate biological data. These are hardware accelerators mostly made up of Field Programmable Gate Array (FPGA) or multiple FPGA hybrid systems. One bioinformatics application in need for acceleration is the haplotype inference application. This application is essential in producing maps used to identify complex diseases. It is also used in finding phylogenetic trees that provide relationships among populations. The main objective of this paper is to build an FPGA-hybrid system connected to a host PC that will accelerate PHASE (one important haplotype inference application) and enhance its processing time while maintaining the same accuracy and functionality.*

Keywords: Hardware accelerators, bioinformatics, FPGA, haplotype inference.

1. Introduction

Bioinformatics is an area with strong demand for high performance computing. The solutions for this high performance demand are by using cluster implementations. In a cluster implementation, the applications' processes, that are what the application is supposed to do by means of functions, are executed in parallel [1].

Increase in bioinformatics research activities has led to a huge increase in data stored in public databases like NCBI or EMBL GenBank [2]. This increase in data is caused by three main reasons [1]. One is the increase in research institutes all specialized in various biological research fields with a large amount of data to be generated and so much processing time. Thus causing more demand on higher storage areas and faster processing machines. Another reason is in modern high throughput experiments and workflows. Modern biological studies and experiments, like microarrays, generate huge amounts of data and information about gene expression

because of the thousands of experiments performed simultaneously. The third reason comes from the combination of data and information from different independent sources or databases. This means more processing time just to get these data into the application's needed format. In addition to the increase in data amounts, new bioinformatics applications have been developed to provide more accurate or better quality results than existing solutions.

New forms of distributed and parallel computing were developed to tackle the problem of long processing times. The basic idea is to make the application that runs on the cluster be parallelized on either the process or thread levels and distributed over the available computing nodes. Note also that parallelized code applications can benefit from the use of an FPGA by means of speed. Meaning that, each parallel thread will be executed faster on an FPGA than on a host PC. Since an FPGA is made up of hardware gates, that makes it a very high speed functional block capable of being programmed by almost any function. So, one good candidate for accelerating bioinformatics applications is by implementing bottleneck functions on FPGA(s).

One good example to demonstrate the capability of having an FPGA for accelerating bioinformatics software is the one found in [3]. In this paper, the authors accelerated the Smith-Waterman implementation in the European Molecular Biology Open Software Suite (EMBOSS) suite for publicly available bioinformatics code. This software is widely used to screen gene databases for sequence similarities with many different applications in bioinformatics research areas. They achieved dropping in the processing time from 50,000 sec to 2,000 sec (98%) for huge datasets [3].

In our paper, we present an FPGA based accelerator for haplotype inference application. We start by presenting the target application in Section 2. Following, we analyse and discuss the candidate function in the application suitable for acceleration in Section 3. In Section 4, we show in details how the hardware accelerator is designed and mapped on an FPGA. Results and analysis of the our system are presented in Section 5 before concluding in the last section, Section 6.

¹The author wish to acknowledge the Optimization for Live Interactive Multimedia Processing (OLIMP) project and the Lebanese National Council for Scientific Research who supported this work through grant number 03-08-06.

Table 1: Comparison between PHASE and HAPLOTYPYER in terms of error rates and execution times

SNPs	PHASE error rate	HAPLOTYPYER error rate
1-8	0.0000	0.0298
10-14	0.0000	0.0106
16-24	0.0209	0.0230
25-35	0.0016	dnf
36-40	0.0193	0.1159
Processing Time (seconds)		
SNPs	PHASE PT	HAPLOTYPYER PT
1-8	46.89	4.29
10-14	44.58	54.56
16-24	52.47	7.93
25-35	60.99	dnf
36-40	34.64	20.36
dnf: did not finish, PT: Processing Time in seconds		

2. Haplotype Inference and PHASE Application

Haplotype inference is a way to infer haplotypes from a given genotype sample dataset. This process is essential in producing Single Nucleotide Polymorphism (SNP) maps used to identify different genes associated with complex diseases. It is also used in finding phylogenetic trees that provide relationships among different populations. The determination of haplotypes from a large dataset was very expensive and nearly infeasible. But as the technology improved, computers became faster processing units, and so haplotypes determination became more and more reachable. Researchers became interested in the topic of haplotype inference after questioning how are different populations are related to each others.

PHASE [4] is considered to be the most commonly used application for haplotype inferences due to three major advantages: increased accuracy compared to other applications (like HAPLOTYPYER [5]), wider applicability, and the facility to assess accurately the uncertainty of PHASE calls [4]. This means that at each run iteration (if many iterations were to be executed; optional), the uncertainty in the previous iteration will be taken into account when executing the next iteration for result enhancement. A comparison of PHASE with its nearest competitor HAPLOTYPYER is shown in Table 1 [6]. Table 1 gives for each certain SNP sequence [7] and for each application software the error rate, which is if the whole haplotype for a certain individual has not been inferred correctly. Table 1 also summarizes the processing time needed by each application to produce its output. Regarding the measure of accuracy, PHASE substantially outperforms the HAPLOTYPYER application [6], but as for the processing time, PHASE takes more processing time than the HAPLOTYPYER.

3. Application Study and Analysis

The application we want to accelerate is PHASE. It is considered one of the best haplotype inference algorithms

Table 2: PHASE execution time for different datasets

Dataset	Processing Time (seconds)
5 Individuals, 5 SNPs	1
5 Individuals, 10 SNPs	1.28
9 Individuals, 20 SNPs	2.54
20 Individuals, 10 SNPs	6.46
20 Individuals, 20 SNPs	86.8
34 Individuals, 20 SNPs	304.18
5 Individuals, 93 SNPs	7510.12

Table 3: Profiling results for a sample of 5 individuals, 93 SNPs dataset

%Time	CS	SS	SC	Name
69.9	6128.91	6128.91	6M	ForwardsAlgorithm
6.01	6679.73	550.82		ieee754_exp
5.7	7201.84	522.11	4M	FDLSProb
CS: Cumulative Seconds, SS: Self Seconds, SC: Self Calls				

in terms of error rate [6]. In order to accelerate PHASE, profiling of the application should be implemented in order to pin out bottleneck functions.

3.1 PHASE Profiling

PHASE requires long processing times due to the high computation demanding algorithms implemented, especially when the input dataset becomes large. Table 2 shows PHASE's processing time as the input dataset varies. Note that the datasets used throughout the entire study are for Lebanese individuals extracted when studying the relationship between the Lebanese population and other populations. All measurements were taken when running PHASE on a 3 GHz Pentium IV machine with 1 GByte in RAMs.

To accelerate PHASE, we profiled the application in order to find out where most of the processing time is taking place. Furthermore, this will lead in finding out what functions consumed most of the processing time. This will also help in picking up candidate functions to be accelerated. Since PHASE is run on a Linux machine, we used the Linux GNU Profiler (GPROF) profiling tool to profile it. A GPROF sample output summary can be seen in Table 3.

From results in Table 3 we have chosen the candidate for acceleration to be the *ForwardsAlgorithm* function. This is due to the facts that this function does not call any other function (it has no branch functions), it does not use any file transactions, and it consumes around 70% of all of PHASE's processing time (most time consuming function in PHASE).

3.2 ForwardsAlgorithm

ForwardsAlgorithm is the computation function inside the function *FDLSProb*. Each time *ForwardsAlgorithm* is called, it fills out either *Alpha* and *AlphaSum* arrays or *Beta* and *BetaSum* arrays with minor changes among both arrays calculations. The final returned output of *ForwardsAlgorithm* will be a one element of the *AlphaSum* or the *BetaSum* array. The naming of the variables inside *ForwardsAlgorithm* is

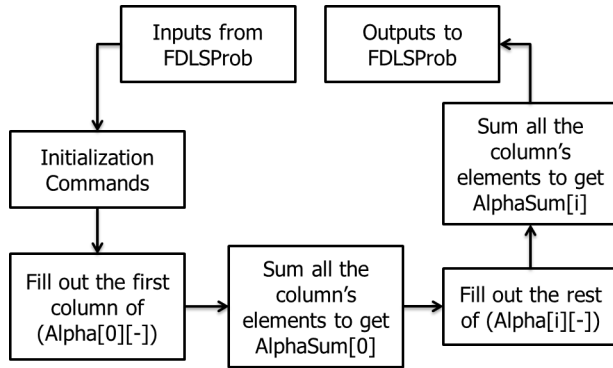


Fig. 1: General flow of the *ForwardsAlgorithm* function.

always *Alpha* (declaration wise). The flow of *ForwardsAlgorithm* function is described in Figure 1.

In general, *ForwardsAlgorithm* calculates a two dimensional array, *Alpha*, and a one dimensional array, *AlphaSum*. *Alpha* is of size $2 * positiveHapsSize \times NLocs$. *AlphaSum* is of size $NLocs$. First, $Alpha[0][n]$ is calculated and then $AlphaSum[0]$ can be calculated by summing all the elements of $Alpha[0][n]$. The function then calculates $Alpha[i][n]$ that depends on $Alpha[i-1][n]$ and $AlphaSum[i-1]$. And then sum all $Alpha[i][n]$ elements to get $AlphaSum[i]$. At the end of the function, *ForwardsAlgorithm* returns the last *AlphaSum* element.

3.3 Software Optimization

In PHASE, the aim is to accelerate this application software using FPGAs. And the most important function we need to accelerate is the *ForwardsAlgorithm* function. But in this function, some redundancy is found through the following:

In *FDLSProb*, it is calculating a big vector called *TransProb*. This array is used for further calculations within the same function. Now, within *FDLSProb* we are calling the function *ForwardsAlgorithm* that calculates the same vector *TransProb* using the same calculations and inputs. *TransProb* is also used by *ForwardsAlgorithm* for some calculations. The problem rising here is if *TransProb* is the same vector used in both *FDLSProb* and *ForwardsAlgorithm* so there is no need to do the calculations twice especially inside *ForwardsAlgorithm* since it is being called by *FDLSProb*. The design alternative is to make a new function with different input arguments in which the vector *TransProb* is in these arguments.

As an alternative solution of this problem, we changed only one header file that is the one containing only *FDLSProb*. Also, we calculated *TransProb* only once in *FDLSProb* rather than doing it two times and then, send *TrasProb* as an input to *ForwardsAlgorithm*.

4. Hardware Accelerator Design

In this section, we will present the steps followed to build up the hardware accelerator for PHASE, presenting also the functional and interfacing blocks. But, before building up the hardware accelerator, some tests should be implemented to assure system correct functionality by means of arithmetic precision.

4.1 Arithmetic Precision Issues

One of the most important variables used to calculate some relevant elements in PHASE and that affects the system's final output, is a two element vector *WEIGHTS*. It was found out that the best number of representative bits that could not affect the system's final output is 51 bits. Since, as the number of bits decrease the value of *WEIGHTS* change, it was relevant to study the effect of this change on PHASE's final output.

The reason for studying this vector is that it is not an input to the *ForwardsAlgorithm* function but rather embedded in it. PHASE's input variables require 32 bits to be represented, but *WEIGHTS* require more.

Before presenting the analyses done, it is important to understand some major points found in the output files of PHASE. At each heterozygous (1 at haplotype1 and 0 at haplotype2 or vice versa), the system will give a certain order of SNPs for each individual with a certain probability that this is the correct combination. PHASE also gives an '=' character for each position it was 100% certain that it is of a correct combination (that is given for two cases, the first is for normal homozygous; 0 at haplotype1 and 0 at haplotype2, and the second is for abnormal homozygous; 1 at haplotype1 and 1 at haplotype2). When changing the values of *WEIGHTS*, the change was only at the heterozygous positions and no change in *WEIGHTS* affected the homozygous (normal or abnormal) final outcome compared to the original output file. While analysing, we measured the number of flips occurrence at the heterozygous positions for all the population sizes. Figure 2 shows the total number of flips for each population versus the number of representative bits for *WEIGHTS*.

Notice that at a representation of 51 bits, the flips are 0 in all populations since the value represented by these bits is still the same as the original value. But as the number of representing bits decreases, more flips start to appear in the genotype of each individual. The change is not uniform in PHASE when the *WEIGHTS* change, sometimes this will cause a flipping and in turn change the whole haplotype of an individual. And in case this haplotype was the most frequent one in the population, or repeated more often in other individuals, then this haplotype will appear in other individuals with the flipping in it and thus increasing the number of flips in this data set. But sometimes there would not be so much flips, talking only about the occurrence of a flip, since when the number of representing bits decrease,

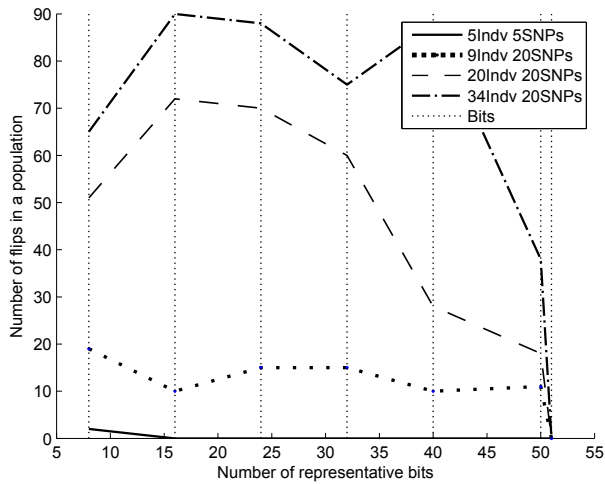


Fig. 2: WEIGHTS Change Effect on PHASE's Accuracy.

this may not give a flip. This will rather affect the probability that this order (original order; not flipped) is correct.

4.2 Hardware Building Blocks

Referring back to Figure 1, the figure shows that the function is divided upon two parts: Filling out $Alpha[0][n]$ and $AlphaSum[0]$ part (that will be called Command 1), and filling out $Alpha[i][n]$ and $AlphaSum[i]$ part (that will be called Command 2).

4.2.1 Command 1

Command 1 starts by testing a variable called *usequad* whether it is 1 or 0. If it is a 1 (execute Block 1), then two elements of the $Alpha[0][n]$ array will be filled using the two elements in the WEIGHTS vector. And if it is a 0 (execute Block 2), one element of the $Alpha[0][n]$ array will be filled but without the use of the WEIGHTS vector. After this is done, all the Alpha elements will be summed together to give out $AlphaSum[0]$.

Block 1 takes as inputs: *SS*, *nchr*, *Freq* array, *PrHitTarg1* array, *ismissing[0]*, and the WEIGHTS vector and gives as an output, two elements in the Alpha array. Each time an $Alpha[0][n]$ element is calculated, a test of the *ismissing[0]* value is implemented, if it is a 0, then the old value of $Alpha[0][n]$ is multiplied by $PrHitTarg1[n][0;SS=1/1;SS=2]$ to give an updated Alpha value. Else the value of $Alpha[0][n]$ is not altered. Figure 3 shows the exact calculations done inside Block 1 when *ismissing[0]* is 1.

Block 2 takes as inputs: *nchr*, *Freq* array, *PrHitTarg2* array, and *ismissing[0]* and gives as an output, one element in the Alpha array. Each time a $Alpha[0][n]$ element is calculated, a test of the *ismissing[0]* value is implemented, if it is a 0, then the old value of $Alpha[0][n]$ is multiplied by $PrHitTarg2[n]$ to give an updated Alpha value. Else the

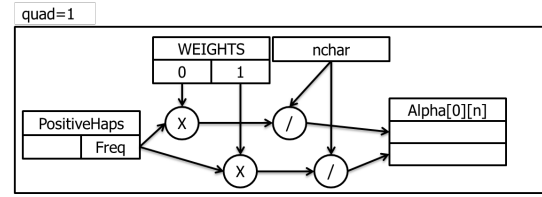


Fig. 3: Block 1 computations.

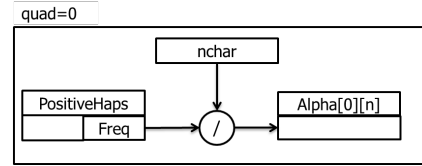


Fig. 4: Block 2 computations.

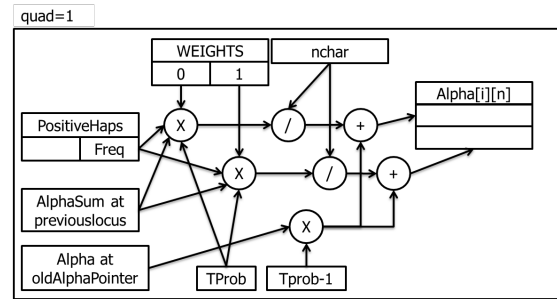


Fig. 5: Block 3 computations.

value of $Alpha[0][n]$ is not altered. Figure 4 shows the exact calculations done inside Block 2 when *ismissing[0]* is 1.

4.2.2 Command 2

Command 2 starts by testing the variable *usequad* whether it is 1 or 0. If it was a 1 (execute Block 3), then two elements of the $Alpha[i][n]$ array will be filled using the two elements in the WEIGHTS vector. And if it is a 0 (execute Block 4), one element of the $Alpha[i][n]$ array will be filled by deducting the WEIGHTS vector. After this is done, all the Alpha elements will be summed together to give out $AlphaSum[n]$.

Block 3 takes as inputs: *SS*, *nchr*, *Freq* array, *PrHitTarg3* array, *ismissing[i]*, *TProb[i]*, $AlphaSum[i-1]$, $Alpha[i-1][n]$ and the WEIGHTS vector and gives as an output, two elements in the Alpha array. Each time an $Alpha[i][n]$ element is calculated, a test of the *ismissing[i]* value is implemented; if it is a 0, then the old value of $Alpha[i][n]$ is multiplied by $PrHitTarg3[i][n][0;SS=1/1;SS=2]$ to give an updated Alpha value. Else the value of $Alpha[i][n]$ is not altered. Figure 5 shows the exact calculations done inside Block 3 when *ismissing[i]* is 1.

Block 4 takes as inputs: *nchr*, *Freq* array, *PrHitTarg4* array, and *ismissing[i]*, *TProb[i]*, $AlphaSum[i-1]$, and $Alpha[i-$

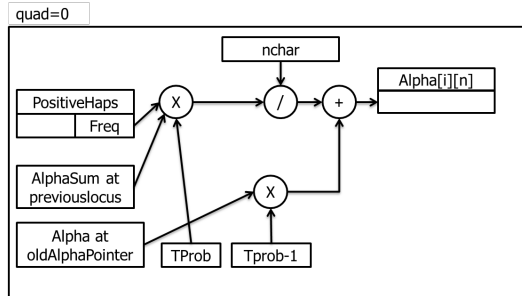


Fig. 6: Block 4 computations.

$1][n]$ and gives as an output, one element in the *Alpha* array. Each time a $Alpha[i][n]$ element is calculated, a test the $ismissing[i]$ value is implemented, if it is a 0, then the old value of $Alpha[i][n]$ is multiplied by $PrHitTarg4[i][n]$ to give an updated *Alpha* value. Else the value of $Alpha[i][n]$ is not altered. Figure 6 shows the exact calculations done inside Block 4 when $ismissing[i]$ is 1.

4.2.3 Building Blocks

Looking back into the blocks described earlier, there are common operations among all blocks. From this concept, the building of the hardware root components became essential. These components are:

- **35 Bits Multiplier:** Since the input data is 32 bits masked, multipliers that can support this width should be built.
- **69 Bits Multiplier:** Since the *WEIGHTS* vector elements need more than 51 bits to be represented, this multiplier was built.
- **Twos Complement Block:** To avoid sending $1-TProb$ to the system, the twos complement of $TProb$ was implemented to give $1-TProb$.
- **Division to Multiplication Converter:** After monitoring the variable $nchar$, its value seems not to exceed 100, so a division by $nchar$ was converted into a multiplication by $1/nchar$.
- **Addition Accumulator Block:** After calculating the *Alpha* array elements, each time one element is calculated, it will be sent to this block in order to add it to the previously calculated *Alpha* element and at the end result in *AlphaSum*.

4.3 Implementation Summary

The hardware accelerator system consists of all the Blocks 1 through 4, an internal control mechanism and the communication core *SV_IFACE*. Additional FIFOs are added (*FIFO_x* and *FIFO_y*) in order to buffer internal values used by proceeding iterations in their calculations. These FIFOs are used as temporary storage spaces only. The flow of the system is described in the following steps:

Table 4: Resource Usage of our hardware design on a Virtex II FPGA

Resource	Available	Consumed	Util. %
Slices	14336	1762	12
Look Up Tables	28672	3154	11
Generated clocks	16	3	18
Flip Flops	28672	274	1
Input/Output Bounds	484	49	10
Block RAMs	96	11	11
18×18 Bits Multipliers	96	96	100

L1– Loop through *PositiveHaps*:

- 1- Read from the FIFO all the necessary data used by B1B2 to produce an *Alpha* element.
- 2- Write each *Alpha* element to a FIFO (*Alpha_x_FIFO*; *FIFO_x*) for future usage and send this element to the addition accumulator to later produce *AlphaSum*[0].

L1– End Loop *PositiveHaps*.

- 3- Send of *AlphaSum*[0] to command 2 in order to begin command 2's calculations (note that now $Alpha[0][n]$ is stored in *Alpha_x_FIFO*).

L2– Loop through *Nloci*:

- 4- Read from the FIFO all necessary data used by the next stage to produce a new *Alpha* column and an *AlphaSum* element. This procedure also includes sending *AlphaSum*[0] to the next stage.

L3– Loop *PositiveHaps*:

- 5- Read an *Alpha* element from *Alpha_x_FIFO*, *AlphaSum*, and all B3B4 needed data.
- 6- Write each *Alpha* element to a new FIFO (*Alpha_y_FIFO*; *FIFO_y*) for future usage and send this element to the addition accumulator to later produce an *AlphaSum*.

L3– End Loop *PositiveHaps*.

- 7- Update the to be sent *AlphaSum* and flip the reading and writings from and to *Alpha_x_FIFO* and *Alpha_y_FIFO*.

L2– End Loop *Nloci*.

- 8- Send the final *AlphaSum* value to the output.

4.4 Resource Management

The hardware accelerator system was implemented on an XtremeDSP Development Kit-II with a Xilinx Virtex-II user FPGA. Table 4 shows the kit's available resources and the resources consumed by the hardware accelerator.

In Table 4, the main and the most important resources consumed by the hardware accelerator are the Input/Output Bounds, Block RAMs and the 18×18 Bits Multipliers. Our system uses 49 Input/Output Bounds all consumed by the *SV_IFACE* component. The system also uses 11 Block RAMs distributed as follows: 7 Block RAMs are used for buffering the inputs, 2 Block RAMs are used in the division to multiplication converter building block and 2 Block RAMs are used by both the *FIFO_x* and *FIFO_y*. Our system consumes all the 96 embedded 18×18 Bits

Multipliers in which 36 multipliers are used in both Blocks 1 and 2, and 60 multipliers are used in both Blocks 3 and 4.

5. Results and Analysis

This section presents how much the *ForwardsAlgorithm* function and PHASE application will measure in processing time on a standalone PC and how much will it measure on the new FPGA-hybrid system. Analysis for these processing times is discussed in details.

5.1 *ForwardsAlgorithm* Execution Time on a Standalone PC

The total time consumed by the function *ForwardsAlgorithm* is a function of the time consumed by both Command 1 and Command 2. Command 1 runs *positiveHapsSize+1* times which will be referred to as X , and Command 2 runs $Nloci(positiveHapsSize+1)$ times which we will be referred to as XY .

Note that, *quad* variable represents when the quadrature option is instantiated. Meaning that if *quad* is 0 (do not use the quadrature option), both Command 1 and Command 2 will be executing their computations once each time one of these commands is being called. In case *quad* is 1 (use the quadrature option), both Command 1 and Command 2 will be executing their computations two times each time one of these commands is being called. When *quad* is 0, the computations done in Command 1 consumed 100 ns to execute and the operations done in Command 2 consumed 1000 ns to execute. When *quad* is 1, then Command 1 consumes 200 ns and Command 2 consumes 2000 ns. Now that for different *quad* values, the execution time by each command is known. In addition to the number of iterations each command is called, two equations can be derived showing the total execution time needed by *ForwardsAlgorithm* on a standalone PC as a function of X and Y for *quad* equals to 0 in (1) and for *quad* equals to 1 in (2).

$$T_{quad=0}(ns) = 100X + 1000XY \quad (1)$$

$$T_{quad=1}(ns) = 200X + 2000XY \quad (2)$$

5.2 *ForwardsAlgorithm* Execution Time on an FPGA-Hybrid System

The time consumed by the accelerated system is divided into two time consuming parts. The time consumed by the operations running in the host PC (register reads, DMA writes) and the time needed by *ForwardsAlgorithm* in hardware to finish its operations and send the final result back to the host PC.

5.2.1 Software Time

The FPGA-hybrid system needs some time to configure, control, send input data, and read the output result from the FPGA; these occur on the host PC. The time needed for

Table 5: Transaction times for *quad=0* and *quad=1*

Operation	quad=0	quad=1	Array Sent
Write to Register R2	25 ns	25 ns	
Write to Register R1	25 ns	25 ns	
Write DMA Burst	25X ns	50X ns	PHT1
Write to Register R1	25 ns	25 ns	
Write DMA Burst	25X ns	25X ns	PHT2
Write to Register R1	25 ns	25 ns	
Write DMA Burst	25X ns	25X ns	FREQ
Write to Register R1	25 ns	25 ns	
Write DMA Burst	25Y ns	25Y ns	TPROB
Write to Register R1	25 ns	25 ns	
Write DMA Burst	25XY ns	50XY ns	PHT3
Write to Register R1	25 ns	25 ns	
Write DMA Burst	25XY ns	25XY ns	PHT4
Write to Register R1	25 ns	25 ns	
Write DMA Burst	25Y ns	25Y ns	ismissing
Read Register R2	0 ns	0 ns	

the configuration of the board will not be calculated since it can be accomplished before even running PHASE. The only times, we are interested in measuring, are the execution times needed by the sending and receiving operations.

There are 16 commands that need to be executed on the host PC: 8 register writing operations (5 ns time consuming each), 7 DMA data writing operations (5 ns time consuming each), and one register reading operation (30 ns time consuming). These commands will end up consuming 105 ns in total.

5.2.2 Hardware Time

Before discussing the time consumed by the hardware, it must be noted that the FPGA clock period is 25 ns. On the hardware side, the time is divided into two parts. One part is consumed by the SV_IFACE while writing data to the input FIFOs. Another part is consumed by *ForwardsAlgorithm* in hardware to do its calculations and produce the final output. Each writing or reading operation to a FIFO takes 25 ns. *ForwardsAlgorithm* in hardware takes 25 ns to produce a single element of *Alpha* array.

From the above, we need 25 ns to buffer each register written while the reading of the final output register takes 0 ns because the data will be already valid and just ready to be read (8 register buffering operations = 7×25 ns = 200 ns). As for the DMA burst buffering, it depends on the size of the buffered data array and the *quad* value. Table 5 summarizes the time consumed by the hardware when *quad=0* and when *quad=1*.

The timings in Table 5 are consumed by the hardware to only write data to the input FIFOs. The execution time of all Command 1 in the hardware takes 25X ns for *quad=0* and 50X ns for *quad=1* while that of Command 2 in hardware takes 25XY ns for *quad=0* and 50XY ns for *quad=1*.

5.2.3 Total Time

The total time consumed by the FPGA-hybrid system is simply the addition of all software and hardware time

consumptions. But note that, not all the software time will be added, but rather only the first 5 ns done for register R2 writing operation. The reason for that is because the time consumed in the hardware part is much bigger than that of the software part. So while the software is consuming some time to send the rest of the data, the hardware has already enough buffered data to process (in the interface FPGA).

So the total time consumed when $quad=0$ is the sum of all what is in Table 5 (for $quad=0$), 5 ns, 25X ns, and 25XY ns and that gives (3).

$$T'_{quad=0}(ns) = 205 + 100X + 25Y + 75XY \quad (3)$$

The total time consumed when $quad=1$ is the sum of all what's in Table 5 (for $quad=1$), 5 ns, 50X ns, and 50XY ns and that gives (4).

$$T'_{quad=1}(ns) = 205 + 150X + 50Y + 125XY \quad (4)$$

While studying the timing of the system and the execution schemes, some gain can be acquired. When measuring the time needed by Command 1 on hardware to finish (including input FIFO writing), we figured out that some pipelining can be implemented and some gain in the speed can be achieved.

The point is to write to the input FIFO the data needed by Command 1 to execute and then write the rest of the data. By doing so, all the operations and cycles done in Command 1 can be processed before the rest of the input FIFO is being filled. Looking back at Table 5, Command 1 can start execution whenever *FREQ* has finished filling its FIFO. From that 25X ns with $quad=0$ and 50X ns with $quad=1$ can be considered as pipelining gain. Hence, by subtracting these values from the new execution time of what is in equation 3 will look like equation 5 and that of what is in equation 4 will look like equation 6.

$$T''_{quad=0}(ns) = 205 + 75 * X + 25 * Y + 75 * X * Y \quad (5)$$

$$T''_{quad=1}(ns) = 205 + 100 * X + 50 * Y + 125 * X * Y \quad (6)$$

5.3 Acceleration of PHASE

After accelerating *ForwardsAlgorithm*, the acceleration's impact on PHASE is studied. As explained earlier, *ForwardsAlgorithm* consumes around 70% of PHASE's total execution time. It is important to mention that the *quad* value flips between 1 and 0 throughout the entire running time of PHASE. For different datasets, different X, Y, and *quad* values were deducted in order to calculate the speedup factor (SUF) of *ForwardsAlgorithm* (FA) and in turn find the total SUF acquired. Table 6 summarizes the speedup factor of PHASE running on the new FPGA-hybrid system. As a conclusion of the results in Table 6, for big datasets, our system can achieve around 16 times speedup for *ForwardsAlgorithm*, leading to a maximum speedup of around 3 times of the PHASE total execution time.

Table 6: Accelerated PHASE Speedup Factors for different datasets.

Dataset	FA SUF	PHASE SUF
5Indv_5SNPs	11.47	2.77
5Indv_10SNPs	12.61	2.81
9Indv_20SNPs	14.21	2.86
20Indv_10SNPs	14.34	2.86
20Indv_20SNPs	15.17	2.88
34Indv_20SNPs	15.88	2.90
5Indv_93SNPs	15.9	2.98

6. Conclusion

The aim of this paper, was to present an educational like flow on how to accelerate a target application using FPGAs following a step-by-step approach. Our work was based on a FPGA based accelerator for haplotype inference application PHASE. We start by providing an overview of related biological topics and an overview of previously implemented FPGA-based accelerators. We then discussed some important existing algorithms that perform haplotype inference. By selecting PHASE as our target application, we analyse the application by means of input and output parameters providing a brief analysis of its functionality. Then, we profiled PHASE and picked up candidate functions for acceleration. A full analysis of the acceleration candidate function *ForwardsAlgorithm* was followed. In the implementation phase, we investigated the FPGA used and tested its functionality in terms of resources, clocking schemes, memory, etc. Following that, we dissected *ForwardsAlgorithm* into small blocks and implement them on the FPGA and use an interface scheme between the host PC and the FPGA. Finally, we built and joined the whole system and test it using real datasets. After implementing the system, the results show that the developed FPGA based accelerator managed to accelerate PHASE by an average factor of 2.9. This means that we managed to eliminate around 63% of PHASE's total execution time.

References

- [1] A. Y. Zomaya, "Parallel Computing for Bioinformatics and Computational Biology", *John Wiley and Sons Inc.*, 2006.
- [2] "National Center for Biotechnology Information NCB", available at <http://www.ncbi.nlm.nih.gov>.
- [3] A. B. Wilkinson, A. Mukherjee, A. Ravindran, C. Gibas and R. K. Karanam, "Using FPGA-based Hybrid Computers for Bioinformatics Applications", *Xcell Journal Third Quarter*, vol. 58, pp. 80-831, Jul. 2006.
- [4] M. Stephens, N. J. Smith and P. Donnelly, "A New Statistical Method for Haplotype Reconstruction from Population Data", *Am. J. Hum. Genet.*, vol. 68, pp. 978-989, 2001.
- [5] T. Niu, Z. S. Qin, X. Xu and J. S. Liu, "Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms", *Am. J. Hum. Genet.*, vol. 70, pp. 157-169, Nov. 2002.
- [6] E. Halperin and E. Eskin, "Haplotype Reconstruction from Genotype Data Using Imperfect Phylogeny", *OUP Journals*, vol. 20, pp. 1842-1948, Aug. 2004.
- [7] M. Daly, J. Rioux, S. Schaffner, T. Hudson and E. Lander, "High-Resolution Haplotype Structure in the Human Genome", *Nat. Genetics*, vol. 29, pp. 229-232, Oct. 2001.

Methodological Procedure for Decision-Making Using Fuzzy Inference for SNP Discovery

Wagner Arbex¹, Marta Martins¹, Marcos Vinícius Silva¹ and Luis Alfredo Carvalho²

¹Brazilian Agricultural Research Corporation, Juiz de Fora, MG, Brazil

²Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

Abstract—Problems when dealing with imprecise or uncertain features, e. g., problems of decision-making, can be designed as fuzzy systems, since these systems allow subjective and qualitative arguments, which are usually intrinsic in such problems, to be processed. Research involving the discovery of single nucleotide polymorphisms (SNPs) requires bioinformatics tools to be applied to different cases with an ability to analyze “reads” from different sources and levels of coverage and also to establish reliable measures. These tools work with different methodologies in regards to distinct attributes. When dealing with the same data set, similar results are expected. However, sometimes such different methodologies may yield different results, which leads to uncertainty in the decision-making process. This paper presents a methodology based on the fuzzy inference decision model applied to bioinformatics, based on results from two other tools for SNPs discovery.

Keywords: Fuzzy inference, decision support, single nucleotide polymorphism, SNP, SNP discovery

1. Introduction

Data generation technologies for molecular biology challenge the development of appropriate computer systems and require accurate bioinformatics tools for analyzing such data. In this sense, machine learning appears as a promising alternative for knowledge discovery in genomic databases, using both decision-making and data mining techniques, among other resources of artificial intelligence.

In the fields of genomics and bioinformatics, the already great amount of data continues to grow very quickly, widening the gap between the generation and interpretation of such data. Therefore, different ways to reduce the problem of huge quantities of data as opposed to the ability to interpret them are studied. For instance, fuzzy inference systems implement computational models for data mining aimed at discovering knowledge in databases. Such models are capable of processing imprecise and qualitative information and, therefore, they are suitable in situations that require decision-making [1].

This paper aims at describing a computational model that uses fuzzy logic as the basis for the implementation of an inference system aimed at assisting decision-making. More information about the inference model proposed and its

applications can be found in the research project “Computational models for the identification of genomic information associated to the resistance to cattle tick” [2].

In support to such description, the concept of single nucleotide polymorphism (SNP) and the use of fuzzy inference to deal with uncertainty, imprecision and decision-making problems will be presented. Following, the fuzzy inference model and the methodological approach will be presented and discussed. They work on previous results obtained by different SNP discovery tools that have possibly conflicting results; therefore, the methodology is applied to assist decision-making in cases when information is conflicting and also in the confirmation of coincident information.

2. Background

2.1 Single Nucleotide Polymorphisms

Sequencing projects have shown that genomes have more variations and more complexity than initially expected. One of such variations and peculiarities are the SNPs, that is, base pairs in a single position in genomics DNA that are presented in sequences with different alternatives [3]. SNPs can be found in the genome of a single individual or groups of individuals, in a given population (Fig. 1).

```

... GGGCAACTCCAG... .. GGGCAACTCCAG... .. GGGCAACTCCAG...
... GGGCAACTCCAG... .. GGGCACTCCAG... .. GGGCACTCCAG...
... GGGCAACTCCAG... .. GGGCACTCCAG... .. GGGCACTCCAG...
... GGGCACTCCAG... .. GGGCACTCCAG... .. GGGCACTCCAG...
... GGGCACTCCAG... .. GGGCACTCCAG... .. GGGCACTCCAG...

```

Fig. 1: Hypothetical instances of SNPs bi, tri and tetra-allelic, respectively. The first line, in bold, shows the consensus sequence and the underlined bases are the SNPs. Actually, the occurrence of bi-allelic SNPs it's not only more common, but almost absolute in relation to the others [4].

Individuality is a result of genetic expression, that is, in essence, the nucleotide sequences form DNA and RNA, as well as protein sequences, which interact and, in turn, form cells, which also interact and form tissues, organs, until, eventually, make individuals. In this relies the importance of SNPs: if a single nucleotide, a single base in a given sequence, is changed, it may alter the formation of proteins and, altogether, these changes may cause variations in the individuals.

2.2 Fuzzy Inference Approach

Classical approaches are insufficient to analyze values very close to the limits of a given category; therefore, one may get results that are questionable, though mathematically and logically accurate. For instance, the Polyphred Score (PPS) [5] determines six classes with precise intervals (Tab. 1). Assuming that the scores 70 and 89 were taken for two points, respectively, then, when deciding whether these two points are SNPs, a 35% of true positives rate (Rank 4) would be considered for both.

Table 1: Accuracy by PPS and rank.

Rank	PPS	True positives rate
1	99	97%
2	95 – 98	75%
3	90 – 94	62%
4	70 – 89	35%
5	50 – 69	11%
6	0 – 49	1%

This logically and mathematically precise decision can be questioned because of the subjectivity involved. Both scores, 70 and 89, are very close to the limits of the classes to which they belong, and, therefore, different interpretations are supported for these scores. However, traditional approaches to logics and mathematics do not have the necessary tools to handle threshold values, or even imprecision or uncertainty. Specifically, threshold values result in doubt when it comes to deciding whether a given base is polymorphic or non-polymorphic, which suggests a fuzzy inference system for handling this uncertainty.

Usually, the problem with threshold values is not as simple as it may seem, if it were so, classical approaches could easily solve it. However, the closer to the subjective reasoning for the interpretation and the extraction of an answer or a decision, the more complex it becomes and the apparent simplicity is given by fuzzy logic modeling and by its basis in the theory of fuzzy sets.

3. Decision-Making with Fuzzy Inference

The subjectivity inherent to reasoning is capable of dealing with complex situations, based on inaccurate, uncertain or approximate information and, therefore, the strategy is to use human operators of an also imprecise nature, which are expressed in linguistic terms or variables. In order to describe or handle problems, such essentially human proposal, generally, does not generate a solution in terms of exact numbers, but, for instance, leads the solution to a qualitative classification, clustering or aggregating results into categories or possible solutions sets [1]. These solutions can be seen as a result of the “principle of incompatibility” [6].

The linguistic terms or variables increase the complexity of traditional models and computational systems concerning

their ability to handle exact numbers and discrete values – which are, sometimes, mutually exclusive. Hence, working with uncertain values may enable the modeling of complex systems, even if they reduce the accuracy of the result, without, thought, leading to loss of credibility.

If uncertainties, when viewed in isolation, are undesirable, when they are associated with other characteristics, they generally allow the reduction of system complexity and increase the credibility of the results [7].

Fuzzy sets theory and fuzzy logics are appropriate to represent, in mathematical terms, the inaccurate information that can be expressed by a set of linguistic rules. Also, if there is the possibility for human operators to be organized as a set of conditional statements (in the if ANTECEDENT then CONSEQUENT form); thus, subjective reasoning can be expressed in the form of computationally executable algorithms [8] [6] with the ability for imprecisely classifying the antecedent and consequent variables of conditional statements as qualitative (instead of quantitative) concepts, which represents the idea of a linguistic variable [1].

Hence, since they are capable of efficiently processing inaccurate and qualitative information, fuzzy inference models are suitable in situations that require decision-making [1].

4. Fuzzy Inference model for identification of SNP candidates

4.1 Methodological Procedure with Fuzzy Inference System for Decision-Making

The function structure of the machine learning model for decision-making is represented in Fig. 2 and 3, in which there is emphasis on the division of the system’s workflow in well-defined stages:

- 1) initial processing of the chromatograms, when bases are read, and, consequently, sequences (“reads”) and contiguous sequences are originated and, besides, the quality of the bases of these sequences is determined. This stage is done by phredPhrap pipeline [9], and many files are generated, such as the “ace” file and several “phd” files, from each sequence read (Fig. 2);
- 2) Polyphred [10] and Polybayes [11] software run on “ace” and “phd” files, and each of these programs, following its own methodology, identifies the SNP candidate bases and determines a probability for each of these bases. These results are recorded in “polyphred.out” and “report.out” files, which will be used as input to the learning procedure (Fig. 2);
- 3) in this next stage, preparation of the data is carried out. Data from Phrap [9] – generated by the phredPhrap pipeline – Polyphred and Polybayes are extracted and selected from their respective files and, if necessary, they are complemented. This stage of preparing the data is done by parsepolyBayes.pl, parsepolyPhred.pl, parsephrapQuality.pl and joinparsersOut.pl scripts [2].

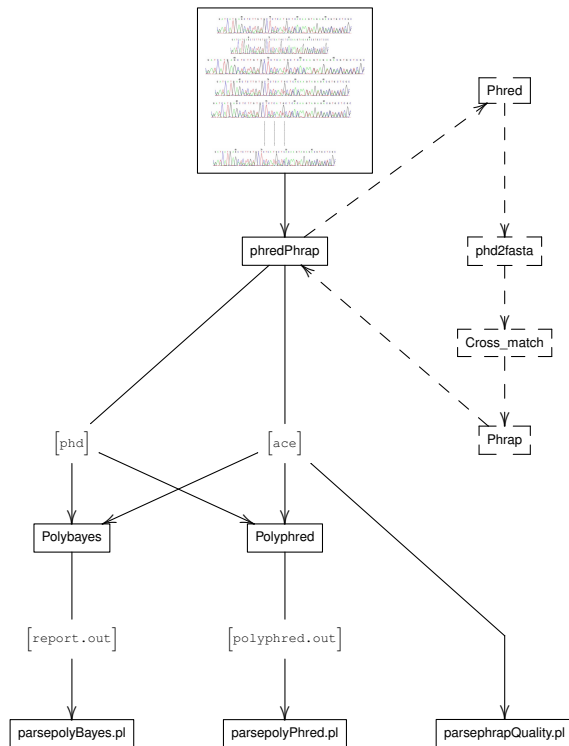


Fig. 2: Synthesis of the functional structure of the model of machine learning (I).

Furthermore, the joinparsersOut.pl script forms the file in a specific structure to fuzzyMorphic.pl [12] software (Fig. 3);

- 4) while running fuzzyMorphic.pl, the machine learning procedure is performed, implementing a fuzzy inference system to make an output file with the same input data, adding the inferred value about the investigated feature (Fig. 3);
- 5) in order to analyze and assess the outcome, we use certified techniques and tools so as to check the inferred results. In this case, a cluster analysis is carried out in the resulting data set, which arises from the fuzzy inference system (Fig. 3).

4.2 Review and Discussion of the Methodological Procedure

The machine learning model implemented, functionally speaking, explores the data set which was created by connecting Polyphred and Polybayes output data sets. Then, it checks the probabilities for each element of this data set, as specified by their different proposals. Next, this model defines, for each element, a new attribute, which should be used as a reference in the attempt to cluster data set into groups of elements that can be seen as confirmed polymorphic points (SNP confirmed), non-polymorphic points (SNP

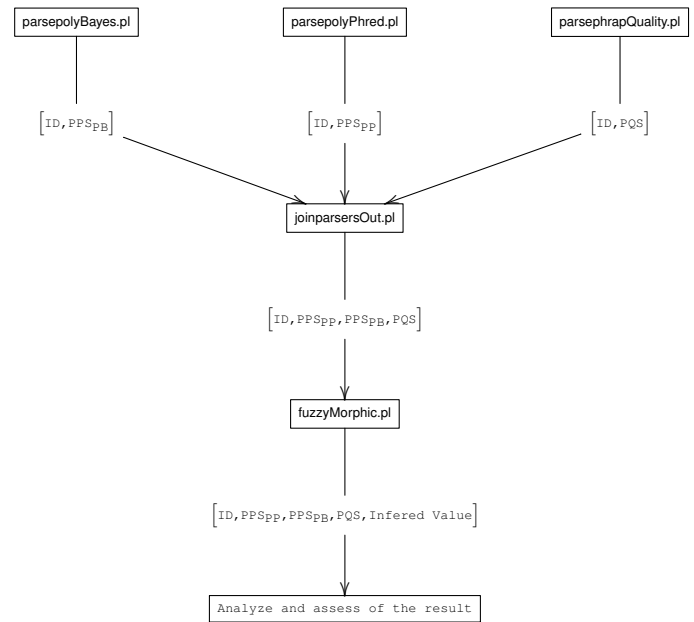


Fig. 3: Synthesis of the functional structure of the model of machine learning (II).

discarded), and also points without sufficient evidence for a conclusive definition (SNP not confirmed).

However, any classification one could propose might be influenced by the data “form” or “behavior”. Also, in regards to classes defined by exact limits, questionable decisions may arise when the value is very close to the limits of classes. These issues, among others, suggest the adoption of non-hierarchical and non-supervised partitioning methods, because these methods do not refer to any external premises to establish the classes that may divide a set, but, rather, its premises are established by specific features, which are internal and inherent to the data set evaluated. Therefore, the adoption of these methods removes or reduces the action of external agents, such as a priori definition of precise limits for the classes, on the model.

Premises of partitioning methods from non-hierarchical algorithms are based on their own set of values assessed, searching for maximum internal cohesion of a group of objects and for maximum detachment between groups [13]. From another perspective, analyzing the set itself, they try to identify the elements that, concerning the attribute evaluated, are closer to the other elements of the group, and, once the groups are established, the elements with a given feature should be as far as possible from the elements belonging to the elements in another group. Thus, as these premises are due to their own values analyzed, the effect of data behavior is reduced, that is, assuming that the attribute evaluated presents a certain trend, all elements have the same behavior and an undirected partitioning taken from elements

themselves can reduce or eliminate this tendency.

The exclusion of external premises as well as the reduction of the models adopted for assessing the results can be advantageous, insofar as they simplify the answers, reducing the risk of them being manipulated. If possible, these models should be self-contained, independent of external components and use as few variables and parameters as possible, avoiding “boundary conditions”, which enable the “accommodation” of a result, instead of truly finding it.

Determination of data clustering is a complex and hard to implement task, because it is necessary to find out how the data are and into how many classes the data are distributed, without any previous knowledge about them. Classes may not even exist, if the elements are distributed equitably over the space and do not feature any category, for clusters or classes are based on the similarity between elements. Eventually, the verification of resulting classes is performed so as to assess whether there is some sort of useful meaning [13].

Following this analysis, the model implemented from machine learning techniques replaces, through fuzzy inference, a continuous probability measure in the interval [0,1] associated with the probability of the point becoming an SNP, by another attribute, which allows clustering the points into three partitions: SNP confirmed, SNP discarded and SNP not confirmed. Thus, after data processing by the fuzzy inference system, which aims at clustering the resulting data through a non-supervised algorithm and dynamically establishing the number of groups, hoping that the result obtained confirms the partitioning of the set into three groups based on the new attribute.

Operationally, this procedure is done by fuzzyMorphic.pl software, which implements the fuzzy inference system and determines this new attribute, while the clustering analyses are aided by Weka (Waikato Environment for Knowledge Analysis) [14] software.

Among clustering algorithms, Weka implements the Expectation-Maximization (EM) algorithm, which has the feature of determining, in runtime, the number of clusters which fits better the elements analyzed, without this information being previously provided to it. EM algorithm was developed for statistical inference problems in general, and it seeks to locate the value for a parameter that maximizes the likelihood function. For the clustering procedure, the data standard division was adopted, that is, 2/3 and 1/3 for training and testing, respectively.

5. Conclusion

Generally, fixed and precise criteria of classification are not suitable when studies show results very close to a certain limit, for instance, a classes division. Nevertheless, these cases can be approached by fuzzy inference systems, which are also convenient, as well as able to handle uncertain and imprecise problems in decision-making.

When adding a new attribute to previous results, the fuzzy system is able to decide, uniquely among the three possibilities resulting from the model, and then it clusters them through a non-supervised algorithm with dynamic establishment of the number of clusters, hoping that the outcome of this clustering confirms set partitioning into three clusters, and requiring no fixed and/or precise limits to classify and, thus, identify potential SNPs.

Acknowledgements

The authors would like to express thanks to the State of Minas Gerais Research Support Agency (Fapemig) for the partial support for the accomplishment of this paper.

References

- [1] P. E. M. de Almeida and A. G. Evsukoff, “Sistemas fuzzy,” in *Sistemas inteligentes: fundamentos e aplicações*, S. O. Rezende, Ed. Barueri: Manole, 2005, pp. 169–202.
- [2] W. Arbex, “Computational models for the identification of genomic information associated to the resistance to cattle tick,” Systems Engineering and Computer Science Program, PhD thesis, Federal University of Rio de Janeiro, Rio de Janeiro, 2009.
- [3] A. J. Brookes, “The essence of SNPs,” *Gene*, vol. 2, no. 234, pp. 177–186, 1999. DOI: 10.1016/S0378-1119(99)00219-X.
- [4] T. Brown, *Genomes*, 2nd ed. New York: John Wiley & Sons, 2002.
- [5] D. A. Nickerson, S. L. Taylor, N. Kolker, J. Sloan, T. Bhangale, M. Stephens, and I. Robertson, *Polyphred users manual*, Version 6.15 Beta, University of Washington, Seattle, 2008.
- [6] L. A. Zadeh, “Outline of a new approach to the analysis of complex systems and decision processes,” *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-3, pp. 28–44, 1973. DOI: 10.1109/TSMC.1973.5408575. [Online]. Available: <http://www-bisc.cs.berkeley.edu/Zadeh-1973.pdf>.
- [7] G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: theory and applications*. Upper Saddle River: Prentice Hall, 1995, p. 592, ISBN: 0131011715.
- [8] R. Tanscheit, “Sistemas fuzzy,” in *Inteligência computacional: aplicada à administração, economia e engenharia em Matlab*, H. A. e Oliveira Júnior, Ed. São Paulo: Thomson Learning, 2007, pp. 229–264.
- [9] P. Green, *Phrap*, 1 CD, C. Linux environment with C compiler., 1999. [Online]. Available: <http://www.phrap.org/index.html>.

- [10] D. A. Nickerson, V. O. Tobe, and S. L. Taylor, "PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing," *Nucl. Acids Res.*, vol. 25, no. 14, pp. 2745–2751, 1997. DOI: 10.1093/nar/25.14.2745. eprint: <http://nar.oxfordjournals.org/cgi/reprint/25/14/2745.pdf>.
- [11] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitzel, L. Hillier, P.-Y. Kwok, and W. R. Gish, "A general approach to single-nucleotide polymorphism discovery," *Nature Genetics*, vol. 23, pp. 452–456, 1999. DOI: 10.1038/70570.
- [12] W. Arbex, *fuzzyMorphic.pl*, 1 CD, Perl. UNIX-like environment with GUI and Perl 5.0 interpreter or newer., Juiz de Fora, 2009.
- [13] L. A. V. de Carvalho, *Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração*. Rio de Janeiro: Ciência Moderna, 2005.
- [14] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005, p. 525, ISBN: 0-12-088407-024884070.

Bioinformatic Analysis of Cyanobacterial Mercuric Resistance Related Genes and Identification of *Synechococcus* sp. IU 625 Putative Mercuric Resistance Genes

Lee H. Lee¹, Chiedozie Okafor¹, Matthew J. Rienzo², and Tin-Chun Chu²

¹Department of Biology & Molecular Biology, Montclair State University, Montclair, NJ, USA

²Department of Biological Sciences, Seton Hall University, South Orange, NJ, USA

Abstract - Due to high levels of heavy metal pollution in the environment, there has always been a high interest in organisms that have developed resistance to heavy metals. Extensive work has been done with respect to mechanisms of resistance to heavy metals in a wide array of microorganisms. However, mechanisms of resistance are yet to be explored in some microorganism such as cyanobacteria *Synechococcus* sp. IU 625. This microorganism has been known to show levels of resistance to Cu^{2+} , Hg^{2+} , and Zn^{2+} . Understanding the mercuric resistance mechanism in this microorganism would enhance the development of bioremediation systems for toxic waste cleanup. In this study, genomic and proteomics analysis of currently identified mercuric resistance genes in prokaryotes were carried out to determine their relationship to putative mercuric resistant genes in *S. IU 625*. Primers for genes encoding putative mercuric resistance were designed, amplified and attempted identified those genes from *S. IU 625*.

Keywords: Cyanobacteria, *Synechococcus* sp. IU 625, mercury resistance, heavy metal

1 Introduction

Cyanobacteria are aquatic photosynthetic microorganisms [1]. They are the oldest known fossils; being in existence for more than 3.5 billion years. They are however still part of the environment today and are one of the largest and most important groups of bacteria [2]. They have been attributed to being responsible for the oxygen rich atmosphere that most life forms on earth depend on today [2]. Cyanobacteria occur in an enormous diversity of habitats, freshwater and marine, as plankton, mats, and periphyton [1]. They have many beneficial functions such as nitrogen fixation and cycling of nutrients in the food chain. Despite their beneficial roles, cyanobacteria are the major causing agent for blooms [3]. Blooms are associated with eutrophic water, especially with levels of total phosphorus > 0.01 mg/L and levels of ammonia- or nitrate-nitrogen > 0.1 mg/L.

Optimal temperatures for blooms are 15-30 °C, and optimal pH is 6-9.

Synechococcus sp. IU 625 (*S. IU 625*), formerly *Anacystis nidulans*, is a non-motile, unicellular, rod-shaped organism, which is similar to gram-negative bacteria in cell wall structure, replication and ability to harbor plasmids [4]. *S. IU 625* is an obligate photoautotroph whose photosynthetic apparatus is similar to that of plants. Due to its potential generation of algal blooms it serves as a good indicator of environmental pollution. *S. IU 625* has been used in numerous studies to assess the effects of heavy metals as environmental pollutants [5-9].

Mercury (Hg) is a heavy metal that transitions between several forms, most of which produce both chronic and/or acute toxic effects [10]. These forms include its zero oxidation state Hg^0 , which exists as a vapor or liquid metal, its mercurous state Hg^+ , which exists as inorganic salts, and finally its mercuric state Hg^{2+} which is able to form either inorganic salts or organomercury compounds [10]. Mercury toxicity is caused by exposure to either ionic mercury or one of its compounds. The effects of mercury toxicity include damage to neurological, gastrointestinal, and renal organs. Results of mercury poisoning include several diseases such as Acrodynia (pink disease), Hunter-Russell syndrome, and Minamata disease [10]. Many studies have reported on bacteria resistance to mercury compounds. Enzymatic reduction of Hg^{2+} is encoded by genes of the *mer* operon [11]. Bioinformatic analysis of the sequences of many of these operons cloned from a diverse range of bacterial species reveals considerable similarity of genetic organization from both Gram-positive and Gram-negative microorganisms. Most of the operons contain a regulatory gene, *merR* at one terminus, which is divergently transcribed from the structural genes from a *mer* O/P region (except in the cases of some Gram-positive organisms) [11]. Proximal to the *mer* O/P region are genes which encode transport functions: most operons possess *merT* and *merP*, and in other operons *merC* and an open reading frame, which has been suggested to encode a transport function due to its homology to *merT* [12].

Research has shown that MerP retrieves mercuric ions from the extracellular environment and passes it along to MerT, which then passes it to mercuric reductase encoded by *merA*; which lies further downstream from the genes encoding transport functions. *merD*, encodes the presumptive down regulatory protein MerD [12]. And finally, in operons conferring resistance to both organic and inorganic mercury, *merA* and *merD* are separated by *merB*, encoding organomercurial lyase. The organomercurial lyase cleaves CH₃-Hg bonds, which leaves the Hg²⁺ ion able to be detoxified by mercuric reductase [12].

In this study, genomic and proteomic analysis of all currently identified mercuric resistance genes in *Synechococcus elongatus* PCC 6301, a very closely related strain, was performed and their relationship to putative mercuric resistance genes of *S. IU 625*, as well as other species of microorganisms was determined. In addition mercuric resistance genes, related genes and/or operon in *S. IU 625* were analyzed by using the primers designed from bioinformatics data. These primers were then used for PCR-based assay to identify and sequence these genes.

2 Materials and Methods

2.1 Cyanobacteria strain, media, and growth conditions

S. IU 625 stock cultures were obtained from the American Type Culture Collection, Manassas, VA (ATCC No. 27344). The cultures were grown and maintained in Erlenmeyer flasks containing 100 mL of sterilized Mauro's Modified Medium (3M medium) [13]. Flasks with cells were maintained in an Innova™ 4340 incubator (New Brunswick Scientific, Edison, NJ) at a constant temperature of 30°C with constant fluorescent light and continuous agitation at 100 rpm.

2.2 Genomic and proteomic analysis

Bioinformatic analysis was performed using BLAST (Basic Local Alignment Search Tool), BLAST Tree View, PSIPRED (Protein Structure Prediction Server) and, MEMSAT3/MEMSAT-SVM (Membrane Protein Structure and Topology Prediction).

2.3 Primer design

Primers were designed using the PrimerQuestSM oligo design software from Integrated DNA Technologies, Inc. All primers were optimized to have melting temperatures between 55°C and 60°C. GC content of all primers was also optimized to fall between 47% and 50%. Finally primers were designed for amplification of both intergenic and intragenic regions of genes that were of interest. All primers were designed based

on the published genomic sequences of *S. elongatus* PCC 6301 (Genbank Accession# NC_006576) [14].

2.4 Polymerase chain reaction, gel electrophoresis and DNA sequencing

Extracted DNA was subjected to polymerase chain reaction in order to determine the presence of putative mercuric resistance genes. The designed forward and reversed primers were used in PCR-based assay using a pre-heated Labnet MultiGene II thermal cycler (Labnet International, Edison, NJ). The reaction profile typically used was: initial denaturation at 95°C for 15 minutes followed by 35 cycles of denaturation at 95°C for 1 minute, primer annealing at 56-64°C for 1 minute, and extension at 72°C for 10 minutes. At the end of the 35 cycles, the reaction tubes were subjected to a final extension at 72°C for 5 minutes. The size of PCR products was estimated by 1% agarose gels in TAE buffer. The gels were analyzed under UV light using a Kodak Image Station 440CF (Perkin Elmer Life Sciences, Waltham, MA). The sequences of the amplicons were obtained using 3130 Genetic Analyzer sequencer (Applied Biosystems, Carlsbad, CA). The homologues searches of the obtained sequences were using NCBI Blast searches.

3 Results

In order to determine the presence of the mer operon and/or individual mercuric resistance genes on the genome of *S. IU 625*, the reference strain of *S. elongatus* PCC 6301 was used for the *in silico* analysis. Strain 6301 was chosen due to its close phylogeny with *S. IU 625*, its completely sequenced and highly annotated genome. The search for putative mercuric resistance genes on strain 6301's genome was done based on published genomic sequences of generic mer operons. Identification of transport (*merT*, *merP*), detoxification (*merA*), and regulation (*merR* and *merD*) genes was carried out. The proximity of the above mentioned genes from each other was also determined. In order to assess final protein structure of identified genes, most of the comparisons were also performed with the amino acid sequences.

3.1 *In silico* analysis of the mercuric reductase (*merA*) gene on genome of *S. elongatus* PCC 6301

The search for the mercuric reductase gene in strain 6301 revealed that the gene has previously been identified and assigned the Genbank accession number YP_171141.1 [14]. It was revealed that the presence of two genes flanking this mercuric reductase gene. These genes have no homology to any of the identified genes on the generic mer operon of other prokaryotes. The first gene, *syc0430_d*, located 30 base pairs upstream from the mercuric reductase gene, encodes a hypothetical protein, which is yet to be identified. In addition,

this gene is transcribed in the same direction as the mercuric reductase gene. The second gene, *syc0432_c*, also unidentified, is located 21 base pairs downstream from the mercuric reductase gene and is transcribed in the opposite direction. This gene also has no homology to any of the identified genes on the generic mer operon of prokaryotes (Fig.1).

Further bioinformatics analysis of *syc0430_d*, *merA*, and *syc0432_c* entire sequence on the structure of this mini

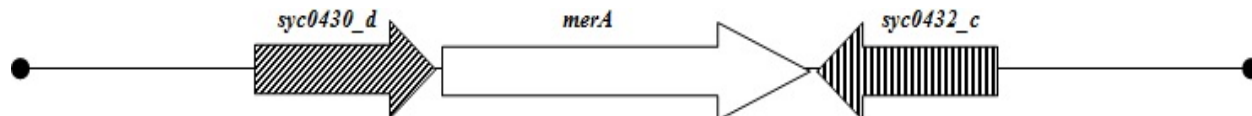


Figure 1. Mercuric reductase gene of *S. elongatus* PCC 6301 and two genes (*syc0430_d*/*syc0432_c*) upstream and downstream, respectively. Visual representation of putative operon: *syc0430_d* and *merA* get transcribed unidirectionally, while *syc0432_c* gets transcribed in the opposite direction.

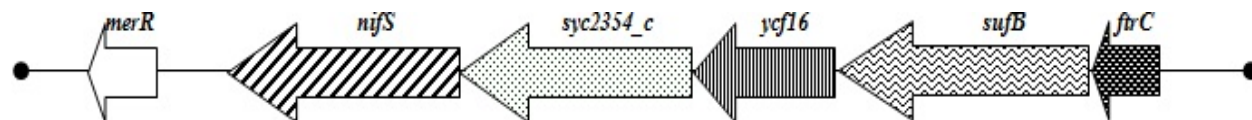


Figure 2. Identified *merR* gene in *S. elongatus* PCC 6301 along with five genes located 297 base pairs downstream. Visual representation of five genes within proximity of *merR*. All genes are differentiated by arrow fill. Note the distance of *merR* from other cluster of genes.

3.2 *In silico* analysis of the mer operon regulatory gene (*merR*) on genome of *S. elongatus* PCC 6301

Compared with the amino acid sequence, analysis of the *merR* nucleotide sequence of strain 6301 shows it only possesses significant sequence homology to the *merR* sequence of strain 7942. However, further analysis of the amino acid sequence shows that it belongs to the helix-turn-helix superfamily of *merR* transcription regulators, and bears homology to the MerR amino acid sequence of other cyanobacterial species.

As previously mentioned, the *merR* gene of other mercuric resistance operons are usually located a few base pairs upstream from the group of other mercuric resistance genes, which collectively make up the mer operon. However, in the case of *S. elongatus* PCC 6301, the identified *merR* gene, which is assigned the Genbank accession number YP_173062.1, is located about 2 million base pairs downstream of *merA* and its flanking genes. More intriguing is that 297 base pairs downstream from *merR* of strain 6301 is a group of genes that are evidently involved in the biosynthesis of iron-sulfur clusters (Fe-S) (Fig. 2). Fe-S clusters are involved in a wide range of functions such as

operon shows that flanking of *merA* by these two genes of unknown function is only conserved among certain other species of cyanobacteria (Fig. 2). Flanking by *syc0430_d* is seen conserved in cyanobacteria such as *Microcystis aeruginosa* NIES-843, *Nostoc punctiforme* PCC 73102, *Microcystis aeruginosa* PCC 7806, and *Anabaena variabilis* ATCC 29413. Flanking by *syc0432_c* is only observed in *Synechococcus elongatus* PCC 6301 and *Synechococcus elongatus* PCC 7942.

controlling protein structure, acting as environmental sensors, serving as modulators of gene regulation, and participating in radical generation. However, their direct involvement of Fe-S clusters in mercuric resistance is yet to be fully researched. It should also be noted that unlike the mer operon of most prokaryotes, the direction of the transcription of *merR* is in the same direction as the transcription of the Fe-S genes located a few base pairs upstream.

The above mentioned genes located 297 base pairs downstream from *merR* are *nifS*, *syc2354_c*, *ycf16*, *sufB*, and *ftrC*. An *in silico* analysis of the genes reveals their close relationship to Fe-S assembly genes of other microorganisms. However, their direct relationship of each individual gene to mercury detoxification is unclear. Bioinformatic analysis of the whole nucleotide sequence BLAST results of *merR* together with the cluster of five genes reveal that the highest degree of conservation is seen from about sequence 2509998 to 2512246. This homology falls within the nucleotide sequence coding for *ycf16* and *sufB*, an ABC transporter ATP-binding protein and a cysteine desulfurase activator complex subunit, respectively. Also revealing a high degree of homology are the sequences that fall between 2507378 and 2508367. Located within this nucleotide sequence is the gene *nifS* which codes for cysteine desulfurase, an enzyme involved in the formation of Fe-S clusters.

3.3 Probing for presence of putative mercuric resistance genes on genome of *S. IU 625*

3.3.1 PCR amplification and gel electrophoresis of putative genes

Upon completion of the bioinformatic analysis of the several putative mercuric resistance genes of strain 6301, we proceeded to probe the genome of strain 625 for the presence of these genes. Specific PCR primers were designed for use

in amplification of distinct regions of each of the above mentioned genes. All the primer sequences used for PCR analysis were designed based on the genome of *S. elongatus* PCC 6301. Gel electrophoresis results for *merA*, *ctaA*, *pacS*, *syc0430_d*, and *syc0432_c* showed that they fell within the range of the expected amplicon sizes based on designed primers (Fig. 3). In addition, gel electrophoresis results for PCR reactions run on *merR*, *nifS*, *sufD*, *ycf16*, and *frtC* also showed that their sizes fell within the expected ranges (Fig. 4). All these primers were able to prime *S. IU 625* DNA.

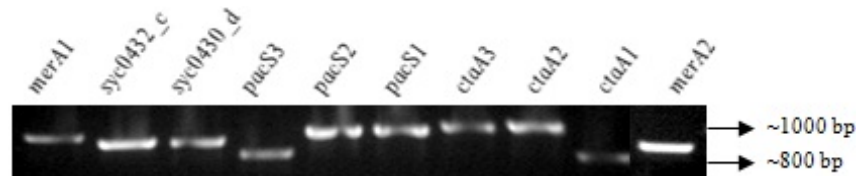


Figure 3. Priming for *merA*, *syc0430_d*, *syc0432_c*, *ctaA*, and *pacS* with *S. IU 625* DNA. Gel electrophoresis results of PCR reaction run on all five genes. Results show positive priming for all genes. Genes fall within expected range according to 1kb DNA ladder. Two sets of primers were designed for *merA*, and three sets each were designed for *ctaA* and *pacS*.

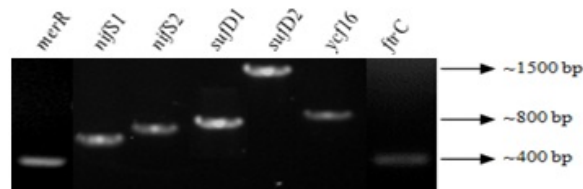


Figure 4. Priming for *merR*, *nifS*, *sufD*, *ycf16* and *frtC* of *S. elongatus* PCC 6301 with *S. IU 625* DNA. Gel electrophoresis results of PCR reaction for all seven genes. All amplicon sizes fall within their expected ranges.

3.3.2 Sequencing of identified genes

After the presence of the genes was confirmed through gel electrophoresis, the PCR products from each identified gene were sequenced. Results revealed that sequence results of all identified genes were at least 95% homologous to strain 6301. In addition, close analysis of the nucleotides of the overlapping regions of genes *syc0430_d*, *merA*, and *syc0432_c*, showed that the sequences were identical to those of strain 6301. The same was observed for overlapping sequences of *merR*, *nifS*, *sufD*, *ycf16*, and *frtC*. Finally, the sequence results of the *ctaA* and *pacS* genes also show a high level of homology to the same genes of strain 6301.

Based on the genome of *S. elongatus* PCC 6301 several hypothetical genes, including mercuric reductase, and the mercuric reductase regulatory gene, were able to be identified. Phylogenetic analyses of *merA* genes in cyanobacteria are shown in Figure 5. Unsurprisingly, the closet relative of *S. elongatus* PCC 6301 is *S. elongatus* PCC 7942 regarding *merA*; followed by *Synechocystis* sp. PCC 6803 and *Microcystis aeruginosa* NIES 843. Mercuric reductase in *Nostoc punctiforme* ATCC 29133, *Nostoc* sp.

PCC 7120 and *Anabaena variabilis* ATCC 29413 are also very close related to both *S. elongatus* PCC 6301 and *S. elongatus* PCC 7942. Microcystis and Anabaena are toxin-releasing, bloom-causing cyanobacteria that are usually the causing agents for harmful algal blooms in freshwater bodies. Bioinformatic analysis of the identified genes seemed to indicate that mercuric reductase was localized to the plasma membrane of the microorganism. In addition, the arrangement of the genes also seemed to indicate that along with detoxification by mercuric reductase, *S. IU 625* might also utilize synthesis of its peptidoglycan layer in reducing the permeability of the mercuric ions, thereby reducing the amount of detoxification the cell needs to perform.

Unlike mer operon of other prokaryotes, the mercuric resistance regulatory gene was located about 2Mbps upstream from the location of the mercuric reductase gene. This left open the question as to how the mercuric reductase gene was being regulated. In addition, because of the close vicinity of iron-sulfur cluster genes and peptidoglycan biosynthesis genes to the mercuric resistance regulatory gene, it also led to the question of their involvement in mercuric resistance.

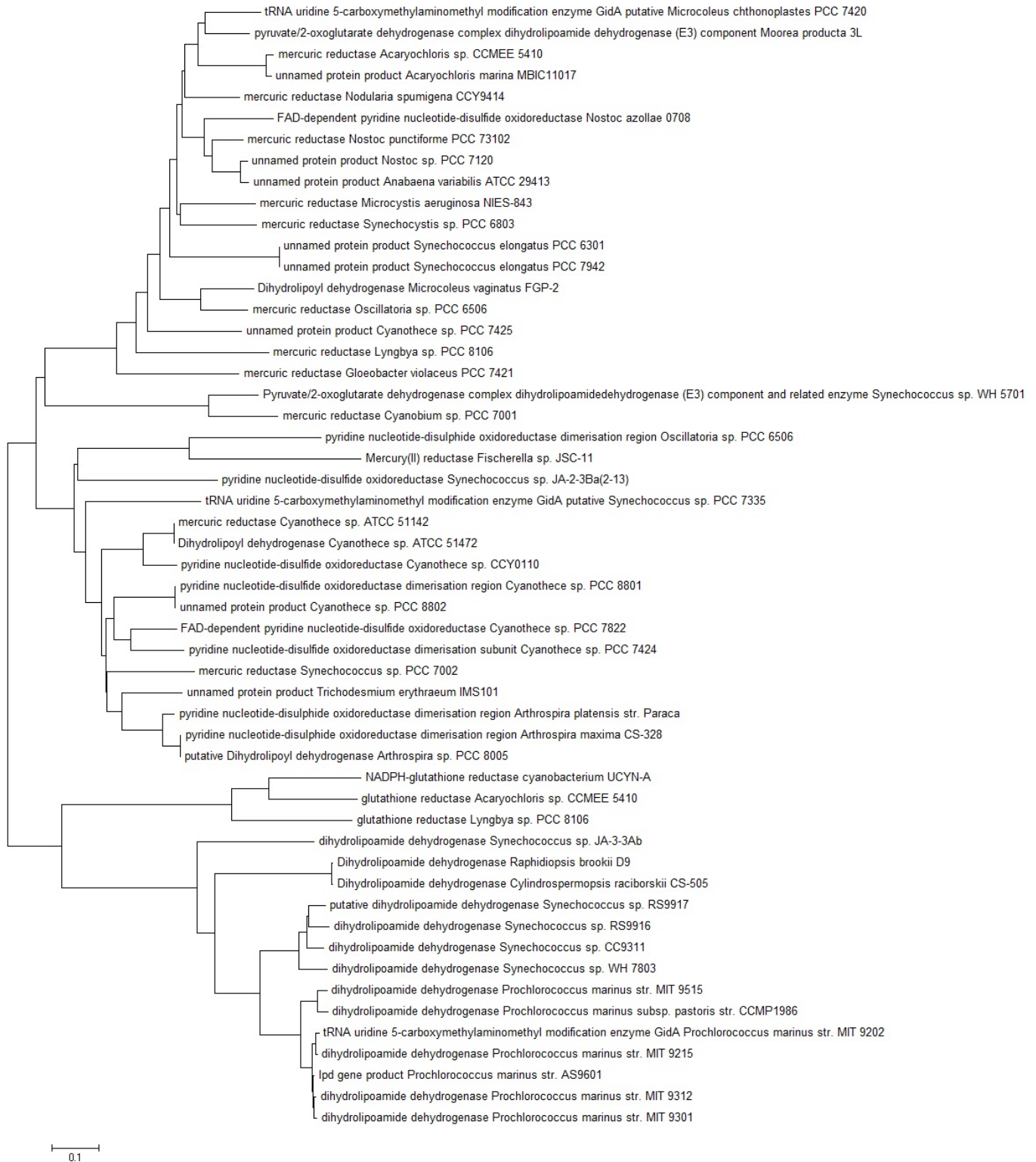


Figure 5. Phylogenetic analyses of *merA* gene in cyanobacteria.

4 References

- [1] Bernard Beall, and Joe Lutkenhaus. "Sequence analysis, transcriptional organization, and insertional mutagenesis of the *envA* gene of *Escherichia coli*"; J Bacteriol., 169(12): 5408-5415, Dec 1987.
- [2] Gail Fletcher, Carleen A. Irwin, Joan M. Henson, Cynthia Fillingim, Molly M. Malone, and James R. Walker. "Identification of the *Escherichia coli* cell division gene *sep* and organization of the cell division-cell envelope genes in the *sep-mur-ftsA-envA* cluster as determined with specialized transducing lambda bacteriophages"; J Bacteriol., 133(1): 91-100, Jan 1978.
- [3] Laura S. Busenlehner, Mario A. Pennella, and David P. Giedroc. "The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance"; FEMS Microbiol. Rev., 27(2-3): 131-143, Jun 2003.
- [4] Reginald H. Lau, Carmen Sapienza, and W. Ford Doolittle. "Cyanobacterial plasmids: their widespread occurrence, and the existence of regions of homology between plasmids in the same and different species"; Molec. gen. Genet. 178, 203-211. Apr 1980.
- [5] Tin-Chun Chu, Sean R. Murray, Jennifer Todd, Winder Perez, Jonathan R. Yarborough, Chiedozie Okafor, and Lee H. Lee. "Adaption of *Synechococcus* sp. IU 625 to growth in the presence of mercuric chloride"; Acta Histochem., 114(1): 6-11, Jan 2012.
- [6] Lee H. Lee and Bonnie Lustigman. "Effect of barium and nickel on the growth of *Anacystis nidulans*"; Bull. Environ. Contam. Toxicol., 56(6): 985-992, Jun 1996.
- [7] Lee H. Lee, Bonnie Lustigman, and Diane Dandorf. "Effect of manganese and zinc on the growth of *Anacystis nidulans*"; Bull. Environ. Contam. Toxicol., 53(1): 158-165, Jul 1994.
- [8] Lee H. Lee, Bonnie Lustigman, Sean Murray, and Stephen Koepp, "Effect of selenium on the growth of the cyanobacterium *Anacystis nidulans*"; Bull. Environ. Contam. Toxicol., 62(5): 591-599, May 1999.
- [9] Lee H. Lee, Bonnie K. Lustigman, and Sean R. Murray, "Combined effect of mercuric chloride and selenium dioxide on the growth of the cyanobacteria, *Anacystis nidulans*."; Bull. Environ. Contam. Toxicol., 69(6): 900-907, Dec 2002.
- [10] Tamar Barkay, Susan M. Miller, and Anne O. Summers. "Bacterial mercury resistance from atoms to ecosystems"; FEMS Microbiol. Rev., 27 (2-3): 355-384, Jun 2003.
- [11] Nigel L. Brown, Tapan K. Misra, Joseph N. Winnie, Annette Schmidt, Michael Seiff, and Simon Silver. "The nucleotide sequence of the mercuric resistance operons of plasmid R100 and transposon Tn501: further evidence for mer genes which enhance the activity of the mercuric ion detoxification system"; Mol. gen. Genet., 202(1): 143-151, Jan 1986.
- [12] Tapan K. Misra, Nigel L. Brown, David C. Fritzinger, R. David Pridmore, Wayne M. Barnes, Linda Haberstroh, and Simon Silver. "Mercuric ion-resistance operons of plasmid R100 and transposon Tn501: the beginning of the operon including the regulatory region and the first two structural genes"; Proc. Natl. Acad. Sci. U S A, 81(19): 5975-5979, Oct 1984.
- [13] William A. Kratz, and Jack Myers. "Photosynthesis and Respiration of Three Blue-Green Algae"; Plant Physiol., 30(3): 275-280. May 1955.
- [14] Chieko Sugita, Koretsugu Ogata, Masamitsu Shikata, Hiroyuki Jikuya, Jun Takano, Miho Furumichi, Minoru Kanehisa, Tatsuo Omata, Masahiro Sugiura and Mamoru Sugita. "Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization"; Photosynth Res., 93(1-3): 55-67, Jul-Sep 2007.

Accelerating the Smith-Waterman Algorithm for Bio-sequence Matching on GPU

Qianghua Zhu, Fei Xia, and Guoqing Jin

Electronic Engineering College, Naval University of Engineering, Wuhan, P. R. China, 430033

Abstract—Nowadays, GPU has emerged as one promising computing platform to accelerate bio-sequence analysis applications by exploiting all kinds of parallel optimization strategies. In this paper, we take a well-known algorithm in the field of pair-wise sequence alignment and database searching, the Smith-Waterman (S-W) algorithm as an example, and demonstrate approaches that fully exploit its performance potentials on GPU platform. We propose the combination of coalesced global memory accesses, shared memory tiles, and loop unfolding, achieving 50X speedups over initial S-W versions on a NVIDIA GeForce GTX 470 card. Experimental results also show that the GPU GTX 470 gains 12X speedups, instead of 100X reported by some studies, over Intel quad core CPU Q9400, under the same manufacturing technology and both with fully optimized schemes.

Keywords: Bio-sequence DB Searching, the Smith-Waterman Algorithm, Code Performance Tuning, GPU, Bioinformatics

1. Introduction

There is fierce market competition for high-performance computing platforms. General-purpose microprocessors, usually called central processing units (CPUs), such as X86 series from Intel and AMD, Power series from IBM, have entered the multi-core era, dominating the market of mainstream computing platforms. Hardware accelerators, particularly general-purpose graphics processing units (GPGPUs), are increasingly becoming important, especially in computation-intensive disciplines, such as realistic 3D computer graphics and high-performance scientific computing. Three supercomputers have entered the top 10 positions in the supercomputer top 500 list and are all hybrid designs. In contrast with the "heavy" core of CPUs equipped with large caches and a rich instruction set, GPUs have hundreds of "lean" processors with reduced instruction sets, small local memories, and in-order execution mechanisms.

Despite many reports on the superiority of GPU acceleration over CPU, there are still many open questions that cause confusion, including debates in some academic papers and website discussions [1, 2, 3]. In this paper, we take the sequence alignment application, the Smith-Waterman algorithm as an example, on GPU platform, to explore several optimization schemes, including coalesced global memory accesses, shared memory tiles, and loop

unrolling. The optimization schemes release the computing power of hundreds of GPU cores completely, taking the performance increase from 0.54~0.73 GCUPS (10^9 Cell Units Per Second) of initial version to 28.23 GCUPS, over 50X of speedups.

2. Background

In the area of modern molecular biology and bioinformatics, the S-W algorithm is a well-known algorithm for performing pair-wise local sequence alignment. It has become the kernel algorithm in the process of bio-sequence matching, multiple sequence alignment, and database searching to discover similarities between sequences and to further explore the evolutionary history, critical preserved motifs, and even the details of the tertiary structure and important clues about protein functions [4].

The algorithm was first developed by Temple F. Smith and Michael S. Waterman in 1981 [5] to determine the optimal local alignment of two sequences. The bio-sequences (i.e., DNA, RNA or protein) serve as the input of the S-W algorithm, and the output is an alignments score representing the degree of similarity of the two input sequences. In the alignment process, a two-dimensional matrix H , as a temporary data structure, is used to store alignment scores of subsequences. Consider two sequences S and L of length M and N . The two subsequences in sequences S and L are $S_1...S_i$ and $L_1...L_j$, respectively. The maximum similarity score of subsequence $S_1...S_i$ and $L_1...L_j$ is $F(i, j)$. The gap-penalty scheme provides the option of gaps being introduced within the alignments. In our implementation, we consider an affine gap penalty scheme that consists of two types of penalties, the gap-open penalty α and the gap-extension penalty β . The computation of $H(i, j)$ for grid cell (i, j) is given by the following recurrences:

$$\begin{cases} \text{for } 1 \leq i \leq M, 1 \leq j \leq N \\ H(i, 0) = E(i, 0) = \bar{H}(0, j) = F(0, j) = 0; & (1) \\ \text{for } 2 \leq i \leq M, 2 \leq j \leq N \\ \begin{cases} H(i, 0) = \max\{0, E(i, j), F(i, j), H(i-1, j-1) \\ \quad + sbt(S[i], L[j])\}; \\ E(i, j) = \max\{H(i, j-1) - \alpha, E(i, j-1) - \beta\}; \\ F(i, j) = \max\{H(i-1, j) - \alpha, E(i-1, j) - \beta\}; \end{cases} & (2) \end{cases}$$

sbt is the character substitution cost table. We make some observations on the characteristics of the S-W algorithm. These observations suggest details of parallel

implementation.

Observation 1: Inter-task parallelization

There is no data dependency in the multiple task of executing the alignment of several database sequences with a single query. Pair-wise alignment, called inter-task parallelization or coarse-grained parallelization, can be performed independently. The shared data are the query sequence and the substitution cost matrix. Multiple sequence alignment tasks can be distributed to GPU platforms to utilize computing resources efficiently. These implementations focus on database partitioning and load balance, instead of the parallelism of pair-wise sequence alignment.

Observation 2: Intra-task parallelization

Data parallelism also exists in pair-wise sequence alignment; it is called intra-task parallelization or fine-grained parallelization. Recurrence Formula 2 implied regular data dependency; that is, each cell $H(i, j)$ depends on its left neighbor $E(i, j-1)$, $H(i, j-1)$, upper neighbor $F(i-1, j)$, $H(i-1, j)$, and upper left neighbor $H(i-1, j-1)$ in filling matrix H . If we fill the score matrix in a row column order in turn, the process must be executed serially. Moreover, there is strict data synchronization among adjacent anti-diagonals. However, there is no data dependency among the elements located in each anti-diagonal. Therefore, all cells along the anti-diagonal k can be computed parallel from the anti-diagonals $k-2$, $k-1$, which can be arranged in a wave-front mode along the diagonal from up-left to down-right (Figure 1).

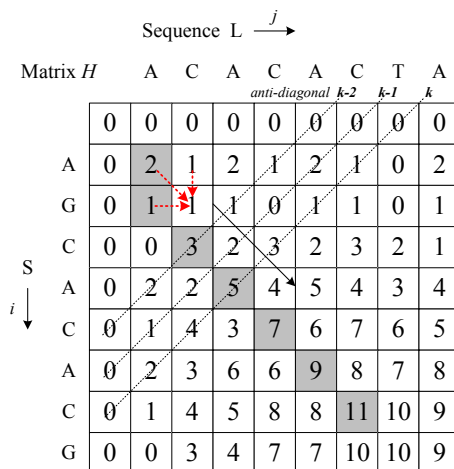


Fig. 1: Example of the S-W algorithm to compute the local alignment between two DNA sequences. The red arrows represent the data dependency in the S-W algorithm and the black arrow marked the computing order. The wave-front computation moves along diagonals from up-left to down-right.

The computational and spatial complexity of pair-wise sequence alignment for two sequences of length M and N is $O(M \times N)$, and the computational complexity of the multi sequence alignment is $O(K \times N^2)$ for K sequences with average length N . Bio-sequence databases are undergoing exponential growth; GenBank, for example, now stands with over 100 million sequences and 100 billion base pairs [6]. Hence, although comparing two sequences using the S-W algorithm is efficient in the classical sense, the execution time is still intolerable for pair-wise sequence alignment on the whole genome scale (more than 10^9 bases). Due to differences in manufacture technology, hardware structure, computing resource, and clock frequency across all kinds of platforms, we use the standard measurement unit, GCUPS (10^9 Cell Updates Per Second), to measure actual computing power. The cell represents the workload for computing one element of the score matrix.

3. Optimizations on GPU

Modern GPUs are designed as programmable processors employing a large number of processor cores. It contains a processor array, which consists of a number of streaming multiprocessors (SMs) and hierarchical memory architecture for programmers to utilize. GPUs are especially well-suited to address problems that can be expressed as data-parallel computations in which the same program is executed on many data elements in parallel, by mapping data elements to parallel processing threads. Thus, to achieve reasonable parallel efficiency for GPU parallel computing, memory optimization schemes have to be adopted carefully to utilize fully the three layers of memory hierarchies: register, shared memory, and global memory [7].

Considering the computation searching of the optimal local alignment between a query sequence and a subject sequence as a task, there are two approaches to the parallel processing of sequence database searches using CUDA. The first approach, as shown in Figure 2(a), is intra-task parallelization, which assigns one task to a grid. In the approach, all threads cooperate to perform the task in parallel by calculating the alignment scores of cells within the same diagonals.

The second approach is inter-task parallelization, which assigns one task to exactly one thread. T1 thread takes task 1 to calculate the optimal alignment score between query sequence and database sequence 1, T2 thread takes task 2, and so on (Figure 2(b)). The method of intra-task parallelization, reported by Liu [8], occupies less device memory space, but suffers from frequent barrier synchronizations between GPU cores. In the following optimizations, we only consider the inter-task parallelization scheme that occupies more device memory because of the need to store the intermediate alignment results but achieves better performance than intra-task parallelization, as presented in [9, 10] by Manavski and Ligowski.

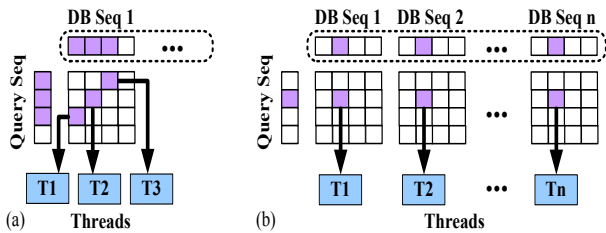


Fig. 2: GPU parallel strategies based on single-thread. (a) Intra-task parallelization. (b) Inter-task parallelization.

Based on inter-task parallelization, the following sections focus mainly on memory optimizations. The basic data structure of the S-W algorithm includes one two-dimension substitution cost table, one query sequence, and one sequence database consisting of a number of subject sequences. During the execution of the S-W algorithm, additional memory is required to store intermediate alignment results. To support a much larger database, the global memory is used to store the sequence database and the intermediate alignment results. The substitution cost table and query sequence are read-only, and are accessed by all threads in the grid. Therefore, we store the substitution cost table and query sequence into the texture memory and the constant memory, respectively.

3.1 Coalesced global memory accesses

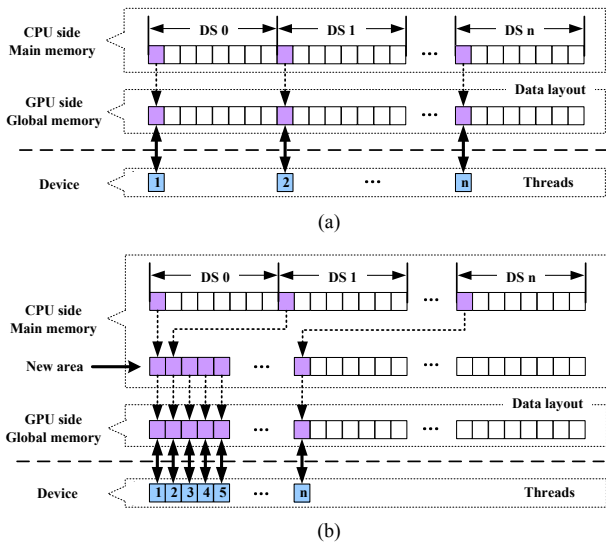


Fig. 3: Global memory layout and access pattern. (a) Naive implementation. (b) Interleaved implementation.

Coalesced global memory accesses can always have significant improvements in performance over non-coalesced global memory accesses due to the effective usage of global memory bandwidth [7]. However, in the initial GPU version of S-W algorithm, none of the memory accesses are coalesced (Figure 3(a)). First, the database sequences loaded

from a disk file are allocated a contiguous area of memory in the CPU side, where one sequence is represented by the data structure of a one-dimension character array, and the next sequence occupies the side contiguous area, as depicted in Figure 3(a) (DS 0, DS 1...DS n). Second, the database sequences area is copied directly to the GPU side global memory, where all symbols of the sequences are kept in order in the CPU side. All GPU threads begin to calculate the alignment scores from the first symbol, the second, and so on. The memory addresses generated from the same thread warp are non-contiguous. The column of non-optimization in Table 1 shows that no coalesced load or store operations occur, other than 181.12×10^8 and 161.69×10^8 times of scattered loads and stores respectively.

In the optimized version, we change the data layout of database sequences in the GPU global memory, creating coalesced memory accesses. Before the area of database sequences is copied to the GPU side global memory, we interleave the database sequences to a new area in CPU side memory. The symbols with the same index of all sequences are collected together. Thereafter, the new area is copied to the GPU side global memory (Figure 3(b)). The threads in the same warp access the global memory addresses in a contiguous way, resulting in the removal of all non-coalesced load/store. The rearranged data layout scheme helps significantly in achieving better performance. The GPU performance is increased from 0.54 to 5.79 GCUPS, over 10X of speedup.

Table 1: Performance improvement using coalesced global memory accesses.

		without opt.	with opt.
memory access	non-coalesced load	181.12	0
	non-coalesced store	161.69	0
number (10 ⁸)	coalesced load	0	11.32
	coalesced store	0	10.09
performance (GCUPS)	GTX 280	0.73	5.54
	GTX 470	0.54	5.79

3.2 Shared memory tiles to reduce global memory accesses

After coalescing the contiguous global memory accesses, we can estimate the total number of global memory accesses. Figure 4(a) shows the order in which one thread calculates the matrix, first in row, then in column order. This process, which involves scanning the matrix from left to right and then from top to bottom, requires $2 \times M \times N$ global memory transactions for the $M \times N$ matrix. Calculating $H(i, j)$ involves two global memory accesses, one for loading $H(i-1, j)$, $F(i-1, j)$, and the other for storing $H(i, j)$, $F(i, j)$.

We utilize tiles in the shared memory layer to reduce global memory accesses. The whole matrix computation is

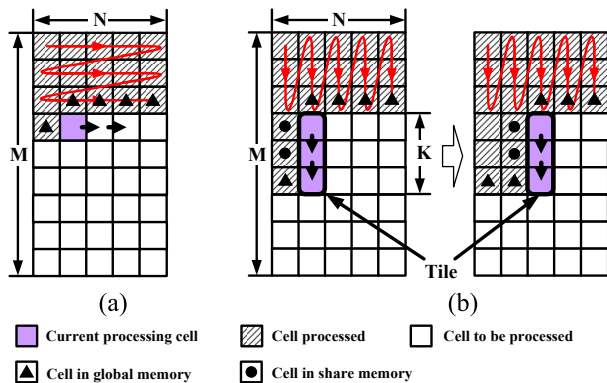


Fig. 4: Memory layout and access pattern. (a) Without tiling. (b) With tiling.

first divided into sub-matrices with $K \times N$ elements. Thereafter, the execution order for calculating the sub-matrices is changed, first in column, then in row order (Figure 4(b)). One column in the sub-matrix is called a "tile" (circled by black line in the figure). A tile consists of K vertically contiguous cells, where K is the tile size. The sub-matrix is processed horizontally tile by tile. The $K - 1$ elements of a tile are stored in the shared memory layer, which will be used in the calculation of the next tile. All operations within the tile are performed in the shared memory and registers, except for the loading of one element from the previous sub-matrix and the storing of one element to the global memory for the next sub-matrix calculation. Thus, for one sub-matrix calculation, only two rows of global memory accesses are performed, one for loading the top row and the other for storing the bottom row. The total global memory transactions are reduced to $2MN/K$.

Table 2: Performance improvement using coalesced shared memory tiles

Tile size	2	4	6	8	10
Num. of load (10^8)	5.66	2.83	1.91	1.42	1.13
Num. of store (10^8)	5.05	2.54	1.71	1.28	1.02
GTX280 (GCUPS)	7.95	8.96	9.35	—	—
GTX470 (GCUPS)	11.36	15.65	16.98	17.64	18.07

The increase in tile size decreases the number of global memory accesses to amortize the global latency cost, but the tile size is strictly limited by the size of the shared memory and the number of threads because each thread occupies a tile area in the shared memory. As shown in the Table 2, for GPU GTX 280, the largest tile size only reaches 6, with over 60% of performance gain, and for GTX 470, only two elements in one tile brings nearly 2X speedups, and the largest tile size can be 10 with over 3X speedups. The decrease in the total number of global memory accesses with the increase in tile size supports above performance gains.

3.3 Register-level optimization with loop unrolling

In the version of tiled GPU S-W algorithm, an additional innermost loop has to be introduced to organize the tile calculation. This loop has a small body and constant iteration count. The tiling scheme reduces the number of global memory accesses at the expense of additional shared memory accesses, branch instructions, and address calculations. When the threads within a warp diverge via data-dependent conditional branches, the warp has to execute each branch path serially, causing severely performance bottlenecks [7].

Table 3: Performance improvement using register optimization with loop unrolling.

	Tile size	2	4	8	10
GTX280	Register	19	25	—	—
	Branch(10^7)	8.41	4.21	—	—
	GCUPS	10.96	13.11	—	—
GTX470	Register	24	28	32	32
	Branch(10^7)	25.24	12.64	6.35	5.07
	GCUPS	11.49	18.76	25.15	28.23

The best performance can be achieved by unrolling the loop completely and removing all innermost loop branches, induction variable increments, and inner loop address calculation instructions. Experiment results show that the loop unrolling scheme greatly decreases branch instructions (Table 3). Finally, with the help of all optimization schemes, the most powerful performance on the GTX 470 platform reaches 28.23 GCUPS, over 50X speedups compared with the initial GPU version without any optimization schemes.

4. Result and Discussion

4.1 Environment and Test Methods

Our prototype system for performance evaluation consists of a host PC and a GPU card. The host is equipped with an Intel Q9400 Quad CPU, 2GB memory, and ASUS P5Q Dulex motherboard [P45 chipset running Windows XP SP3 with Visual Studio 2008 development environment (Visual C++ Compiler 15.00.30729.01)]. We use Redhat enterprise Linux 5.4 operating system with GCC 4.1.2 compiler for testing the Xeon CPU platform. We choose two commercial graphics cards, Geforce GTX280 and GTX470 with CUDA toolkit 3.1, as our GPU experimental platforms. The original S-W source code is derived from the kernel of the ClustalW [11], the famous program for multi-sequence alignment application.

Table 4 shows the experimental results of the S-W algorithm for protein database searching application on CPU and GPU platforms. For each kind of platform, we list the performance measurements of two chips with different manufacturing technologies.

Table 4: Results on CPU and GPU platforms with different manufacturing technologies.

	CPU		GPU	
Device Chip	Dual-Core E2140	Quad Q9400	GTX280	GTX470
Development Year	2007	2008	2008	2010
Chip Technology (nm)	65	45	65	40
PE(Core) Number/Chips	2	4	240	448
Frequency (MHz)	1600	2660	1296	1215
Peak Memory Bandwidth	3.2 GB/s	10.6 GB/s	141.7 GB/s	133.9 GB/s
Cache Capacity (KB)	1,024	6,400	728	792
GCUPS (average)	0.35	2.18	13.71	28.23

Table 5: Compare the average performance (GCUPS) with related works.

		Platforms	Approach	Performance
CPU	Alpern [12]	Intel Paragon i860	Single thread, SIMD	< 0.01
	Wozniak [13]	Sun Ultra SPARC 167MHz	Single thread, SIMD	0.02
	Rognes [14]	Intel Pentium III 500MHz	Single thread, SIMD	0.15
	Jacob [15]	Intel Pentium 4 2.8GHz	Single thread, SIMD	0.49
	Ours	Intel Q9400 Quad 2.66GHz	Multi-thread, SIMD	2.18
GPU	Manavski [9]	GeForce GTX 8800	Global memory optimization	1.75
	Ligowski [10]	GeForce 9800 GX2	Shared memory optimization	8.67
	Liu [8]	GeForce GTX 295	Global and shared memory optimization	9.51
	Ours	GeForce GTX 470	Global, shared memory and register opt.	28.23

4.2 Compared to CPU Implementation

4.2.1 GCUPS Performance Comparison

We implement the S-W algorithm with the affine gap penalty model for protein database search application on CPU and GPU platforms with 65 nm and 45 nm manufacturing technology, respectively. For each computing platform, we test the average performance (GCUPS) of the S-W algorithm with different optimization grades. As shown in Figure 5, the horizontal axis is the average performance represented by GCUPS, and the bar with different colors and letters represent different optimization grades. The right part with black oblique lines in each bar is the performance improvement on the new generation 45 nm manufacture technology as compared to 65 nm computing platforms using the same optimization scheme.

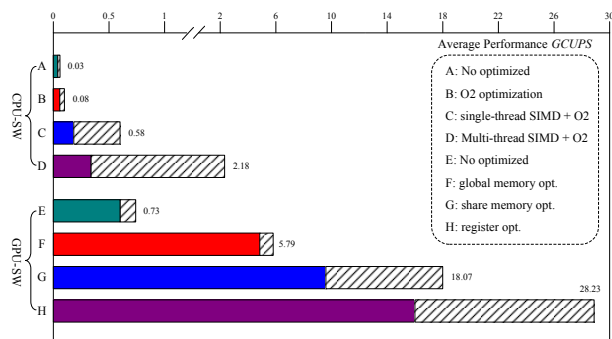


Fig. 5: Average performance growth of the S-W algorithm with different optimization grade on CPU and GPU platforms.

The naive S-W version running on the state-of-the-art multi-core CPU platform is only 0.03 GCUPS. The performance is improved steadily by adopting different optimization strategies, including compiler auto options, single-thread SIMD, and multi-thread SIMD. The performance reaches 2.18 GCUPS, over 70X speedups, using multi-thread SIMD on Intel Q9400 Quad CPU. On the GPU platform, the performance of the naive version on Geforce GTX 470 without any optimization is just 0.73 GCUPS, lower than that of the multi-thread SIMD implementation on the Q9400 Quad CPU. However, the final performance is increased by nearly 40 times, reaching 28.23 GCUPS, on condition of the highest optimization effort.

4.2.2 Misunderstandings on GPU and CPU comparison

We find three misunderstandings on performance comparison between GPU and CPU implementations.

(1) Optimized version GPU vs. naive version CPU.

From Figure 5, we observe that the GPU implementation with register optimization shows a factor of more than 900X speedup over the naive version running on CPU. However, compared to the optimized version with multi-thread SIMD with loop unrolling, the speedup factor is only by 12X.

(2) New manufacturing technology GPU vs. old manufacturing technology CPU.

If we compare the performance of the S-W algorithm tested on GPUs with 45 nm manufacture technology to that on 65 nm dual-core CPU, a factor of more than 80X speedup can be achieved. However, the performance on the GPU platform is only improved by 12X compared to the CPU with the same 45 nm technology.

(3) Optimized version on new GPU platform vs. naive version on old CPU platform.

The most unfair comparison is that of the performance of the optimized version on the new generation GPU platform to the naive version running on the old CPU. If we adopt this measure approach, the GPU implementation would show a speedup factor of more than 1000X over the origin CPU version without optimization. But this is beyond scientific evaluation.

Conclusively, the performance of the GPU is superior to the CPU version. Instead of hundreds of times speedup, GPU shows a speedup factor of 12X over the CPU when both are running the optimized version under the same manufacture technology.

5. Comparison with Relatedworks

There are a number of efficient implementations of the S-W algorithm on GPU platforms, as listed in Table 5. Most work on GPU acceleration discussed intra-task parallelization or global memory optimization schemes separately, and none compared GPU with a fully optimized CPU version. Our work combines three levels of optimizations and reports fair comparison results. Recently, there have been several papers evaluating the performance of CPU and GPU computing platforms. Both [1] from Intel Lee and [2] from IBM reported performance comparisons between carefully tuned CPU version with optimized GPU version collected from published papers. However, the comparisons were taken from 45 nm CPUs to 65/55 nm GPUs.

6. Conclusion

This paper explored the parallel schemes on GPU platforms to accelerate the S-W algorithm for pair-wise sequence alignment. We tried various optimization schemes, including coalesced global memory accesses, shared memory tiles, and loop unfolding, and obtained over 50X speedups. The experimental results show that GPU is obviously superior to CPU. However, the performance difference does not reach 100X, only 12X on the condition of fair Comparative Study.

Acknowledgments.

We would like to thank the anonymous reviewers for their detailed revising directions and constructive comments. This work is partially sponsored by the NSFC (Hardware Acceleration and Parallelism Research for Complex Biological Sequence Analysis on Heterogeneous Computing Platform).

References

[1] Victor W. Lee, Changkyu Kim, et al, "Debunking the 100X GPU vs. CPU Myth: An Evaluation of Throughput Computing on CPU and GPU," in *Proc. International Symposium on Computer Architecture ISCA'10*, 2010, pp. 451–460.

[2] Rajesh Bordawekar, Uday Bondhugula, and Ravi Rao, "Believe It or Not! Multi-core CPUs Can Match GPU Performance for FLOP-intensive Application," IBM Thomas J. Watson Research Center, Technical Report RC24982, Apr. 2010.

[3] Andy Keane. (2010) "GPUs are only up to 14 times faster than CPUs" says Intel. [Online]. Available: <http://blogs.nvidia.com/intersect/2010/06/gpus-are-only-up-to-14-times-faster-than-cpus-says-intel.html>

[4] G.L. Moritz, C. Jory, H.S. Lopes, C.R.E. Lima, "Implementation of a Parallel Algorithm for Protein Pairwise Alignment Using Reconfigurable Computing," in *Proc. IEEE International Conference on Reconfigurable Computing and FPGA's ReConFig'06*, 2006, pp. 1–7.

[5] Smith, Temple F. and Waterman, Michael S., "Identification of Common Molecular Subsequences," *IEEE J. Mol. Biol.*, vol. 147, pp. 195–197, 1981.

[6] NCBI. (2009) GenBank Growth Statistics. [Online]. Available: <http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html>

[7] NVIDIA Corporation. (2010) NVIDIA CUDA Best Practices Guide 3.1. [Online]. Available: <http://developer.nvidia.com/cuda/>

[8] Y. Liu, D. Maskell, B. Schmidt., "CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units," *BMC Research Notes*, vol.2(1), p. 73, 1981.

[9] S. A. Manavski and G. Valle., "CUDA-compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment," *Bioinformatics*, vol.9(2), pp. 10–19, 2008.

[10] L. Ligowski, W. Rudnicki, "An efficient implementation of smith waterman algorithm on GPU using cuda for massively parallel scanning of sequence databases," in *Proc. ACM SIGPLAN symposium on Principles and practice of parallel programming*, 2010, pp. 137–146.

[11] J. Thompson, D. Higgins, T. Gibson, "Clustalw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position specific gap penalties and weight matrix choice," in *Nucleic Acids Res.*, vol.22, pp. 4673–4680, 1994.

[12] Bowen Alpern, Larry Carter, and Kang Su Gatlin, "Microparallelism and high performance protein matching," in *Proc. ACM/IEEE Supercomputing Conference SC'95*, 1995, p. 24.

[13] A. Wozniak, "Using video-oriented instructions to speed up sequence comparison," *Comput. Appl. Biosci.*, vol.13(2), pp. 145–150, 1997.

[14] T. Rognes, E. Seeberg, "Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors," *Bioinformatics*, vol.16(8), pp. 699–706, 2000.

[15] Arpith Jacob, Marcin Paprzycki, et al, "Applying SIMD approach to whole genome comparison on commodity hardware," in *Proc. International conference on Parallel processing and applied mathematics*, 2007, pp. 1220–1229.

Aligning Highly Variable DNA Sequences Using the W-curve and SQL

Steven Lembark¹, Shadi Beidas², Douglas Cork², Workhorse Computing¹, St. Louis, MO, 631101, Illinois Institute of Technology², Chicago, IL. 60616

Abstract

Unidimensional character strings are practical because of their simplicity. This simplicity made it possible to create the initial generation of tools for analyzing DNA. However there are some cases that these tools do not handle gracefully. DNA found in non-correcting viruses or oncology and “cloud” computing pose significant barriers to pushing these tools forward. The alternative we show here uses an abstract geometric representation of the DNA sequence called the W-curve. Geometric properties of these curves offer new avenues that bypass the roadblocks inherent in string-based approaches. Our approach described here uses a database with geometric fields and spatial indexes originally developed for geo-coding. The techniques improve handling of crossover-recombinant sequences and are suitable for distributed computing.

1. Background: Comparing DNA Character Strings.

The tools most commonly used today for analyzing DNA sequences are based on character-string representation of the DNA sequence. Representing the sequences in this manner makes intuitive sense and works for the most common cases. Analysis with these tools assumes largely similar sequences and that any differences between the sequences are significant. This approach works in most cases: for example, humans and chimpanzees have nearly 96% of genetic material in common [1], and nearly 60% of human genes are common to that of the fruit fly, *Drosophila melanogaster* [2].

The assumption of similarity starts to break down in studies of non-correcting RNA viruses or cancerous cells. The best-known example is HIV-1, the group also includes the filoviruses Marburgvirus and Ebolavirus and sequences found in cells damaged by radiation or botched mitoses. The process of studying these sequences today starts with a shotgun sequence from Next Generation Sequencing (“NGS”) machines [3]. The following step is to align the small fragments output by NGS with template sequences. The high

variance makes the fragments difficult to align, often requiring longer sequences for success. [4].

Crossover recombination is a common problem with HIV-1. The virus packages itself with two strands of RNA per viron. This leads to a fairly high rate of crossovers between the adjacent genomes in the viral progeny. Combined high rates of mutation and re-infection leave many patients with multiple distinct strains of HIV-1 infecting the same cells [5]. HIV-1's propensity for recombination makes hybrid strains relatively common compared to other viral infections. Similar problems arise in oncology studies where crossovers may be the cause of cancer: there is no good way to compare fragments to multiple template sequences at once utilizing the current generation of string-based techniques.

Compounding the problem is the scoring mechanism used by existing software, which produces a single value for the entire match. These tools offer no mechanism for comparing fragments piecewise and choosing the most relevant matches for each section.

BLAST, FASTA, and ClustalX2 all perform their comparisons recursively [6]. This works well enough for singly-threaded applications but is difficult to run in parallel. The growth of distributed computing environments makes it important to start looking for algorithms that are adaptable to parallel and highly-distributed execution. This requires an algorithm suitable for a divide-and-conquer approach, with the ability to compare regions separately and combine the results for final analysis.

2. Alternative: The W-curve

The W-curve was originally developed as a visualization tool for comparing large sequences of DNA. It uses a state machine to generate three-dimensional output based on the DNA sequence. The curve has a few useful properties for comparing sequences in the presence of SNPs and gaps, and its geometric result has more detail at the fine scale than a sequence of characters. Most important the location of points on the W-curve are influenced by the prior points on the

curve. This produces a more detailed structure which can be queried independently at each point on the curve [7].

Previous papers have described generating a W-curve in detail [8]. The W-curve is produced by a state machine using four corners of a square with corners labeled for the four bases in DNA (Fig. 2a). The X and Y axis are unit-less, the Z-axis is discrete, matching the sequence's base numbers. At each point on the curve, the next point is determined by going halfway to the corner for the next base in X-Y (Fig. 2b). All

curves begin at the origin, so the first point on all curves is at a distance of $\frac{1}{2}$ along an axis with a Z value of 1.

Two important properties of the W-curve are shown in Figure 2. One is that altering one base in the sequence will change the locations of successive points in the curve. This is an important difference with character-based algorithms: Each point in the W-curve depends to a certain extent on the sequence of prior points. There is no analogous relationship between the bases in a character string. For example, knowing

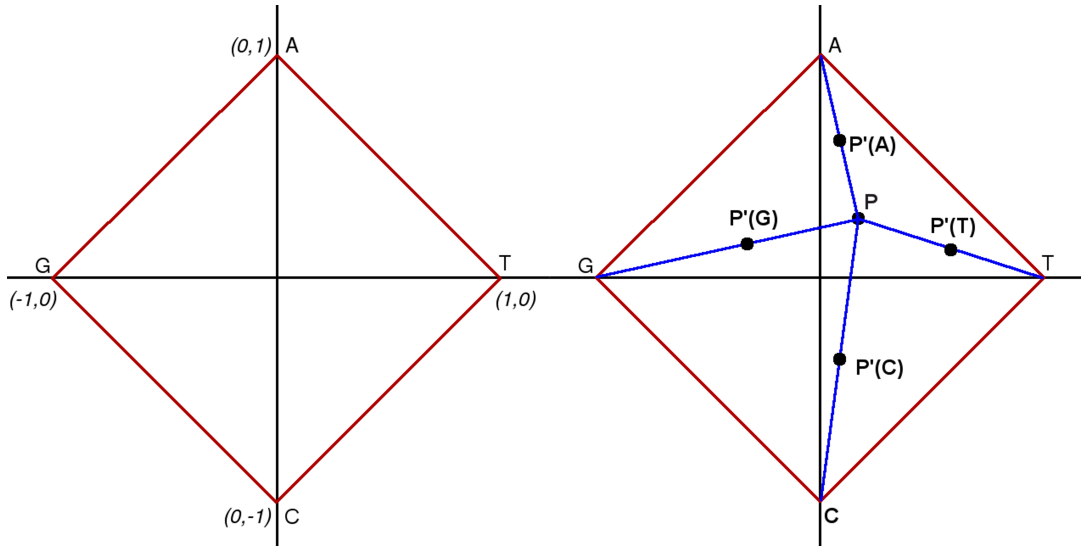


Figure 1: (a) Each corner of a square is labeled with one DNA base. (b) Successive points are halfway from the current point to the corner labeled with the next base. Shown are the next points from P for each possible next base (P'(A), etc).

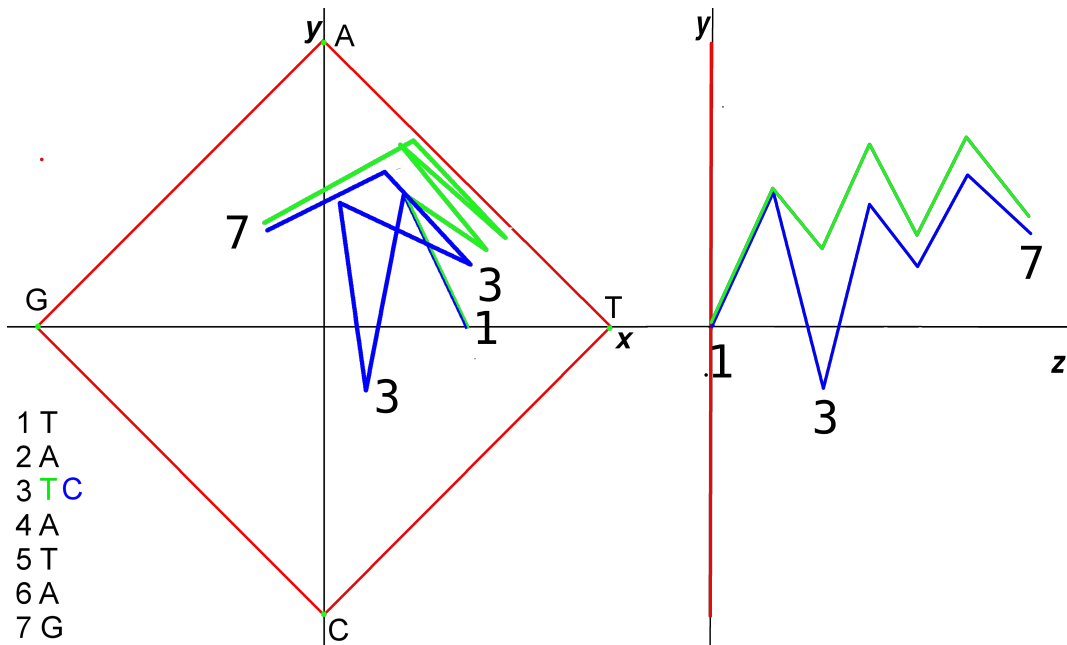


Figure 2: W-curve divergence and convergence after a SNP at base number 3 (T vs. C). The curves diverge noticeably at base 3 but have largely converged by base seven. This combination of divergence with auto-regression makes the curves useful for comparing sequences: local differences are detectable after which the curves converge due to auto-regression.

the x-y location of a point on a W-curve tells us something about the previous points. If nothing else, we know that if any of the last few points were different then the point would not be where we found it. This contrasts with a string-based sequence of characters: replacing any character in the string has no affect on the ones before or after it.

Another important property is that after a few common bases the curves rejoin one another. This property, called “auto-regression”, is how the W-curve handles SNPs and gaps between the sequences. The effect can be seen in bases 4-7 of the curves in figure 2: the curves diverge noticeably at a SNP in base 3 but are nearly re-aligned by base 7. A similar effect is seen with gaps: within a few bases after a gap the curves converge with a phase-shift equal to the gap size.

The balance of local divergence and auto-regression makes it possible to align larger sequences while finding the differences between them. Auto-regression permits piecewise comparison of the curves since the alignment of any number of fragments will match their alignment taken as a whole. Regardless of local divergences from SNPs or gaps, common sequences will still align.

The following section will illustrate how the curves can be stored and compared using geometric extensions for relational databases.

3. Querying a Curve

Recent developments in relational database technology have added queryable geometric objects to the relational vocabulary [9]. We are using Postgres 9.1 with GiST objects (a.k.a. “postgis”). The geometric fields were originally

designed for geographic or astronomical queries: find the roads in a city, or locations of postal codes. The constructs include points, lines, polygons, and circles which can be queried for overlap, intersection, inclusion or distance.

These database extensions also include “spatial indexes” which define bounding boxes for the geometric elements in the indexed fields. The indexes greatly improve performance in intersection or “contained within” queries.

There are any number of ways to apply these database extensions to model and query a W-curve. Our initial approach was similar to ones used with string-based approaches: began by selecting the template points close to the fragment's first base in X-Y. Then looking for points close in X-Y to the next point. However, this approach has problems with SNPs or gaps leaving no points to select in the next iteration or an initial SNP filtering out the correct templates.

One workaround for internal SNPs and gaps is to keep searching with an expanding window until one point is selected, then continue from that point forward. This approach handles internal SNPs or gaps does not account for crossovers or mismatches at the start of a curve. The problem with recombinant fragments is that they stop matching on one template at its midpoint, leaving us with nothing to select going forward. This approach is also not suitable for distributed computing since the process of acquiring each base depends on the previous one selected.

Some limitations of selecting incremental bases can be worked around by proceeding from both ends of the fragment

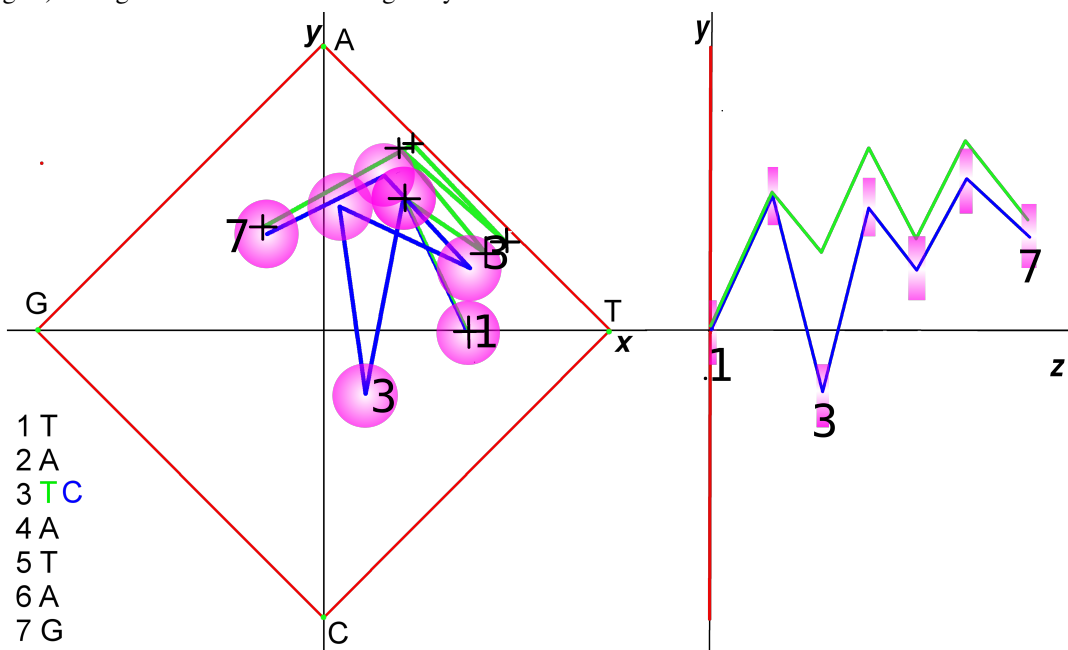


Figure 3: Overlap of fixed (green) point vertexes with fragment (blue) vertexes as circles using a radius of 0.10. After the curves diverge at base 3, the template's points intersect the fragment circles at base 7. Adjusting the circle size allows for fuzzy matching and helps compensate for progressive rounding error, variations caused by the fragment starting at (0,0), or multi-base/ low-quality FastQ entries.

at once: selecting any template points that match either end of the fragment and working back towards the fragment's center. This can handle crossovers but still leaves curves orphaned due to a SNP or gap at the ends of a fragment and is still not really suitable for distributed computation.

Avoiding issues with the endpoints requires dealing with points in the middle. Querying all of the points, will locate all the candidate templates in one pass. A more efficient two-pass approach is to first query a sample of the fragment and use those results to filter out trivial matches. The sampling approach we have developed is derived from the W-curve's original use as a visual tool: uses lining up the curves manually will start by making the extreme points match. Lining up these "peaks" in two curves is the fastest way to visually align the curves. In the database, this starts by selecting fragment points outside of a radius from the X-Y origin, usually 0.5. The first pass sample points are used to select matching template points,. The sample results are then filtered using adjacencies to remove trivial matches, providing a set of templates for complete matches. In the second pass, all points are compared to the templates with added restrictions based on the sample point matches. The result of this second selection are finally arranged for maximum coverage of the fragment and ranked by total coverage to produce candidate alignments.

This approach gracefully handles recombinant matches by simply locating the points on all available template curves. The queries are readily adapted to distributed computing since the point comparisons are independent, depending only on the positions of individual fragment and template point locations. The queries can restrict the locations of points using the relative base numbers of sample points, delivering a manageable amount of data to the central node for filtering. Even the filtering can be parallelized to the number of template sequences selected since those evaluations are independent.

The main remaining issue is defining a database which can be suitably queried.

4. Database Layout & Queries

The database schema that supports these queries has to deal gracefully with rounding errors, gradual effects of auto-regression after a SNP, and phase shifts in the curves after a gap. It also needs to be compact in both for query performance and distribution to nodes in the cloud.

One design that might seem attractive is storing the points as three-dimensional entities and simply querying the curves for intersecting points. This fails on two fronts, however, since it permit querying the X-Y locations of points independently of their base numbers or the direct selection of base numbers.

Comparing the points without reference to their base numbers initially requires storing the curves with an X-Y value and separate base number. At this point storing X-Y values as points and querying the distance might seem reasonable. The storage is simple and compact, but the distance computation is too expensive and points alone are exquisitely sensitive to rounding errors. Storing all of the vertexes as polygons solves rounding errors but requires storing bulky objects with expensive intersect/overlap queries.

The final solution was a mixture of point and circle objects. The template W-curve vertexes are stored using X-Y points. The fragments, however, are stored as circles (Fig. 3). This provides a relatively compact database in both cases, with a simple query for the points contained within the circles. The circles also make effective use of spatial indexes, which store a bounding box containing the geometry. The bounding box for circles is efficient to compute, minimizes rounding error, and is an effective filter for the contained-within queries used with template points.

Storing fragment vertexes as circles also helps solve two issues with the W-curve that have been ignored thus far: initial bases in fragments and multi-base alternatives in the sequences. The former is a problem that W-curves generated from short reads all begin at the origin before their first base, but the template curves are in mid-sequence. This leaves the first few bases of the fragment's curve are guaranteed not to match the corresponding points on the template. One solution is to prefix the curve with the 16 possible two-base alternatives and draw a larger circle around the resulting locations for the first 2-3 bases. These larger circles are essentially a fuzzy-matching approach to the alignment. This approach permits matching in the first few bases of the fragment at the expense of filtering out more points due to extraneous matches.

Multi-base alternatives found in Fast Q output of NGS systems can be handled in a similar fashion: simply draw larger circles or store multiple *circularstring* objects in the database for the alternative bases. Resulting matches from each circle could be weighted according to the quality values for the final match. Again, the filtering process for adjacencies will remove any one-off matches.

A skeleton database supporting these queries has four tables: two of identifiers with any additional non-geometric data and two of W-curve values, one with points one with circles for the geometry (Fig. 4). The identifier tables have a candidate key of the template's external identifier and an integer surrogate key for use in the geometry tables. The template geometry table has a candidate key of the template's surrogate key (SK) and base number; fragment geometry requires a candidate key of the fragment's SK, a base number, and the base AA from which the geometry is defined. The

```

create table sequence
(
  id          serial          not null,
  parent     integer         not null references dna default 0,
  ident      varchar(32)     not null,
  sequence   text           not null default '',

  primary key ( id ),
  unique ( ident, parent )
);
create table template
(
  seq        integer not null references sequence,
  base       integer not null,
  nucleotide char(1) not null,

  primary key ( dna, base_no )
);
create table fragment
(
  seq        integer not null references sequence,
  base       integer not null,
  nucleotide char(1) not null,

  primary key( dna, base_no, base_aa )
);

select AddGeometryColumn( 'template', 'vertex', -1, 'POINT',          2 );
select AddGeometryColumn( 'fragment', 'vertex', -1, 'CIRCULARSTRING', 2 );

create index fragment_vertex_ix on fragment using gist( vertex );

insert into sequence ( id, parent, ident ) values ( 0, 0, 'root' );
insert into sequence ( ident ) values ( "B.K03455" );
insert into sequence ( ident, parent ) values ( "gp120", 1 );

prepare insert_template( integer, integer, char, varchar )
as insert into template ( seq, base, nucleotide, vertex )
values ( $1, $2, $3, St_GeometryFromText($4) );

prepare insert_fragment( integer, integer, char, varchar )
as insert into template ( seq, base, nucleotide, vertex )
values ( $1, $2, $3, St_GeometryFromText($4) );

insert_template( 1, 1, 'T', "POINT(0.5 0)" );
insert_fragment( 2, 1, 'A', "CIRCULARSTRING(-0.05 0.5,0.05 0.5, -0.05 0.5)" );

```

Figure 4: Skeleton database for storing template and fragment W-curve vertices. Example inserts show the first vertex for HXB2 and its gp120 protein.

fragment's larger candidate key is required to accommodate storing multiple circles for handling FastQ results.

The circles are handled via *CIRCULARSTRING* objects. These can describe full circles using three points with the first and last points matching and the second point being opposite the first. Using an offset to the X-axis value for each vertex requires minimal computation for the input data and produces a bounding box without rounding error. For example, in Figure 3 the first vertex for 'T' at (0, 0.5) produces a circle with points (0, 0.55), (0, 0.45); in Figure 4 shows the input format with three points, with the initial vertex for 'A' at *CIRCULARSTRING*(-0.05 0.50, 0.05 0.50, -0.05 0.50), -1.

The general query for alignments selects the dna.id, base_no, and base_aa values from template, fragment tables “where template.vertex && fragment.vertex”. The '&&' operator looks for intersecting geometry and makes efficient use of bounding boxes in the fragment's spatial index.

5. Further Research

Determining the most effective radius for the fragment radius and how to use either variable-radius or multiple-circle designs will be important for matching short sequences provided as FastQ inputs used with most NGS machines. In addition, an efficient, distributed filtering algorithm for the first pass selection from sample fragments will be key to making this approach efficient in cloud-computing environments.

6. Conclusion

The W-curves' abstract, three-dimensional geometry for representing DNA sequences provides more detail than a uni-dimensional character sequence. Its balance of local divergence and global convergence make the representation useful for aligning sequences that character-based algorithms cannot handle well. Geometric data objects now give us the tools to mine W-curve databases effectively, handling highly variable and crossover recombinant sequences. The algorithm described here uses generic tools such as SQL, and is suitable for highly-parallel environments such as cloud computing. Although the examples here use HIV-1, other non-correcting viruses or oncology are also good candidates for its application. We are not saying that the W-curve is a replacement for Fasta or Clustal, but used as an adjunct to them it opens up new opportunities for studying difficult sequences. And that has to be our goal going forward: expanding the range of tools available for studying the complexity of biology.

7. References

1. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome

and comparison with the human genome. *Nature* 2005 Sep 1;437:69–87.

2. Remsen J, O'Grady P. Phylogeny of *Drosophilinae* (Diptera: Drosophilidae), with comments on

combined analysis and character support. *Molecular Phylogenetics and Evolution* 2002, 24:249–264.

3. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs.

Bioinformatics 2002, 18:452-464.

4. Grasso C, Lee C. Combining partial order alignment and progressive multiple sequence alignment

increases alignment speed and scalability to very large alignment problems. *Bioinformatics* 2004, 20:1546-1556.

5. Christophe Fraser. HIV recombination: what is the impact on antiretroviral therapy? *J R Soc Interface*.

2005 December 22; 2(5): 489–503.

6. Biegert A, Söding J (March 2009). Sequence context-specific profiles for homology searching.

Proceedings of the National Academy of Sciences of the United States of America 2009 March;106(10): 3770–5

7. Cork DJ, Toguemf A. Using fuzzy logic to confirm the integrity of a pattern recognition algorithm for

long genomic sequences: the W-curve. *Ann N Y Acad Sci*. 2002 Dec;980:32-40.

8. Cork DJ, Lembark S, Tovanabutra S, Robb ML, Kim JH (2010) W-Curve Alignments for HIV-1

Genomic Comparisons. *PLoS ONE* 5(6): e10829. doi:10.1371/journal.pone.0010829

9. Ahmad N et al. Preserving Data Replication Consistency through ROWA-MSTS.

Communications in

Computer and Information Science 2011;180(1): 244-253.

SESSION

EXPERIMENTAL MEDICINE, COMPUTER-ASSISTED MEDICAL CARE AND SERVICE SYSTEMS, ANALYSIS AND DIAGNOSTIC TOOLS

Chair(s)

TBA

Validating Critical Limits of the Universal Brain Injury Criterion

Igor Szczyrba¹, Martin Burtcher², and Rafał Szczyrba³

¹School of Mathematical Sciences, University of Northern Colorado, Greeley, CO 80639, U.S.A.

²Department of Computer Science, Texas State University–San Marcos, TX 78666, U.S.A.

³Funiosoft, LLC, Silverthorne, CO 80498, U.S.A.

Abstract— We present results of numerical simulations that further validate the critical limits we previously proposed for our universal Brain Injury Criterion (BIC). The BIC extends the applicability of the translational Head Injury Criterion (HIC) to arbitrary head motions by reformulating the acceleration-based HIC formula in terms of the energy transferred locally from the skull to the brain. Our simulations are based on a generalization of the Kelvin-Voigt (K-V) Closed Head Injury model that includes a nonlinear strain-stress relation. We validate the proposed BIC limits against (i) the critical limit $HIC_{15} = 700$, (ii) the Diffuse Axonal Injury Tolerance Criterion (DAITC) for head rotations that has been derived from the K-V model and from experiments with animal brains, and (iii) recent experimental data on strain levels leading to permanent neuronal damage. Our results imply that for head rotations about various fixed axes, the critical BIC_{15} limits coincide with the HIC_{15} critical limit and are in agreement with the DAITC thresholds.

Keywords: brain injury, universal critical limits

1. Introduction

In previous work [1], [2], [3], we have introduced a universal Brain Injury Criterion (BIC) that allows assessing Closed Head Injury (CHI) caused by arbitrary traumatic head motions. Our approach is based on the assumption that if energy is transferred *locally* from the skull to the brain in a similar way, the likelihood and severity of a brain injury in a given location should be similar in any traumatic scenario, including traumatic head translations.

This article makes the following contributions: First, to the best of our knowledge, we are the first to establish that the way in which energy is transferred locally from the skull to the brain can play a crucial role in determining the likelihood and severity of a brain injury during arbitrary traumatic head motions. Specifically, we consider the rate at which energy (i.e., power) is transferred to the brain per unit mass from the moving skull as well as the rate of power transferred to the brain (i.e., whether energy is transferred to the brain in an accelerated or constant way).

Second, by using the energy/power transferred locally from the skull to the brain during traumatic head motions

as a predictor of a brain injury, we introduce a ‘common denominator’ for assessing the severity and likelihood of the injury appearing as a result of traumatic head translations and rotations. This makes it possible to establish a direct link between the translational and rotational critical limits introduced by other researchers.

Third, we demonstrate how the operator norm of the strain matrix can be used to evaluate the time evolution of the spatial distribution of the maximal strain in the brain matter.

Fourth, by numerically simulating various traumatic scenarios using our nonlinear CHI model, we show that, for head rotations about *fixed* axes lasting for 0.015s, the BIC critical limits (i) do not depend in an essential way on the position of the rotational axis, (ii) coincide with the new critical limit $HIC_{15} = 700$, and (iii) are in agreement with the existing rotational Diffuse Axonal Injury thresholds.

1.1 Derivation of the BIC formula

Based on our assumption regarding the local transfer of energy from the skull to the brain, we derive the BIC formula from the well-established translational Head Injury Criterion (HIC) formula:

$$HIC_{1000T} = \max A^{2.5} T, \quad (1)$$

where T is a time subinterval of the head’s translational acceleration time, A is the average (over T) of the acceleration magnitude’s absolute value, and the maximum is taken over all subintervals T .

Specifically, for monotone accelerations, we express A in terms of the energy E and the power P as follows:

$$A = \sqrt{2P} / (\sqrt{E(t_2)} + \sqrt{E(t_1)}), \quad (2)$$

where $E(t)$ denotes the average kinetic energy per unit mass at time t transferred to the brain surface from a translated skull, and $P = |E(t_2) - E(t_1)| / T$ is the absolute value of the average power transferred per unit mass to the brain in the time interval $T = t_2 - t_1$, cf. [1] for details.

The reformulation of the acceleration A in terms of energy and power allows us to generalize the applicability of the HIC formula (1) to arbitrary traumatic head motions, i.e., to introduce the following formula:

$$BIC_{1000T} = \max \left(\frac{\sqrt{2P}}{\sqrt{2E(t_1)} + \sqrt{2E(t_2)}} \right)^{2.5} T. \quad (3)$$

Let us note that, contrary to the case of a head translation, during an arbitrary head motion, energy is transferred non-uniformly from the skull to the brain, i.e., both E and P depend not only on the time t but also on the localization of the brain parcels. For instance, during a head rotation about a *fixed* axis, the energy transferred to the brain is negligible near the axis because the magnitude of the rotational velocity is very small there. Hence, in case of an arbitrary traumatic head motion, the maximum in the formula (3) should be taken not only over all time intervals T but also over the entire brain surface.

If, for a time interval $T=t_2-t_1$ for which the maximum in (3) is assumed, the velocity of an acceleration pulse is zero at t_1 or t_2 , the BIC formula (3) can be simplified to become a function of only the average power P and the duration of the acceleration time T :

$$BIC_{1000T} = \max 2P(2P/T)^{0.25}, \quad (4)$$

where the ratio P/T approximates the rate at which power is transferred to the brain. Thus, our BIC formula (4) exposes a new role that is possibly played in the creation of brain injuries by an accelerated delivery of power from the skull to the brain (second temporal derivative of energy).

1.2 Rotations about fixed axes

In the case of accelerated head rotations about *fixed* axes (which we focus on in this study), the requirement that the maximum should be taken over the entire brain surface can be relaxed by considering only a thin strip of the brain's surface located along the boundary of the 2D brain cross section that is perpendicular to the rotational axis and is characterized by the *highest* value of tangential velocity.

Moreover, if the magnitude of this tangential velocity over time is the same as the magnitude of the translational velocity characterizing a head's accelerated translation, the likelihood and severity of a brain injury appearing in this brain cross-section should be similar to the likelihood and severity of an injury when the head is translated since the local transfer of energy along the cross section's boundary is practically identical in both traumatic scenarios. Consequently, it should be possible to directly use the critical HIC limits introduced in [4], [5] to derive the critical BIC limits for traumatic head rotations about fixed axes.

1.3 Correlation between translational and rotational brain injury criteria

In deriving BIC critical limits for traumatic head rotations about fixed axes, our approach allows us to also use the rotational critical limits introduced by the Diffuse Axonal Injury Tolerance Criterion (DAITC). DAITC has been developed in 1992 based on experiments with baboon brains and the *linear* viscoelastic Kelvin-Voigt (K-V) CHI

model, *cf.* [6]. The DAITC is expressed in terms of the peak rotational acceleration about a *fixed* rotational axis positioned centroidally, and the peak change of the rotational velocity's magnitude.

In fact, considering how energy is transferred locally from the skull to the brain in traumatic situations as a brain injury predictor allows us to link the translational critical HIC limits with the rotational critical DAITC limits. For instance, the maximum translational acceleration of a triangularly shaped acceleration pulse characterized by the critical $HIC_{15}=700$ limit equals $150g=1,472m/s^2$ and the corresponding peak change in the velocity is $5.4m/s$, *cf.* Fig. 2 in Section 3.

If the same tangential pulse is used to centroidally rotate an adult human head with an 'average radius' of $0.1m$ about a fixed axis, the maximum rotational acceleration equals $14,460rad/s^2$ and the peak change in the rotational velocity magnitude equals $55rad/s$. This corresponds to a point that is near the critical region defined by the DAITC analytic model's threshold curve and is inside the critical region defined by the DAITC physical model, *cf.* Fig. 5 in [6].

2. Generalized Kelvin-Voigt CHI model

We further validate the critical BIC limits by conducting simulations using our numerical nonlinear CHI model that generalizes the K-V model used to derive DAITC.

2.1 Nonlinear stress-strain relation

Following experimental data obtained over the last two decades, *cf.* [7], [8], [9], we include a *nonlinear* stress-strain relation in our generalization of the K-V CHI model. Thus, our computational model utilizes the following Partial Differential Equations (PDEs) describing the propagation of shear waves in incompressible viscoelastic materials:

$$\begin{aligned} \frac{\partial \mathbf{v}(\mathbf{x}, t)}{\partial t} &= \Delta(c^2(\mathbf{x}, t)\mathbf{u}(\mathbf{x}, t) + \nu \mathbf{v}(\mathbf{x}, t)), \\ \frac{\partial \mathbf{u}(\mathbf{x}, t)}{\partial t} &= \mathbf{v}(\mathbf{x}, t), \quad \nabla \cdot \mathbf{v}(\mathbf{x}, t) = 0, \end{aligned} \quad (5)$$

where $\mathbf{v}(\mathbf{x}, t) \equiv (v_1(\mathbf{x}, t), v_2(\mathbf{x}, t), v_3(\mathbf{x}, t))$ with $\mathbf{x} \equiv (x_1, x_2, x_3)$ represents the brain matter velocity vector field at time t in an *external* coordinate system, $\mathbf{u}(\mathbf{x}, t)$ is the corresponding displacement vector field, $c(\mathbf{x}, t)$ describes the brain's shear wave velocity that depends on the distribution of strain in the brain matter, and ν is the brain's kinematic viscosity.

Specifically, based on experimental data reported in [7], we model the stress-strain relation as an exponential function, i.e., we set

$$c(\mathbf{x}, t) \equiv c \cdot \exp(r \cdot s(\mathbf{x}, t)), \quad (6)$$

where $c \equiv \sqrt{G/\delta}$ denotes the basic shear wave velocity in the absence of strain (with G and δ being the brain matter shear modulus and density, respectively), $s(\mathbf{x}, t)$ describes the time

evolution of the spatial distribution of the maximum strain within the brain matter, and r is a coefficient determining how the brain matter stiffens under strain.

Since there are no experimental data on the brain matter's strain-stress relation for very large strains, we make the assumption that for strains larger than $m\%$, e.g., exceeding $m = 50\%$, the shear wave velocity $c(\mathbf{x}, t)$ given by (6) 'saturates', i.e., it smoothly becomes proportional to the basic velocity c .

2.2 Spatial distribution of maximal strain

To find the spatial distribution $s(\mathbf{x}, t)$ of the maximal strain, we evaluate the components of the matrix:

$$\mathbf{S}(\mathbf{x}, t) \equiv \nabla \cdot \mathbf{U}(\mathbf{x}, t) + \mathbf{I} \equiv \partial \mathbf{U} \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}, t) + \mathbf{I}, \quad (7)$$

where $\mathbf{U}(\mathbf{x}, t) \equiv (U_1(\mathbf{x}, t), U_2(\mathbf{x}, t), U_3(\mathbf{x}, t))$ denotes the brain matter's displacement vector field *relative* to the moving skull, $\nabla \cdot \mathbf{U}(\mathbf{x}, t)$ is the strain matrix of this field, and \mathbf{I} is the identity matrix in 3D.

The diagonal terms in the strain matrix $\nabla \cdot \mathbf{U}(\mathbf{x}, t)$ determine the contribution of the partial derivatives to the brain deformation by evaluating the brain matter's *strain* with regard to the directions of the base vectors used, whereas the partial derivatives in the non-diagonal terms determine their contribution to the brain deformation by evaluating the *total deformation* of the brain matter.

Adding *one* to the diagonal terms in (7) puts all partial derivatives on 'equal footing', i.e., enables us to evaluate the maximal total deformation in each point \mathbf{x} of the brain at time t by using the operator norm $\|\cdot\|_O$ of the matrix $\mathbf{S}(\mathbf{x}, t)$. Next, by subtracting *one* from this maximal total deformation, we obtain the spatial distribution of the maximal strain at time t . Thus, the function $s(\mathbf{x}, t)$ is given by:

$$s(\mathbf{x}, t) \equiv \|\mathbf{S}(\mathbf{x}, t)\|_O - 1 \equiv \sup \|\mathbf{S}(\mathbf{x}, t) \cdot \mathbf{y}\| - 1, \quad (8)$$

where $\mathbf{y} \equiv (y_1, y_2, y_3)$, $\|\cdot\|$ denotes the vector norm in 3D, and the supremum is taken over all vectors \mathbf{y} with $\|\mathbf{y}\| = 1$. One can easily check that:

$$s(\mathbf{x}, t) \leq \|\nabla \cdot \mathbf{U}(\mathbf{x}, t)\|_O, \quad (9)$$

i.e., the operator norm $\|\nabla \cdot \mathbf{U}(\mathbf{x}, t)\|_O$ of the strain matrix provides an upper bound for the function $s(\mathbf{x}, t)$ describing the spatial distribution of the maximal strain.

3. Simulation setup

In this paper, we present simulation results of head translations in a fixed direction as well as of head rotations about certain fixed rotational axes. As mentioned above, this allows us to directly verify the results of our numerical simulations using the HIC and DAITC critical limits.

In both of these scenarios the forces applied to the head have one zero component, and consequently, one component

of the 3D solutions is zero. Therefore, it suffices to solve PDEs (5)–(8) only in 2D brain cross sections near which the transfer of energy from the skull to the brain is the largest.

Thus, for forward head translations and rotations, we present the results of our simulations in a sagittal brain cross section that is positioned near the falx cerebri, whereas for lateral head rotations, we present the results in a coronal brain cross section that is positioned near the brain's center of mass and that includes the falx cerebri.

3.1 Skull-brain facsimile

As solution domains, we use 2D facsimiles of the skull-brain cross sections consisting of three layers: (i) the skull and dura layer, (ii) the Cerebro Spinal Fluid (CSF) layer, and (iii) the brain matter layer, *cf.* Fig 1. Specifically, we model the skull and the dura mater as a solid body layer, the $4 \cdot 10^{-3}\text{m}$ thick CSF layer representing the pia-arachnoid complex with the fluid is modeled as an incompressible elastic medium, and the brain matter is modeled as an incompressible viscoelastic medium.

Since there exist no conclusive experimental data on how the stress depends on the strain in the gray matter, the brain matter is assumed to be homogenous having the physical characteristics of the white matter.

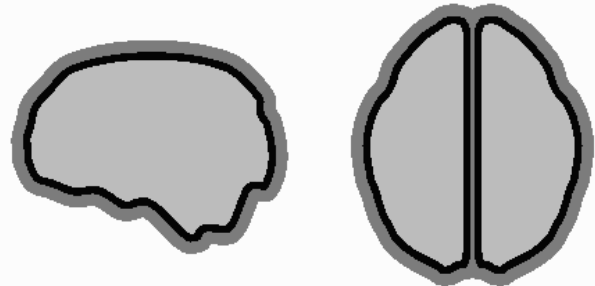


Fig. 1

THE THREE-LAYER SAGITTAL AND CORONAL HEAD CROSS SECTIONS
SOLID BODY SKULL AND DURA MATER - DARK GRAY, ELASTIC CSF
COMPLEX - BLACK; VISCOELASTIC HOMOGENOUS BRAIN - LIGHT GRAY

Experimental data in [7], [10–17] imply that the shear wave velocity in the white matter is approximately 1m/s , the stiffening coefficient $0.5 \leq r \leq 2.5$, the brain's viscosity $0.009\text{m}^2/\text{s} \leq \nu \leq 0.017\text{m}^2/\text{s}$, and neurons can sustain mechanical strain up to 80%.

According to [18], the CSF layer is predominately modeled as an incompressible elastic medium with a shear modulus G_{CSF} as low as 200PA , which reflects the role of the CSF in reducing the strain within the brain matter [19]. The simulation results presented here are obtained with the following values for the constants in the system (5)–(8): $c = 1\text{m/s}$, $\nu = 0.013\text{m}^2/\text{s}$, $r = 1.4$, $m = 50\%$, $G_{CSF} = 225\text{PA}$.

3.2 Acceleration loads used

We simulate forward head translations and forward head rotations about fixed horizontal axes positioned at the head's center of mass, the chin, the neck, and the abdomen as well as lateral head rotations about fixed vertical axes positioned at the head's center of mass, the skull, and at some distances outside of the skull.

We present simulation results obtained using a triangular acceleration load with the acceleration time $T = 0.015s$ and the tangential acceleration and velocity magnitudes corresponding to $HIC_{15} = BIC_{15}$ ranging from 100 to 1000. Fig. 2 depicts the dynamic characteristics of the critical load used with $HIC_{15} = BIC_{15} = 700$.

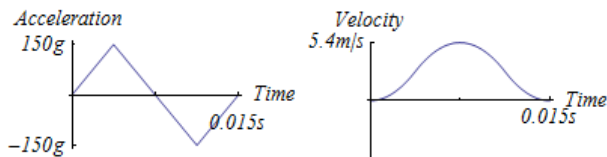


Fig. 2

DYNAMIC CHARACTERISTICS OF THE ACCELERATION LOAD WITH
 $HIC_{15} = BIC_{15} = 700$

4. Simulation results

To evaluate the possible severity of a brain injury, we find the absolute maximum s_{max} of the function $s(\mathbf{x}, t)$, i.e., the maximum strain value attained in a given brain cross section during or some time after the head is accelerated.

4.1 Simulations of forward head translations

Table 1 depicts the values of s_{max} attained in the coronal and sagittal brain cross sections in our simulations of forward head translations under loads characterized by four HIC_{15} values ranging between 100 and 1000.

HIC_{15}	100	400	700	1000
coronal	10%	20%	25%	27%
sagittal	17%	27%	35%	38%

Table 1

MAXIMAL STRAIN s_{max} IN THE CORONAL AND SAGITTAL CROSS SECTIONS ATTAINED DURING OR AFTER FORWARD HEAD TRANSLATIONS WITH HIC_{15} RANGING BETWEEN 100 AND 1000

Experiments imply that neurons sustain permanent damage due to a chemical imbalance when stretched by 25%-30% [6], [7], [20]. For the $HIC_{15} = 700$ load, the average of the s_{max} values attained in both brain cross sections equals 30%. Thus, our translational results are in good agreement with this critical HIC limit, which validates the predictions of our computational CHI model.

The disparity between the s_{max} values in the coronal and sagittal brain cross sections are most likely due to the fact that the simulations with the sagittal cross section do not take into account the impact of the falx cerebri, which seems to lower the maximal strain, cf. [6].

4.2 Simulations of head rotations

Diffuse Axonal Injuries (DAI) appear predominantly as a result of rapid head rotations, cf. [21]. To derive critical BIC values that can be used to assess the severity and likelihood of DAI, we conduct numerous simulations of head rotations under loads characterized by a variety of BIC values.

Table 2 depicts the values s_{max} attained under BIC_{15} loads ranging from 100 to 1000 in the sagittal brain cross section during or after forward head rotations about fixed horizontal axes positioned at the head's center of mass, the chin, the neck, and the abdomen.

s_{max} values in sagittal cross section	BIC_{15}			
forward rotation about fixed axis at	100	400	700	1000
head's center of mass	17%	32%	37%	39%
chin	17%	32%	36%	43%
neck	13%	29%	35%	40%
abdomen	19%	29%	36%	39%

Table 2

MAXIMAL STRAIN s_{max} IN THE SAGITTAL CROSS SECTION ATTAINED DURING OR AFTER FORWARD HEAD ROTATIONS ABOUT VARIOUS HORIZONTAL AXES WITH BIC_{15} VALUES BETWEEN 100 AND 1000

Table 3 shows the values s_{max} attained under the same BIC_{15} loads but in the coronal brain cross section when the head is rotated laterally, counter-clockwise about fixed vertical axes positioned at the head's center of mass, the skull, 0.1m from the skull, and 0.2m from the skull.

s_{max} values in coronal cross section	BIC_{15}			
lateral rotation about fixed axis at	100	400	700	1000
head's center of mass	11%	25%	39%	40%
skull	17%	35%	40%	41%
0.1m from the skull	16%	35%	40%	40%
0.2m from the skull	15%	30%	35%	37%

Table 3

MAXIMAL STRAIN s_{max} IN THE CORONAL CROSS SECTION ATTAINED DURING OR AFTER LATERAL HEAD ROTATIONS ABOUT VARIOUS VERTICAL AXES WITH BIC_{15} VALUES BETWEEN 100 AND 1000

These simulation results provide maximal strain values s_{max} that are slightly higher than (but still in line with) the values obtained for head translations. Let us note, however, that during rapid head rotations the absolute maxima s_{max} of strain are, in general, attained in a pointwise manner in very small regions of the brain matter during a very short period of time lasting for 0.01s to 0.02s.

Such localized high strain values lasting for such short periods of time should only be used to obtain an upper bound

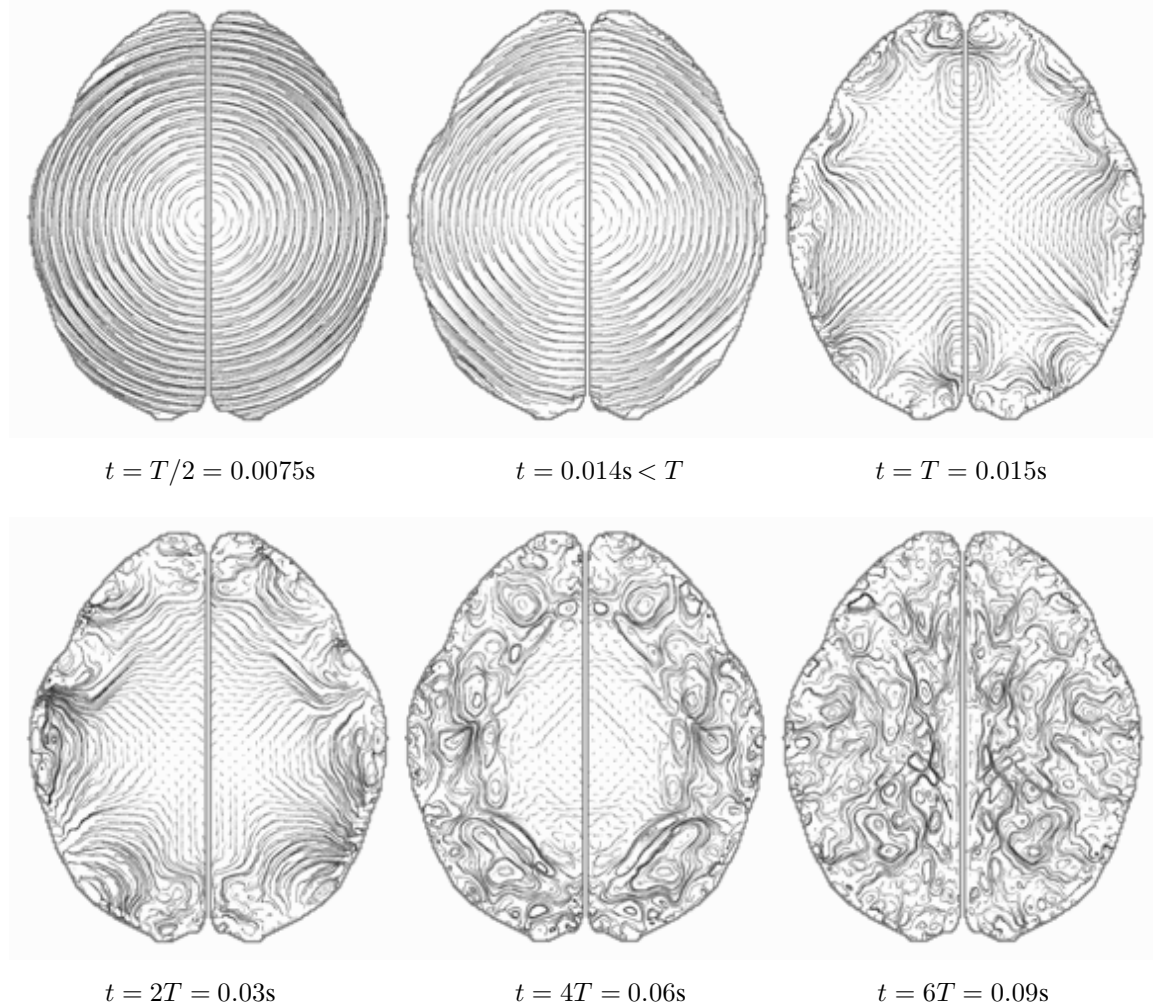


Fig. 3

TIME EVOLUTION OF THE VELOCITY CURVED VECTOR FIELD $V(x_1, x_2, t)$ RELATIVE TO THE SKULL IN THE CORONAL BRAIN CROSS SECTION DURING AND AFTER A LATERAL HEAD ROTATION WITH $BIC_{15} = 700$ ABOUT A VERTICAL AXIS POSITIONED AT THE BRAIN'S CENTER OF MASS

estimate for predicting DAI likelihood and severity, since the loss of axonal transport in a single axon does not properly reflect the spatial scattering of DAI [22].

Instead, the Cumulative Strain Damage Measure (CSDM) introduced in [23] has been accepted as a good DAI predictor [24]. An initial analysis of our simulation results from the point of view of the CSDM suggests that the critical HIC value of 700 can be used as the BIC critical value for head rotations about fixed axes and as a starting point for establishing critical BIC limits for arbitrary head rotations.

Since commercial software cannot adequately depict highly localized oscillations of vector fields, we have developed animated Curved Vector Field (CVF) plots [25]. CVF plots use curved, dark-to-light shaded lines instead of arrows

to indicate the motion's direction. They provide a good depiction of vectors and portray potential trajectories of brain parcels. Animated versions of our CVF plots are available at <http://www.funiosoft.com/brain/> in form of MPEG movies.

Fig. 3 (resp. 4) depicts time snapshots of CVF animations representing the brain matter's velocity vector field $V(x_1, x_2, t)$ relative to the moving skull at various times t in the coronal (resp. sagittal) 2D brain cross section when the head is rotated laterally, counter-clockwise (resp. forward) under the $BIC_{15} = 700$ load about an axis positioned at the brain's center of mass.

The highly localized brain matter oscillations depicted in Figs. 3 and 4 create multiple local strain maxima that are scattered over the entire brain cross section.

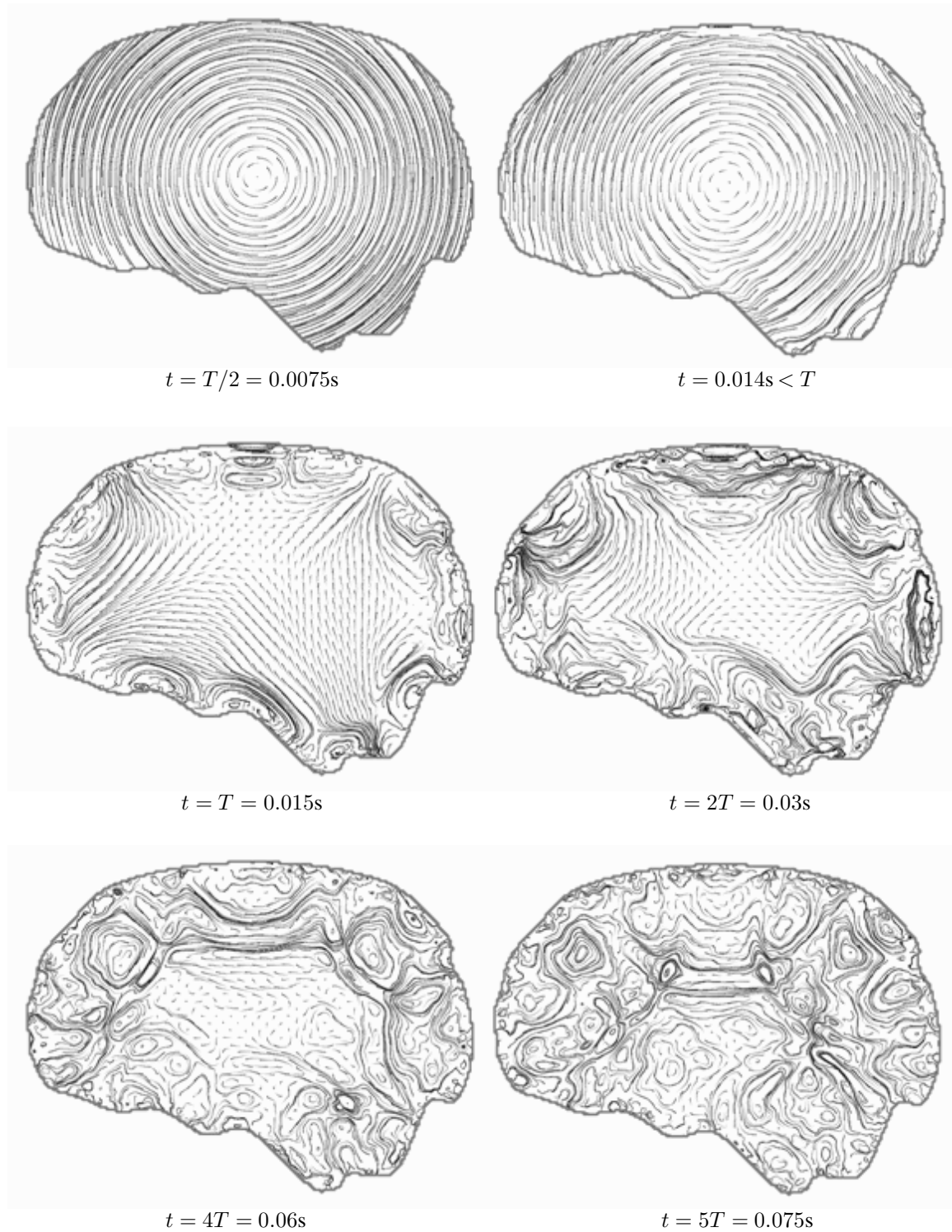


Fig. 4

TIME EVOLUTION OF THE VELOCITY CURVED VECTOR FIELD $V(x_1, x_2, t)$ RELATIVE TO THE SKULL IN THE SAGITTAL BRAIN CROSS SECTION DURING AND AFTER A FORWARD HEAD ROTATION WITH $BIC_{15} = 700$ ABOUT A HORIZONTAL AXIS POSITIONED AT THE BRAIN'S CENTER OF MASS

Note that, in the case of the lateral head rotation, the brain matter oscillations 'spread' throughout the entire cross section at a later time in comparison to the forward head rotation. This shows again that the falx cerebri plays a role in shaping the DAI features.

5. Conclusions

Our idea that the severity and likelihood of brain injuries can be assessed, regardless of whether a head is translated or rotated, based on the analysis of how the energy is locally transferred from the skull to the brain enables us to develop a universal Brain Injury Criterion applicable for arbitrary traumatic head motions. Our approach further allows to correlate the new Head Injury Criterion critical limits derived in [4], [5] with the Diffuse Axonal Injury Tolerance Criterion critical values established in [6].

The results from numerical simulations based on our viscoelastic Closed Head Injury model that includes a nonlinear strain-stress relation imply that, for centroidal and non-centroidal head rotations about *fixed* axes with an acceleration time period $T=0.015s$, the critical BIC_{15} limits:

- do not depend in an essential way on the position of the fixed rotational axis,
- coincide with the new critical $HIC_{15} = 700$ limit, and
- are in agreement with the critical limits implied by the DAITC threshold curves.

These results suggest that the critical $BIC_{15} = 700$ limit may be valid for arbitrary traumatic head motions.

6. Acknowledgment

The authors would like to thank Intel Corporation for providing two multiprocessor servers that were used for conducting a portion of the numerical simulations presented in this article.

References

- [1] I. Szczyrba, M. Burtscher, and R. Szczyrba, "A Proposed New Brain Injury Tolerance Criterion Based on the Exchange of Energy Between the Skull and the Brain," in *Proc. 2007 Summer Bioengineering Conf.*, American Society of Mechanical Engineers, SBC 2007-171967, 2007.
- [2] I. Szczyrba, M. Burtscher, and R. Szczyrba, "Computational Modeling of Brain Dynamics during Repetitive Head Motions," in *Proc. 2007 Conf. on Modeling, Simulation and Visualization Methods*, pp. 143-149, CSREA Press 2007.
- [3] I. Szczyrba, M. Burtscher, and R. Szczyrba, "On the Role of a Nonlinear Stress-Strain Relation in Brain Trauma," in *Proc. 2008 Conf. Bioinformatics and Computational Biology*, vol. 1, pp. 265-271, CSREA Press 2008.
- [4] M. Kleinberger *et al.*, "Development of Improved Injury Criteria for the Assessment of Advanced Automotive Restraint Systems," (1998) The NHTSA website. [Online]. Available: <http://www-nrd.nhtsa.dot.gov/pdf/nrd-11/airbags/criteria.pdf>
- [5] R. Eppinger *et al.*, "Development of Improved Injury Criteria for the Assessment of Advanced Automotive Restraint Systems-II," (2000) The NHTSA website. [Online]. Available: http://www-nrd.nhtsa.dot.gov/pdf/nrd-11/airbags/finalrule_all.pdf
- [6] S. S. Margulies, and L. Thibault, "A Proposed Tolerance Criterion for Diffuse Axonal Injury in Man," *J. of Biomechanics*, vol. 25, pp. 917-923, 1992.
- [7] B. R. Donnelly, and J. Medige, "Shear Properties of Human Brain Tissue," *J. of Biomechanical Engineering*, vol. 119, pp. 423-432, 1998.
- [8] E. G. Takhounts, J. R. Crandall, and K. Darvish, "On the Importance of Nonlinearity of Brain Tissue under Large Deformations," *Stapp Car Crash J.*, vol. 47, pp. 79-92, 2003.
- [9] S. Mehdizadeh, *et al.*, "Comparison between Brain Tissue Gray and White Matters in Tension Including Necking Phenomenon," *American J. of Applied Sciences*, vol. 5, no. 12, pp. 1701-1706, 2008.
- [10] G. T. Fallenstein, V. D. Hulce, and J. W. Melvin, "Dynamic Material Properties of Human Brain Tissue," *J. of Biomechanics*, vol. 2, pp. 217-226, 1969.
- [11] C. Ljung, "A Model for Brain Deformation Due to Rotation of the Skull," *J. of Biomechanics*, vol. 8, pp. 263-274, 1975.
- [12] Y. Tada, and T. Nagashima, "Modeling and Simulation of Brain Lesions by the Finite-Element Method," *IEEE Engineering in Medicine and Biology*, pp. 497-503, 1994.
- [13] B. R. Donnelly, "Brain tissue material properties: A comparison of results. Biomechanical research: Experimental and computational," in *Proc. 26th Int. Workshop*, pp. 47-57, 1998.
- [14] K. Paulsen, *et al.*, "A Computational Model for Tracking Subsurface Tissue Deformation During Stereotactic Neurosurgery," *IEEE Transactions on Biomechanical Engineering*, vol. 46, pp. 213-225, 1999.
- [15] J. A. Wolf, *et al.*, "Calcium Influx and Membrane Permeability in Axons after Dynamic Stretch Injury in Vitro," *J. of Neurotrauma*, vol. 16, p. 966, 1999.
- [16] A. Bain, and D. Meaney, "Tissue-Level Thresholds for Axonal Damage in an Experimental Model of Central Nervous System White Matter Injury," *J. of Biomechanical Engineering*, vol. 122, pp. 615-622, 2000.
- [17] A. Bain, *et al.*, "Dynamic Stretch Correlates to Both Morphological Abnormalities and Electrophysiological Impairment in a Model of Traumatic Axonal Injury," *J. of Neurotrauma*, vol. 18, pp. 499-511, 2001.
- [18] J. Xin, H. K. Yang, and A. J. King, "Mechanical properties of bovine pia-arachnoid complex in shear," *J. of Biomechanics*, vol. 44, pp. 467-474, 2011.
- [19] Y. H. Chu, "Finite Element Analysis of Traumatic Brain Injury," 2002 [Online]. Available: at: <http://www.ruf.rice.edu/preors/Chu-YH.pdf>
- [20] Y. Matsui, and T. Nishimoto, "Nerve Level Traumatic Brain Injury in Vivo/in Vitro Experiments," *Stapp Car Crash J.*, vol. 54, pp. 197-210, 2010.
- [21] J. Meythaler, "Amantadine to Improve Neurorecovery in Traumatic Brain Injury-associated Diffuse Axonal Injury," *J. of Head Trauma and Rehabilitation*, vol. 17, no. 4, pp. 303-313, 2002.
- [22] T. A. Gennarelli, *et al.*, "Diffuse axonal injury and traumatic coma in the primate," *Annals of Neurology*, vol. 12, no. 6, pp. 564-574, 1982.
- [23] F. A. Bandak, and R. H. Eppinger, "A three-dimensional finite element analysis of the human brain under combined rotational and translational accelerations," in *Proc. 38th Stapp Car Crash Conf.*, pp. 145-163, 1994.
- [24] E. G. Takhounts, *et al.*, "Investigation of Traumatic Brain Injuries Using the Next Generation of Simulated Injury Monitor (SIMon) Finite Element Head Model," *Stapp Car Crash J.*, vol. 52, pp. 1-31, 2008.
- [25] M. Burtscher, and I. Szczyrba, "Numerical Modeling of Brain Dynamics in Traumatic Situations — Impulsive Translations," in *Proc. 2005 Conf. on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, pp. 205-211, CSREA Press 2005.

XML in Health Information Systems

Justin Brewton, Xiaohong Yuan, Francis Akowuah

Department of Computer Science, North Carolina A&T State University, Greensboro, North Carolina, USA

Abstract

Advancing technologies in the healthcare industry has led to the idea of an electronic health record. This form of document will allow healthcare institutions to store patient information more efficiently. The technology that allows hospitals to create such a document is XML. This paper discusses the emergence of XML in the healthcare field and also the HL7 standard, which provides guidelines for the creation and sharing of these documents. Also discussed will be current issues regarding securing the XML language.

Keywords

Health information systems, security and privacy, XML, HL7

1. Introduction

The development of the Hypertext Markup Language (HTML) brought about a significant change in the way electronic documents were exchanged. The flexibility and simplicity of the language was key part in the growth of the World Wide Web. HTML focuses on separating text information from presentation information through the use of a tagging system. As websites became more widespread, the shortcomings of HTML began to be exposed. The major problem was that HTML had no means of representing structured data. Data elements that had a hierarchical relationship could not be efficiently represented in the language. In an effort to mitigate these problems, the Extensible Markup Language (XML) was created [11].

Initially XML was to take the place of HTML as the norm for the exchange of data and documents over the internet. However, HTML remained the standard for internet exchanges and XML found it's calling in facilitating exchanges in transaction-based systems and various other disparate systems. XML is considered a meta-language, meaning that it can be used to define a language [11]. A user constructs a new language by creating custom tags that are tailored for the type of data being manipulated.

In recent years, there has been a rapid increase in the development of health information systems motivated by legislation intended to protect patients' information and privacy, and the government's interests in reducing the cost and improving the quality of healthcare. Electronic health record allows healthcare institutions to store patient information more efficiently. XML has become a basic

technology for implementing electronic health record and health information systems.

This paper introduces the basics of XML, and discusses Health Level 7 (HL7), an organization that sets standards. The Clinical Document Architecture (CDA) defined by HL7 is introduced, which provides guidelines for the creation and sharing of electronic health records. Current issues regarding securing the XML language is also discussed.

This paper is organized as follows. Sections 2 and 3 introduce the basics of XML and the advantages and disadvantages of XML. Section 4 discusses the history of patient records. HL7 is introduced in Section 5. Section 6 discusses security issues in XML and Section 7 concludes the paper.

2. XML Basics

The creation of an XML document may consist of three parts. The first of which is the data type definition (DTD). This layer describes the version of the data format, element descriptions, data structures, and some of the restrictions placed on the data. Essentially the overall format of the document is specified by the DTD. Here is a very simple example of a DTD that could hold a list of basketball players on a team:

1. <!ELEMENT player_list (player) *>
2. <!ELEMENT player (name, age, school? , country)>
3. <!ELEMENT name (#PCDATA) >
4. <!ELEMENT age (#PCDATA) >
5. <!ELEMENT school (#PCDATA) >
6. <!ELEMENT country (#PCDATA) >

Line one says that player_list is a valid element name and any instance of such element contains any number of player elements. The * signifies that there can be 0 or more player elements within the player_list element. The next line states that player is a valid element and any instance of this element should be followed by elements of type name, then age, then school (optional), and finally country. The ? character following an element signifies that the element is optional. Lines three, four, five, and six merely declare the elements name, age, school, and country as valid element types. The tag (#PCDATA) stands for parsed character data, meaning that the data is taken from what is entered by the author of the document. The following is an example of a document that conforms to this DTD:

```
<?xml version="1.0" encoding="UTF-8"
standalone="no"?>
<!DOCTYPE people_list SYSTEM "example.dtd">
<player_list>
  <player>
    <name>John Hooper</name>
    <age>23</age>
    <country>USA</country>
  </player>
</player_list>
```

The second part of the document is a detailed explanation of what the user created tags mean. The last layer of an XML document, which is optional, defines how the information will be presented [1]. Documents can be linked to use CSS or XSLT style sheet.

3. Advantages and Disadvantages of XML

The advantages of XML make it a viable solution to many of the data exchange problems that plague modern systems. There are many advantages to using a language like XML, but the major ones are:

- The ability to support user created tags allows the language to be fully extensible and void of any type of tag limitations. Since the language does not actually “do” anything, compatibility between systems is not an issue. As long as both systems can support the XML application that actually uses the document then the exchange of data is possible.
- Another key advantage of XML is its versatility. Any type of data can be modeled and tags can be created for very specific contexts.

There are also limitations to the XML that must be considered, such as:

- The lack of powerful applications that can process XML data and actually make the data useful is a primary disadvantage of the language. Only in recent history have browsers began to have the ability to read XML. Even now, these browsers still make use of HTML to render the XML document. This means that as of now, XML cannot be used as a language that is independent of HTML.
- Another disadvantage of XML results from the unlimited flexibility of the language. The tags implemented in a document are solely chosen by the creator. There is not a standard or generally accepted set of tags to be used in an XML document. As a result

of this, designers can not just create general applications because each company will invariably have their own set of special tags and unique meaning for those tags.

4. History of Patient Records

The majority of medical institutions initially used paper to record various transactions that occurred. Doctors used and still use the traditional pen and pad to record any medical notes about a patient. The notes included general observations, possible diagnosis, and information about any follow up visits that need to be scheduled. In addition to medical notes, medical centers also needed to keep financial information about each patient for billing purposes. When considering the potentially high number of patients a doctor’s office or hospital could encounter, the cost of materials to store their records could easily reach a very high value.

The first step to solving the cost problem was to incorporate information technology into the health care industry. Offices began to electronically deal with back-office operations such as billing. Dramatic reductions in cost resulted from this shift towards the use of electronic business systems. The success of the electronic business model sparked an even stronger focus on finding ways to integrate the latest technologies in information systems. The next major advancement was the creation of a system to digitize the process of Admitting, Discharging, and Transferring of a patient (ADT). These ADT systems provided health care facilities the ability to not only locate patients but also keep an accurate count of them [10].

The next logical step to create a fully digitalized health care system is to develop an electronic system that is capable of storing a patient’s entire health history. It is at this point that the idea of the Electronic Health Record (EHR) becomes the focus of research. Imagine an electronic record that displays a patient’s lab results, billing information, allergies etc. This type of record would serve to minimize costs and medical errors while increasing data accuracy and integrity. The ultimate goal of all this work on EHR is to create a system where information can be shared between patients and medical institutions and also back and forth between independent medical practices. This system does not require some huge data center because each practice will store its information remotely. However, there is a need for a standard that details how EHRs should be formatted.

5. Health Level 7 (HL7)

Health Level 7 (HL7) is an organization that sets standards and is accredited by the American National Standards Institute. This group is responsible for many communication standards used across America. Some of the standards created by this organization consist of:

- Arden Syntax – a grammar for representing medical conditions and recommendations
- Structured Product Labeling – the published information that accompanies a medicine
- Clinical Context Object Workgroup – an interoperability specification for the visual integration of user applications
- Claims Attachments – a standard health care attachment to augment another healthcare transaction

Their goal for the healthcare field is to provide standards for the exchange, management and integration of data that support clinical patient care and the management, delivery, and evaluation of healthcare services. In addition to creating messaging standards HL7 is also working on developing standards for the representation of clinical documents such as discharge summaries and progress notes. As a whole, these standards collectively make up the HL7 Clinical Document Architecture (CDA) [2].

The CDA aims at solving the previously discussed problem of finding a reliable and standardized means of storing and exchanging clinical documents. By specifying a mark-up and semantic structure through XML, the architecture works toward creating a universal way of allowing medical institutions to share clinical documents.

5.1 Clinical Document

A clinical document is defined as having these qualities:

- *Persistence* – A clinical document remains in an unaltered state for a user specified amount of time
- *Stewardship* – An entrusted person or party must have the responsibility of maintaining the document
- *Potential for authentication* – The document is intended to be legally authenticated
- *Wholeness* – Authentication applies to the whole document and not to just portions of the information

- *Human readability* – A clinical document should be human readable

5.2 Reference Information Model (RIM)

Currently HL7 version 3 is being developed. This family of standards includes The Clinical Data Architecture as well as rules for messaging. The newly developed version 3 allows clinical documents to contain not only text but also images, sounds, and other types of multimedia [3]. Both standards are implemented with XML and are derived from the Reference Information Model. The Reference Information Model or RIM is an object-oriented graphical depiction of clinical data and aids to understanding the life-cycle of events that messages and documents go through [6]. It focuses on five major themes:

- Ensure coverage of HL7 version 2.x. It ensured that it included all the information content of HL7 version 2.x.
- Remove unsubstantiated content from the model. It removed content from the draft that the technical committee did not originate and could find no rationale for retaining.
- Unified service action model (USAM). It introduced a concise, well-defined set of structures and vocabularies that address the information needs of a wide variety of clinical scenarios.
- Ensure quality. It addressed inconsistencies in the draft model and conflicts between the model and the modeling style guide.
- Address the "left hand side" of the model. It introduced powerful structures and vocabularies for the non-clinical portions of the model (patient administration, finance, scheduling).

Figure 1 [4] shows an example of RIM represented in graphical form.

5.3 The Hierarchical Structure of CDA

The actual architecture of the CDA can be thought of as a set of hierarchically related XML Document Type Definitions. As of now, only the top node, known as Level One, has been defined. Level one is designed to include enough detail to mark up narrative clinical notes. The objective of this level is to ease users into RIM. It is intentionally not very complex to allow deeper levels the ability to mark up the document even more. As seen in Figure 2 [6], level one material consists of the raw data gathered from an encounter. There are no high level medical codes or terminologies used [8].

Level Two, which has not been developed, will be a set of templates that can be layered on top of Level One. Level Two is envisioned to provide constraints to documents by

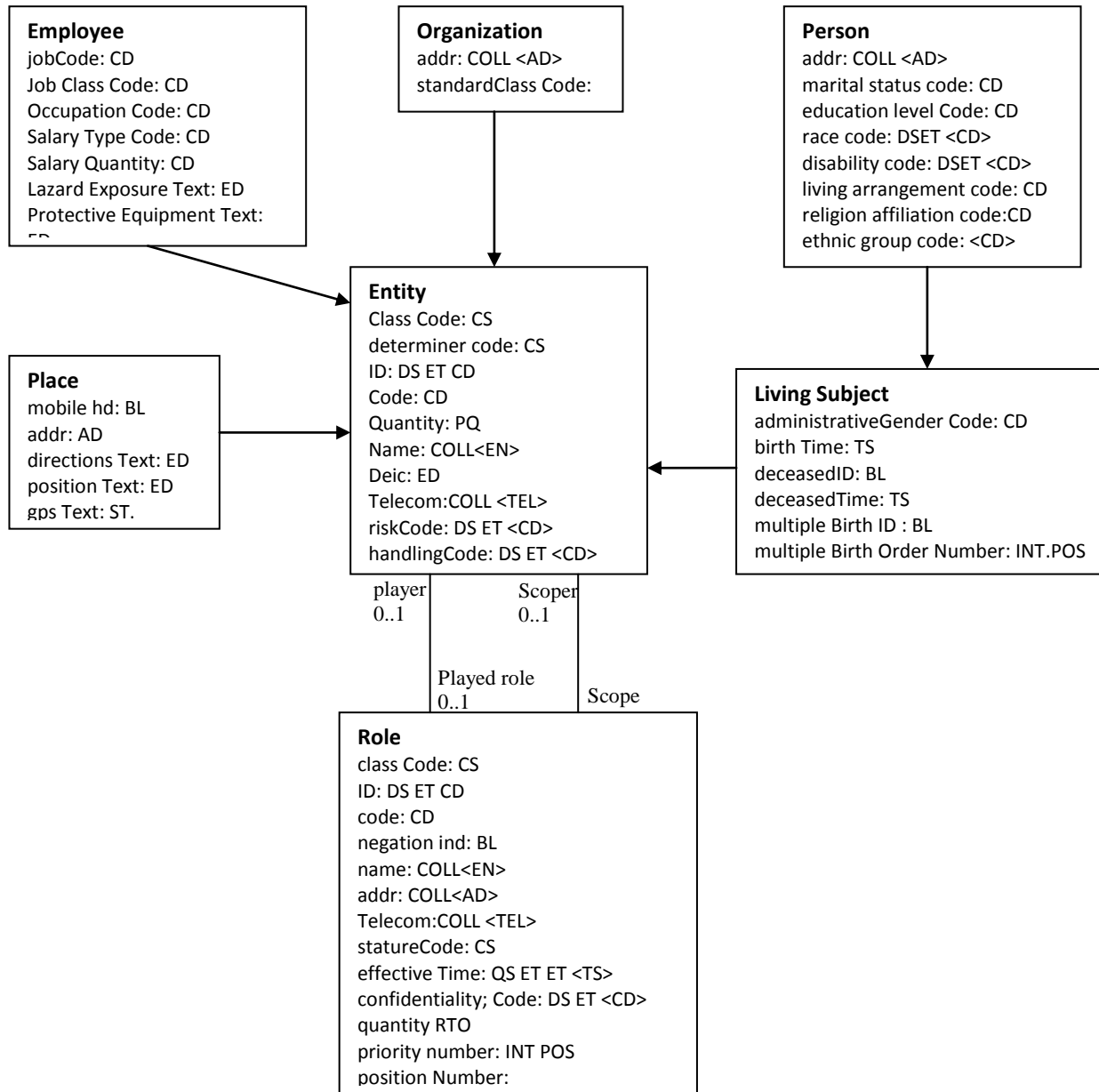


Figure 1. An example of Reference Information Model

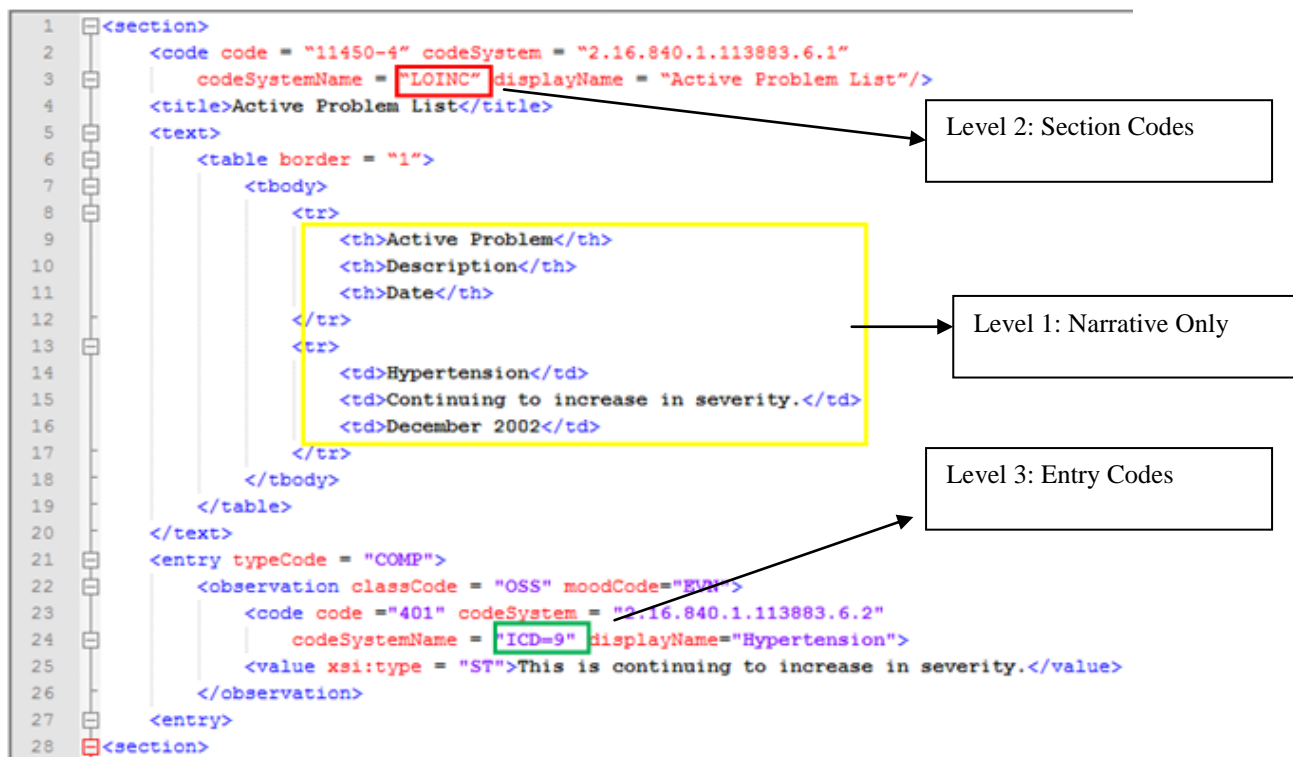


Figure 2. The hierarchical structure of CDA

requiring that specific types of medical documents contain a certain piece of information. For example one of these templates might require that a document of type “blood work” requires an “insulin level” section. As imagined, this type of structure would necessitate input from various professional groups to come up with an agreed upon template. Level Three information consists of specific medical codes used by healthcare institutions. A proper structure for information of such a deep level will require extensive collaboration between healthcare offices worldwide and the HL7 group.

5.4 CDA Document Structure

CDA documents are composed of a header and a body. The header is used to describe the context in which the document was created. CDA document headers serve three purposes:

1. Make document exchange possible within the same institutions and between separate institutions
2. Facilitate document management
3. Facilitate the compilation of an individual’s complete medical history

The body of the document is made up of paragraphs, lists, and tables. Each of these sections can contain data, medical codes, and multimedia that describe the patient health care based transactions [4].

6. XML Security

When considering the security of XML documents, all the traditional qualities are desired: integrity, confidentiality, authorization etc. To achieve these goals XML data is treated much like any other types of data, in terms of security. XML data that is used to make up an individual’s health record must be secure [9]. A patient’s electronic health record could be potentially sent to many different institutions to be viewed by various doctors. Patients need to be sure that their personal information is only seen by an authorized party.

6.1 Digital Signatures

XML Signatures operate identically to regular digital signatures [5]. A signature contains three sections:

- **SignedInfo:** Contains information about what part of the document is actually signed.
- **SignatureValue:** This is the output of the encryption of the data. It is the actual digital signature.
- **KeyInfo:** Provides the key or information on finding the key that validates the signature.

An XML signature allows the signing of a whole or specific section of a document. This standard provides integrity, message authentications as well as authentication for the signer of the document. Consider a patient that has been instructed by her physician to see a cardiologist. If the cardiologist makes any changes to the patient's EHR, only the section changed should be digitally signed by the cardiologist.

6.2 XML Encryption

The recommended encryption techniques to provide confidentiality for XML documents are not a replacement for security protocols such as SSL/TLS. Instead, XML encryption mandates requirements for areas not covered by SSL. More specifically XML allows for certain parts of the data to be encrypted and also provides security for sessions between more than two parties. Along with those two new areas covered, XML encryption still provides the traditional encryption methods. The need for an efficient encryption method, when dealing with healthcare documents is evident. The ultimate goal of this EHR revolution is to facilitate the exchange and storage of medical information. As new technologies make these tasks easier, the measures for securing this type of sensitive information must be strengthened.

6.3 Attribute Based Encryption (ABE)

Attribute Based Encryption (ABE) is an encryption method that works well with XML. ABE allows only users who have a specific set of attributes, which also match with the attribute set associated with a message, to decrypt the contents of that message. Just like with the traditional Identity Based Encryption method, user will be assigned a secret key by a central authority. However, the ABE secret key is based on the specific attributes of each user. When messages are created, the author creates a policy that corresponds with the ciphertext. The policy is just a Boolean statement that specifies the attributes a user must have to decrypt the information [7]. ABE secures message passing between separate entities. The following example explains the ABE method.

- Mary and John work for a company
- Mary is a *sales manager* and John work in the *IT department*
- When each employee's private key is assigned by the authority, the key contains attributes about their position (Mary- *Sales AND Manager* ; John – *IT*)
- A message is sent with a policy that maintains that only worker in the IT department are allowed to view it
- John's private key fulfils the policy, as a result, his key can decrypt the message. Mary is unable to view the message because her private key attributes do not satisfy the policy of the message.

When considering ABE's application to electronic health records, think about a patient that has all of his medical history contained in one document. After a routine visit of his regular physician an appointment to see a dermatologist is made because of a rash found on the patient's arm. As a result some of the patient's medical history needs to be sent to the dermatologist's office. By making use of ABE, the patient's entire medical record can be sent and the patient can be assured that only the dermatologist is able to view his personal information. To accomplish this, the record needs to be sent with a policy that allows only the individual with the dermatologist's credentials to view the document. In addition, ABE can also be used to guarantee that only information pertinent to the skin problem can be seen by the dermatologist. To achieve this all the information that the dermatologist is not allowed to see should be encrypted with a key that is different from the one used with the ABE.

7. Conclusion

It seems that in the future, the way medical information is stored and shared between institutions will be revolutionized. The extensibility and versatility of XML will be used as a catalyst for this advancement. The overall goal of creating an environment in which medical institutions can freely share information is far from becoming a reality but it is not impossible. The most important key for achieving this goal will be creating a standard for the storage of documents and for the method of sharing these documents between independent medical entities.

Organizations, like the HL7 group are essential to this process. As with many new technologies, as more people began to make use of this system, the necessity for creating a secure environment will increase. Equally important as creating a structure for storing and sharing medical information is the issue of securing this information. As developers continue to create more powerful process and actually make use of XML data, the usefulness of electronic medical applications will grow.

Acknowledgements

This work is partially supported by NSF under grant HRD-1137516, and by Department of Education under grant P120A090049. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and Department of Education

References

- [1] Wasim A Al-Hamdani, "XML Security in Healthcare Web Systems," *Information Security Curriculum Development Conference*, Kennesaw GA, 2010.
- [2] Dolin RH, Alschuler L, Behlen F, et al. HL7 document patient record architecture: an XML document architecture based on a shared information model. *Proc AMIA Symp.* 1999:52–56.
- [3] Dolin RH, Alschuler L. Approaching Semantic interoperability in Health Level Seven. *Journal of the American Medical Informatics Association*; Jan 2011, Vol. 18 Issue 1, p99.
- [4] Corepoint Health. CDA Architecture. [Online]. <http://www.corepointhealth.com/resource-center/hl7-resources/hl7-cda>.
- [5] ABB Corporate Research. Standards for XML and Web Services Security. [Online]. <http://www.tik.ee.ethz.ch>.
- [6] HL7 Standards. CDA Levels of Interoperability. [Online]. <http://www.hl7standards.com/blog/2011/06/02/cda-levels-of-interoperability/>
- [7] Jin Li, Man Ho Au, Willy Susilo, Dongqing Xie, and Kui Ren. 2010. Attribute-based signature and its applications. In *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security (ASIACCS '10)*. ACM, New York, NY, USA, 60–69. DOI=10.1145/1755688.1755697 <http://doi.acm.org/10.1145/1755688.1755697>
- [8] Chien-Tsai Liu, Ann-Ging Long, Yu-Chuan Li, Kuo-Ching Tsai, Hsu-Sung Kuo, "Sharing patient care records over the World Wide Web," *International journal of medical informatics* 1 May 2001 (volume 61 issue 2 Pages 189-205)
- [9] Hsiao, Tsung-Chih. *A Secure Integrated Medical Information System*. *Journal of Medical Systems*. 2011.
- [10] Gupta, V. and Murtaza, M. B. (2009), "Approaches To Electronic Health Record Implementation", *The Review of Business Information Systems*, 13(4), pg. 21
- [11] Seals M. The use of XML in healthcare information management. *J Healthc Inf Manag.* 2000 Summer;14(2):85–95.

Accurate Proton Beam Localization

Y. Chen¹, E. Gomez¹, F. Hurley², Y. Nie², K.E. Schubert¹, R. Schulte²

gracenumerical@gmail.com, ernesto@csusb.edu, ford.hurley@gmail.com,
yingnie@dominion.llumc.edu, keith@r2labs.org, rschulte@dominion.llumc.edu

¹School of Computer Science and Engineering, California State University, San Bernardino,
5500 University Parkway, San Bernardino, CA 92407 USA

²Department of Radiation Medicine, Loma Linda University Medical Center,
Loma Linda, CA, 92354, USA

Abstract—*The goal of this work was to further improve an experimental proton radiosurgery system at Loma Linda University Medical Center to reach sub-millimeter accuracy before proton radiosurgery with narrow beams can be used in a clinical trial. Radiosurgery precisely targets a specific anatomical region with high doses of radiation. We have developed a program that provides correcting translational offsets during target rotation and allows the proton beams to be aimed at the target from multiple directions in the proton research room. This was accomplished by developing and testing advanced image analysis software tools. The targeting accuracy was determined with a commercial stereotactic performance phantom. It was found that sub-millimeter targeting accuracy can be achieved with the current system.*

Keywords: localization, proton radiosurgery, stereotactic target localization, image analysis, sub-millimeter targeting accuracy

1. Introduction

Stereotactic radiosurgery (SRS) is a radiation therapy method that precisely delivers very high dose of external radiation to well-defined targets in the brain or to the target within the body. SRS has many advantages over open surgery. Since there is no incision with this method, there is no risk of bleeding, infection or other possible surgical complications. [2]

Functional radiosurgery is a sub-specialty of SRS that creates small lesions in an area of diseased brain that interrupts pathological functions, such as abnormal movement or pain, one can treat functional disorders. Diseases that are currently treated with functional radiosurgery include, trigeminal neuralgia, Parkinson's disease and essential tremor.

The use of protons for functional radiosurgery will be advantageous when the lesion is in close proximity to critical neural structures. The methodology for functional proton radiosurgery is currently being developed at Loma Linda University Medical Center (LLUMC). With a margin of error of 1-2 mm, there is a high risk of delivering a dose to the incorrect location, which could result in serious complications for the patient. Working in such close proximities to critical

brain features provides very little margin for error. Therefore, a very accurate method for stereotactic target localization must be developed and tested prior to deployment of a new system for functional proton radiosurgery in human patients. An experimental platform for testing these new methods has been built and is used to develop and test new methods of beam localization and verification.

Previous work centered on methods for accurately aligning the target to the proton beam using feedback from a room-fixed camera-based system [3], [4], [5]. The primary goal of the present work was to test a different strategy to improve proton beam targeting accuracy for proton functional radiosurgery by using a system that relies on accurate characterization of 3D-stage movements and rotation relative to a fixed proton beam.

An important task within the work, described in this paper, was to develop an algorithm that will deliver proton beams from multiple directions to the target without the need for checking correct alignment before each beam delivery. The performance of the algorithm was verified by analyzing radiochromic films embedded in a quality assurance phantom (Lucy®, Standard Imaging Inc.). In addition, development of a user-friendly software interface was an important subject of the present research work.

2. Approach

2.1 Experimental Setup

Unlike the proton treatment rooms at LLUMC with their 90-ton, three-story gantries that can be rotated 360 degrees to deliver the proton beams at any angle prescribed by the physician, a research platform for phantoms and small animals (rats) was built and mounted on one of the fixed horizontal proton beam lines that deliver the proton beams through evacuated steel tubes (beam pipes) into the proton research room.

The research system simulates the functional proton radiosurgery treatment, which in the future will take place in one of the proton treatment rooms. The stereotactic setup consists of one rotational and three translational micro-stages (Newport Corporation). The rotational stage rotates the

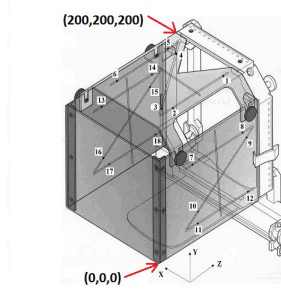


Fig. 1: Leksell coordinate frame and Leksell CT indicator frame.

stereotactic system [1] around an axis that is approximately parallel to the z axis of stereotactic system, thus, simulating the rotation of the proton beam around the gantry axis in the treatment room. The translational stages align the target to the beam axis in longitudinal, horizontal and vertical direction. The program developed within this research is able to control the rotational micro-stage, acquire the position of the target from the stereotactic localization software, and to perform translational moves to bring a preselected target into alignment with the beam axis. The performance of this system was tested using the Lucy phantom.

2.2 Stereotactic and Stage Coordinate Systems

The main goal of this research was to develop and test methodology to accurately align the proton beam to a planned target in a stereotactic coordinate system.

The stereotactic coordinate system is defined by the Leksell coordinate frame (Electa), an instrument often used for clinical stereotactic radiosurgery, and the CT-indicator frame, which used with computed tomography (CT) to define the stereotactic coordinates of the target (see Figure 1). The stereotactic coordinate system is a right-handed Cartesian system. When the Leksell coordinate frame is mounted on a human head, the positive x-axis points from the patient's right to the left, the positive y-axis from the back of the head to the nose, and the positive z-axis from head to feet. The Leksell coordinate frame is engraved with a rectilinear coordinate scale, in which the origin (0, 0, 0) is located superior, lateral, and posterior to the frame on the patient's superior right side. The coordinates are expressed in millimeters. The center of the Leksell CT indicator frame is at stereotactic coordinates (100, 100, 133) mm.

The stage coordinate system is defined by three orthogonal translational stages, which move the stereotactic system relative to the beam axis. Imagine one stands in front of the stage system and the proton beam comes from the left side, refer to Figure 2. The positive x-axis (longitudinal translational stage) coincides with the stereotactic z-axis, the positive y-axis points to the opposite direction of the

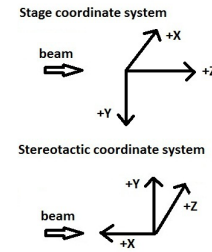


Fig. 2: Coordinate system of micro-stage and stereotactic coordinate system.

stereotactic y-axis, and the positive z-axis points to the opposite direction of the stereotactic x-axis.

In the proton research room, the radiosurgery cart is aligned to the proton beam line. When correctly aligned, the proton beam line passes through the center of the Leksell CT indicator frame, which has stereotactic coordinates (100, 100, 133) mm and stage coordinates (-6.5, 39.5, 0) mm.

2.3 Alignment Software

In order to relate the stereotactic coordinates of a target and corresponding micro-stage coordinates that will align the proton beam to the target, a software algorithm which takes into account the different orientation and relative position of these two systems was required. Assuming the stage coordinates of the home position are (h1,h2,h3) and the corresponding stereotactic coordinates are (s1,s2,s3). The following transformation performs this task.

$$matrix = \begin{bmatrix} 0 & 0 & 1.0000 \\ -0.0034 & 1.0000 & 0 \\ -1.0000 & -0.0034 & 0 \end{bmatrix} \quad (1)$$

$$vector = (h1, h2, h3) - (s1, s2, s3) * matrix \quad (2)$$

$$(h1, h2, h3) = (s1, s2, s3) * matrix + vector \quad (3)$$

To prevent a user from accidentally changing the main program, the home position coordinate and the associated stereotactic coordinates are stored in an external text file. The beam axis was carefully aligned to the stereotactic coordinate system so that it is parallel to the stereotactic x-axis, represented by the vector (-1,0,0). The translational stages can only be moved from -50 mm to +50 mm in x- and z- direction and from -8 mm to +93 mm in y direction. In order to detect whether the movement is beyond these limits, the program first calculates all required translational corrections and validates them against the limitations. If any correction is beyond the translational limits, the user is alerted with a warning message. The translational corrections along the beam direction (z-axis of the stage system) are ignored to prevent collision of the object with the collimator. The color code for "Off limits"(stage could not performed move) and "Error"(stage did not reach destination) is red;

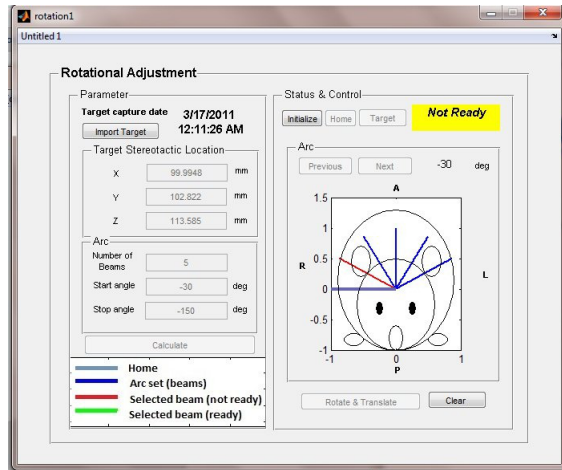


Fig. 3: Rotational system GUI.

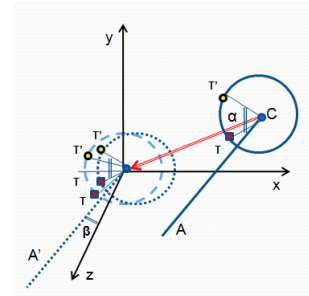
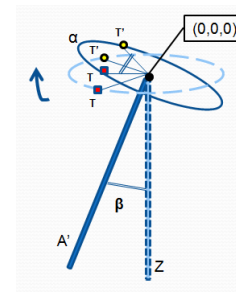
the color code for "Not ready"(moves not yet complete) is yellow, and the color code for "Ready"(all moves completed) is green.

The input parameters of the alignment software includes stereotactic target coordinates, the number of beam angles, and the start and stop angles, so that all the beams can be rapidly delivered in sequence. The software was written to that the user can go to the preselected beam angles in arbitrary sequence. In the home position, the rat is in the orientation shown in Figure 3. This corresponds to an absolute internal rotation of +90 degrees. The stage can rotate counterclockwise only up to an angle of -170 degrees and clockwise up to +360 degrees (a total of 530 degrees range). In order to perform the stage rotation from the home position by an angle of ϕ degrees, the stage needs to be programmed to rotate to an absolute angle of $\phi + 90$ degrees if $\phi > -260$ degrees and $\phi + 450$ degrees otherwise. The GUI shows the beam location relative to the rat, refer to Figure 3. Legend of beam indicator: The beam location at home position, usually at 0 degrees, is indicated by a grey line; the pre-set beam angles is indicated by a blue line, and the selected beam line is indicated by a red line, and once the system is ready changes to a green line.

2.4 Method to Calculate Translational Corrections

This section contains a mathematical description of the translational corrections required after the rotation has been made, one needs to perform a mathematical 3D rotation and calculate the 3D vector that shifts the rotated target point back to the beam axis. In case the rotational axis is parallel to the z-axis, only a 2D rotation in the xy-plane is required and the correction vector becomes a 2D vector.

A 3D rotation describes the motion of a rigid body around a fixed axis in 3D space, while a 2D rotation describes the

Fig. 4: Geometry of the mathematical steps to perform a target rotation by angle α around a micro-stage axis A .Fig. 5: Once the micro-stage rotational axis has been shifted to the origin, it needs to be aligned with the stereotactic z-axis before applying the 2D rotation by angle α .

motion of a rigid body around a fixed point in a 2D plane. It is mathematically convenient to perform a 3D rotation about any axis in space by first making the rotational axis coincide with one of the axes of the coordinate system and then to perform a 2D rotation about that axis.

In the present experimental setup, the rotational micro-stage axis is only approximately parallel to the z-axis of the stereotactic reference system. To calculate a 3D rotation of the target point T around the rotational micro-stage axis A by an angle α mathematically, one first shifts the axis point C that has the same z-coordinate as the target point to the origin of the stereotactic reference system. The rotational axis A' now intersects the origin, but is still rotated by the angle β relative to the z-axis, see Figure 4. The next steps are to align the rotational axis A' with the stereotactic z-axis and then to apply the 2D rotation with angle α , see Figure 5.

One now needs to find the translational vector that shifts the target point back to the fixed beam axis. Instead of rotating the target point, one can also rotate the beam axis, keeping the target point fixed, see Figure 6. The translational correction is represented by the vector v_3 which shifts the target point back to the beam axis. The components of this vector should be expressed along the axes of the translational stages which perform the translational corrections. When performed correctly, a series of rotated beam axis should intersect at the target point, forming a star pattern.

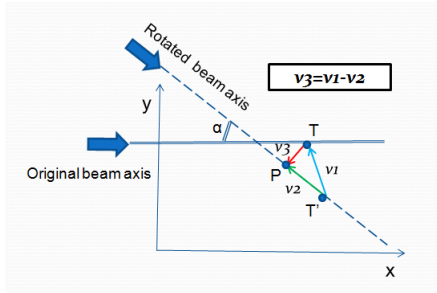


Fig. 6: Geometric representation of the translational correction. The beam axis is rotated by angle α relative to the target point T . After rotation, the point on the beam axis originally intersecting T is now T' . P is the projection of T onto the new beam axis. The translational correction v_3 is calculated as shown.

A program was developed implementing an algorithm that calculates the translational corrections in the stage coordinate system. The algorithm initially assumed that the z-axis is the rotational axis at the beginning, but the result was not ideal. The stereotactic coordinates of the rotational axis were further defined by another experiment and image analysis. The details of this will be published in the thesis of one of the authors. With the defined rotational axis, the structure of the algorithm of the translational corrections is as follows:

- 1) Find the point on the rotation axis that intersects the xy-plane containing the stereotactic target point (i.e., has the same stereotactic z coordinate). Since the z-axis is practically perpendicular to the xy-plane, the intersection point will be the point on the axis closest to the target.
- 2) Find the translation vector that shifts the point found in step 2 to the origin of the stereotactic reference system (SRS) and shift the beam axis point (target point) by adding the same vector to the coordinates of the target point.
- 3) Find the 3D rotation matrix MA that aligns the horizontal rotation axis with the stereotactic z-axis and apply this rotation to the beam axis point and vector.
- 4) Perform the stage rotation for a given angle by using a 2D rotation matrix in the xy-plane and apply it to the beam axis point and beam axis vector.
- 5) Apply the inverse rotation matrix MA and then the inverse translation vector to the shifted and rotated beam axis point and vector found in step 4. This will represent the new beam axis location in the SRS coordinates system.
- 6) Find the vector that represents the shortest distance from the stereotactic target point to the beam axis and convert its components to correctional shifts of the translational stages by applying the reverse rotation matrix.

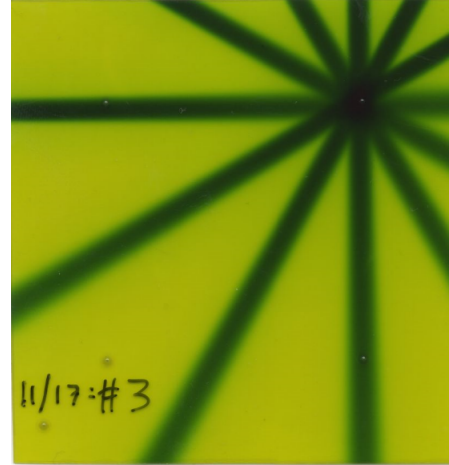


Fig. 7: Six beams from 0 degrees to -150 degrees, aiming at the upper right pin of the Lucy phantom cassette. The location of the marker pin is also visible on the film.

3. Performance Study

To verify the stereotactic targeting accuracy, narrow proton beams were imaged with a radiochromic film (Gafchromic EBT2 film, International Specialty Products), embedded in the Lucy phantom and to find the beam axis in relation to target points, also visible on the film. The 5 steel pins of 0.5 mm diameter, which hold the radiochromic film in place were used as stereotactic targets. The advantage of using these targets was that they could be seen in CT localization images and also created visible perforation holes in the film. An image analysis method was developed to support the data analysis of the performance study.

Each proton beam produces a dark footprint with lateral penumbra on the radiochromic film. In order to digitize the beam image meaningfully and efficiently, the intensity image was converted to a binary image. Then the MatLab Canny Edge Detector was applied to define beam edges in the beam penumbra by looking for local maxima of the intensity gradient of the image. Once the two edge lines of each beam path were found, the beam axes equations were determined by averaging parameters of the edge lines.

In addition to finding the beam axis, it was also necessary to define the location of target pins on the film, which were represented by small pin holes. Besides from serving as target points, they can also be used to determine the length scale (pixel per mm) of the digital image, as four of the five pins in the Lucy phantom form a square with 40 mm side length.

4. Conclusion

The results of the initial image analysis of a proton beam star pattern were not ideal. It was found that the initial assumption that the rotational axis was exactly parallel to

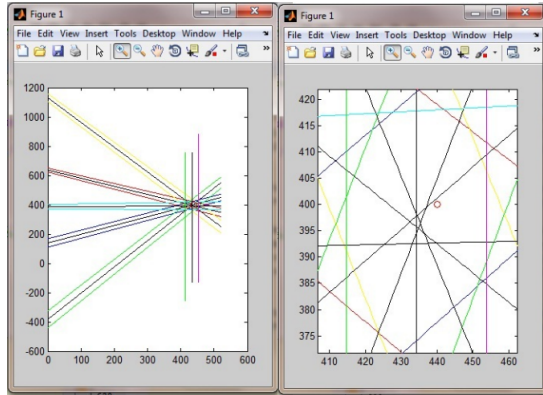


Fig. 8: Beam axes pattern resulting from analysis of Figure 7. The right half of image shows a close-up view of the beam-axes intersection region. 1 pixel = 0.0844 mm

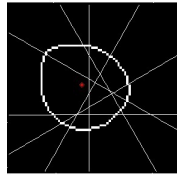


Fig. 9: Results of the six beams image. The largest raw distance from the target to the beam path is 0.211 mm

the z-axis. After the z-axis was determined more accurately by studying the change of a the proton beam position relative to the radiochromic film during a series of discrete rotations, localization accuracy was much improved to better than 0.32 mm. The average target error is significantly better at $0.149 \text{ mm} \pm 0.058 \text{ mm}$, which demonstrates the high repeatability of this method.

During this research, advanced methods of image analysis of beam patterns visualized with radiochromic films were developed. This included a consistent definition of marker pin holes, the use of a high-resolution algorithm for edge definition, definition of the best practice for film scanning, and methods for taking into account the scanner distortion. Most importantly, a rotational correction program was developed that performs translational corrections after each stage rotation. This allows fast and accurate beam delivery from many consecutive directions. A user-friendly GUI for this program was also developed.

References

- [1] Brown, R., Roberts, T., Osborn, A.: Stereotaxic frame and computer software for ct-directed neurosurgical localization. *Invest Radiol* **15**(4), 308–312 (1980)
- [2] Chin, L., Regine, W.: *Principles and Practice of Stereotactic Radio-surgery*. Springer (2008)
- [3] Gomez, E., Karant, Y., Malkoc, V., Neupane, M.R., Schubert, K., Schulte, R.W.: Orthogonal and Least-Squares Based Coordinate Transforms for Optical Alignment Verification in Radiosurgery. In: *Proceedings ITCC*, vol. II, pp. 83–88. IEEE, Los Alamitos, Ca (2005)
- [4] Schulte, R.W., Shihadeh, F., Schubert, K.E.: Performance study of an optoelectronic localization system for functional proton radiosurgery. *International Journal of Radiation Oncology Biology Physics - INT J RADIAT ONCOL BIOL PHYS* **69**(3), S712 (2007)
- [5] Shihadeh, F., Schulte, R., Schubert, K., Chakrapani, P.: Performance analysis of an optoelectronic localization system for monitoring brain lesioning with proton beams. In: *29th IEEE EMBS Annual International Conference* (2007)

The HL7 CDA-Based Electronic Forms for Physical Examination

P.Y. Lee¹, P.Y. Hung¹, J.H. Lin³, A.J. Lee², S.T. Tang¹

¹Department of Biomedical Engineering, Ming Chuan University, Taoyuan, Taiwan

²Department of Health Information and Management, Ming Chuan University, Taoyuan, Taiwan

³Department of Electronic Engineering, Kun Shan University, Tainan, Taiwan

Abstract - In the developed country, it is necessary that the enterprise should arrange the occupational physical examination yearly, there are millions cases, results in a heavy loading to the hospital. The occupational physical examination is a moving process. But the information flow of the general hospital is not designed for the occupational examination. Then the physical examination is usually a paper-based process. As a result, it needs lots of clerks for data key in and confirmation. The paper-based operation is time and manpower consuming, and difficult in future expand and inter-discipline data exchange. The XML (eXtensible Markup Language) is the most popular format, which supports global data exchange, and could be embedded in Web service. Health level 7 (HL7) clinical document architecture (CDA) is a XML based format, which provides a standard form for digitizing a series of medical documents, and cross-discipline data exchange. In this study, we demonstrate the electronic forms for the occupational physical examination basing on HL7 CDA standard, and there are two styles, one is an Android application, and the other is a RIA (Rich Interactive Application) for general web browser.

Keywords: HL7 CDA, Physical Examination, Electronic Form, Android, RIA

1 Introduction

In the developed country, it is necessary that the enterprise should arrange the occupational physical examination yearly, there are millions cases, results in a heavy loading to the hospital. The general occupational physical examination is a moving process, the person undergo examination is moving site to site. The information flow of the general hospital is designed for patient treatment not for physical examination. Then the physical examination is usually a paper-based process. The traditional paper-based workflow is shown in Fig. 1. As a result, there need lots of clerks for data key-in, and additional human resources should be involved for data confirmation [1].

The paper-based operation is difficult in future expand and cross-discipline data exchange. The key problem of data exchange is heterogeneous data format. The XML (eXtensible Markup Language) is the most popular format for global data exchange. Additionally, XML could be embedded in Web service. As a result, the XML document is easy shared via internet, accessed through Web browser. Health level 7 (HL7) clinical document architecture (CDA) is a XML based format, which is constituted by medical objects, including text, image, and voice. And then provides a standard form for digitizing a series of medical documents, and cross-discipline data exchange [2, 3].

In this study, we demonstrate the electronic forms of the physical examination basing on HL7 CDA standard. The examined data and image are described by CDA level 3. It is not only going to transform the paper work to be a part of EHR (Electronic Health Record), but also is an extension of showing the richer medical contents, such as physiological signals, medical images and so on. The Android platform and RIA (Rich Interactive Application) are as well deployed to develop the electronic forms.

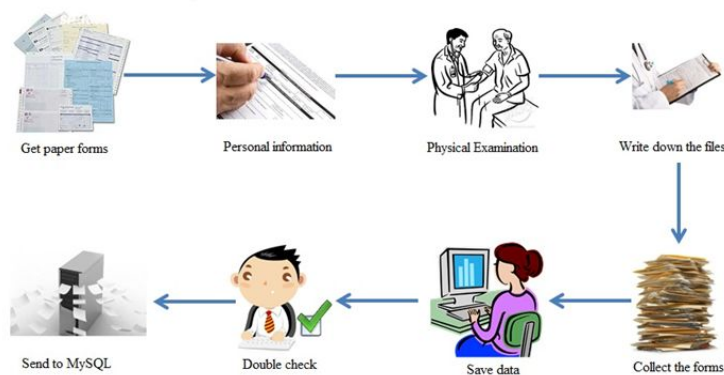


Figure 1. Paper-based working flowchart.

2 Methods

The developed examination forms are referring the current occupational physical examination procedure of Taipei

Veterans General Hospital, and the related examination procedure of other outsourcing medical laboratories. We use C#, RIA and Android SDK to develop the form, the result data would finally be transmitted to MySQL database. The developed data flow is show in Fig. 2. The data format is HL7 CDA compliant.

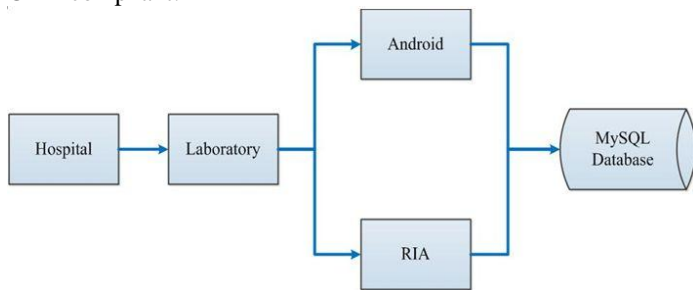


Figure 2. The developed data flow.

2.1 Android

Android is a Linux-based operating system developed by Google. In addition to the operating system, it also provides Android SDK/NDK application software development kit that would facilitate the development of Android applications [4]. Compared with the iPad operating system (iOS), any study or debug of the program are limited by Apple Inc. So it's more difficult to develop application in iPad operating system. Therefore, we utilize the Android operating system in our study. There are lots of facilities involved, include Eclipse, Java Development Kit (JDK), Android Development Tools (ADT), Android SDK, and ASUS TF101.

Eclipse is an open-source community that develops open platforms and products, and began as an IBM Canada project. Eclipse is a flexible environment to experiment with new computer languages or extensions to existing languages. The Java Development Kit (JDK) is an Oracle Corporation product aimed at Java developers. Since the introduction of Java, it has been by far the most widely used Java SDK. The ADT plug-in integrates the emulator into Eclipse so that it's launched automatically when run or debug projects. In case of not using the plug-in or want to use the emulator outside of Eclipse, that can telnet into the emulator and control it from its console. The Android SDK provides the tools and libraries necessary to begin developing applications that run on Android-powered devices [5]. ASUS TF101 is a tablet computer, which is a super slim profile of only 12.98 mm thick in a frame that weighs only 680 g. The ASUS TF101 is comfortable to hold from any position. This provides access to a full keyboard along with unique Android Function keys, turning the tablet Transformer into a full-fledged notebook..

2.2 RIA

C# is an elegant and type-safe object-oriented language that enables developers to build a variety of secure and robust applications that run on the .NET Framework. It can use C# to create traditional Windows client applications, XML Web

services, distributed components, client-server applications, and database applications, etc [6]. There are no separate header files, and no requirement that methods and types be declared in a particular order. C# programs run on the .NET Framework belong to a special non-mechanical code.

The .NET Framework is an integral Windows component that supports building and running the next generation of applications and XML Web services [7]. To provide a consistent object-oriented programming environment whether object code is stored and executed locally, executed locally but Internet-distributed, or executed remotely and to provide a code-execution environment that minimizes software deployment and versioning conflicts. The .NET Framework has two main components: the common language runtime and .NET Framework class library.

Common language runtime is the foundation of the .NET Framework, it can think of the runtime as an agent that manages code at execution time, providing core services such as memory management, thread management, and remoting, while also enforcing strict type safety and other forms of code accuracy that promote security and robustness. The class library is a comprehensive, object-oriented collection of reusable types that can use to develop applications ranging from traditional command-line or graphical user interface (GUI) applications to applications based on the latest innovations provided by ASP.NET [8].

2.3 MySQL

The MySQL database has become the world's most popular open source database, because of its high performance, high reliability and ease of use. It is also the database of choice for a new generation of applications built on the LAMP stack (Linux, Apache, MySQL, PHP/Perl/Python). Many of the world's largest and fastest-growing organizations including Facebook, Google and Adobe rely on MySQL to save time and money, and power their high-volume Web sites, business-critical systems and packaged software. MySQL runs on more than 20 platforms including Linux, Windows, Mac OS, Solaris, IBM AIX, which provide the kind of flexibility [9].

2.4 HL7 CDA

The CDA document is constituted by Header and Body. The Body includes StructuredBody and nonXMLBody. The root element is defined in <ClinicalDocument> [10], and there are three descending element levels. The Header is Level I, and StructuredBody includes Narrative block-Level II and Entries-Level III [11].

The deployed XML editor is Oxygen XML Editor 10 (Academic version). The Java CDA XML content is processed in Eclipse IDE. The Header is described by the logical observation identifiers names and codes (LOINC) [12]. There are three describe structure: 1. in <nonXMLBody> part, use <reference value="xx"> for target file index, 2. the out source

is coded by BASE64, and embeds in the <text> element of <nonXMLBody>, which constitutes the XML file, 3. <structuredBody> for content detail. Additionally, the <reference> and BASE64 codes should be integrated, and represented by XSLT (Extensible Stylesheet Language Transformations) technology, as shown in Fig. 3.

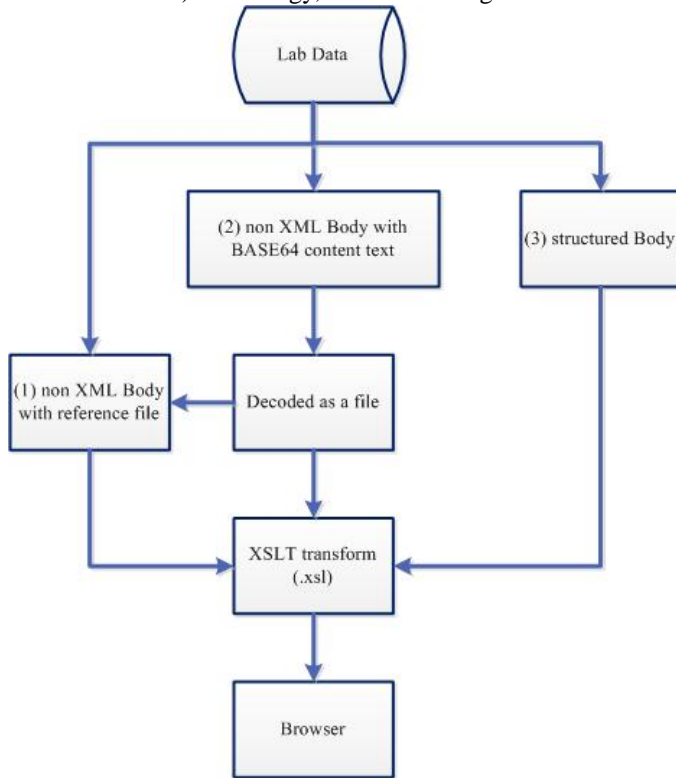


Figure 3. CDA body structure.

The CDA structure confirmation schema is composed by POCD_HD000040.xsd, datatypes.xsd, datatypes-base.xsd, NarrativeBlock.xsd, and voc.xsd, as shown in Fig. 4. The facility for confirmation is CDA Validator developed by Alschuler Associates LLC [13].

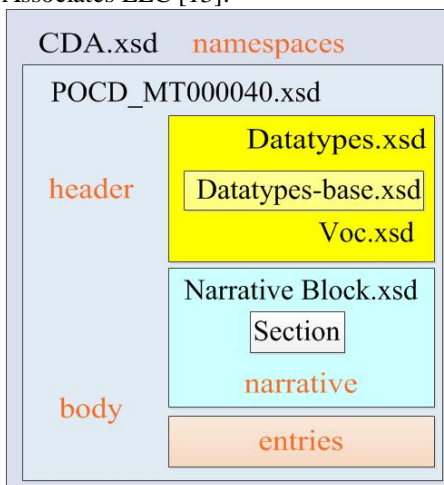


Figure 4. CDA structure confirmation blocks.

3 Results

We have developed the HL7 CDA compliant electronic form to replace the paper form, which saves a lot of time, and reduces the errors in key-in or paper lost. The developed workflow of electronic form is shown in Fig. 5. The Web GUI of the RIA application is shown in Fig. 6. The GUI of the electronic form in Android platform is shown in Fig. 7.

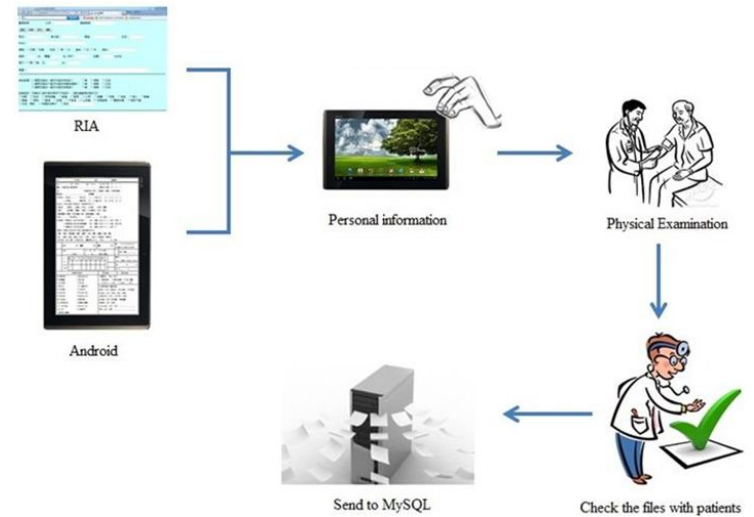


Figure 5. Electronic forms of physical checkup shows on RIA.

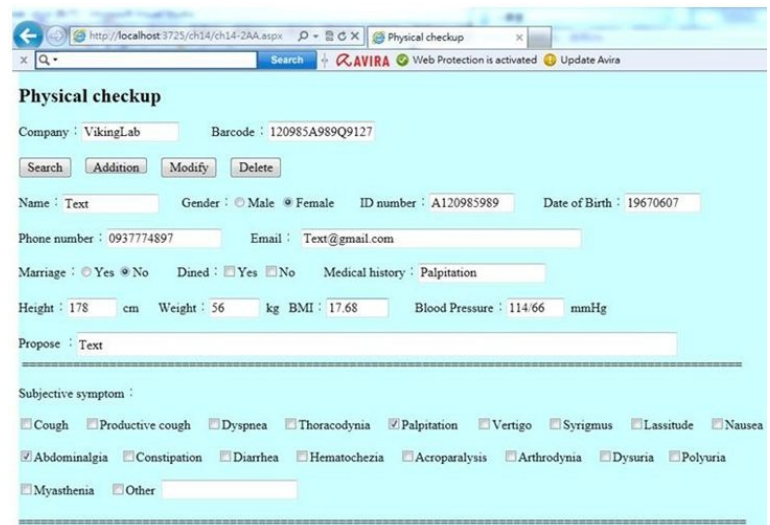


Figure 6. The GUI of the electronic form in RIA technology.

Figure 7. The GUI of the electronic form in Android platform.

The report of physical examination could be shared by general browser, e.g. IE, Firefox, Chrome, which is shown in Fig. 8.

Figure 8. Physical examination report with StructuredBody.

4 Conclusions

In this paper, we adopt RIA and Android technologies to develop a HL7 CDA-based cross-platform and highly mobile interactive tools, which provide the physical examination, and enables subjects to confirm their examination information by their self and upload to the database in the same time [14]. That reduces the back-end human cost and paper storage space. This study use RIA Web form in the status of Internet available, and use Android tablet when Internet is unavailable, then the examination information can be stored locally. We will develop personal health records system in the future, which would provide the patients access for health management by their self, especially for the chronic diseases patients. In addition, the design of interface can be re-optimized makes healthcare providers conveniently in the future.

5 References

- [1] C.C. Wu, "A Reference Architecture and Solution Plan for the Employment of Digital Signature in Medical Information System". Department of Computer Science and Information Engineering, National Dong Hwa University, Taiwan, 2004.
- [2] M.L. Müller, F. Ückert, T. Bürkle, H.U. Prokosch, "Cross-institutional data exchange using the clinical document architecture (CDA)", International Journal of Medical Informatics, vol 74, pp. 245–256, 2005.
- [3] F. Piero, R. Gustavo, Sá, and N.F. Fernando, "Rich Internet Application", IEEE Internet Computing, vol 14, pp. 9–12, 2010.
- [4] W. Enck, M. Ongtang, and P. McDaniel, "Understanding Android Security", IEEE Security & Privacy, vol 7, pp. 50–57, 2009.
- [5] Android. <http://developer.android.com/index.html>
- [6] C.C. Chang, The Development of Remote Monitoring with Embedded System, National Yunlin University, Taiwan, 2007.
- [7] C. Neale, "The .NET Compact Framework Group", IEEE Pervasive Computing, vol 1, pp. 84–87, 2002.
- [8] Microsoft MSDN. <http://msdn.microsoft.com/>
- [9] MySQL. <http://www.mysql.com/>
- [10] CDA_R2_NormativeWebEdition2005/infrastructure/cda/cda.htm#Introduction_to_CDA_Technical_Artifacts
- [11] IHE Laboratory Technical Framework Supplement 2006-2007 Sharing Laboratory Reports (XD*-LAB), <http://www.ihe.net/Laboratory/index.cfm>.
- [12] Logical Observation Identifiers Names and Codes (LOINC), <http://loinc.org/>
- [13] Alschuler Associates LLC–cda support/CDA Tools/CDA Validator, available : <http://www.alschulerassociates.com/validator/>
- [14] M. Weitzel, A. Smith, S. de Deugd, and R. Yates, "A Web 2.0 Model for Patient-Centered Health Informatics Applications", Computer, vol 43, pp. 43–50, 2010.

Cloud Computing System for Integrated Electronic Health Records

Hebah Mirza and Samir El-Masri

Department of Information Systems, College of Computer and Information Sciences / King Saud University, Riyadh, Saudi Arabia

Abstract. *The application of Information and Communication Technology in healthcare environment facilitates healthcare process and improves its service quality. However developing healthcare via technology innovations usually faces many challenges such as fear of cost, maintenance difficulties and security threats. Electronic Health Record systems showed great effects on developing healthcare outcomes and many are adopting it, but still many others fear to use it or face problems during its implementation and maintenance. Cloud computing technology is a new technology that has been used in different life environments and showed large positive changes. Despite the great features of Cloud computing, they haven't been utilized fairly yet in healthcare industry. This paper presents an innovative Healthcare Cloud Computing system for Integrated Electronic Health Records (EHRs). The proposed Cloud system applies Cloud Computing technology on EHR system, to present a comprehensive EHR integrated environment. The proposed Cloud system is composed of three main components; first is the Cloud's Central Database that represents the data repository for EHR's. The second part is the Unifier Interface Middleware; this component remains in the Cloud and responsible for masking the heterogeneity and standardising the communication between different EHR standards and the Cloud EHR system. Third component represents the web portal for the Cloud, it issues request messages and receives responses from the Cloud system via secured network connections.*

Keywords: Cloud Computing, Electronic Health Record, Integration, Middleware

1 Introduction

As healthcare remains one of the most important and expensive sectors in any community; many technologies have emerged and been funded by governments to improve healthcare delivery outcomes. The most common technologies that are designed to improve healthcare services are MHR (Medical Health Record), PHR (Personal Health

Record), and EHR. EHR has many definitions, such as the electronic record that stores patient's medical history information in a health record system, accessible and managed by care providers [1]. Despite its positive impact on healthcare services; its adoption progress is slow in most healthcare institutions in worldwide; especially in developing countries due to several common challenges. Several studies found that the main barriers for its adoption are: 74% because of its high purchase costs, 44% for its high maintenance costs, physician's resistance 36%, Unclear return on investment 32% and Shortage in skilled IT staff 30% [2]. Patients in developing countries or in rural areas suffer from travelling to large hospitals carrying their paper health records and crossing the land to reach the specialized physicians and medical care in large hospitals with EHR systems. Moreover, patients registered in independent EHR systems in different hospitals also suffer from transferring their files to other hospitals. Such difficulties can be defeated by integrating EHR systems in healthcare institutions. But EHR integration (the process of patient information sharing among health care providers and exchanging them over the internet with other healthcare providers) remains a challenge and a serious concern since it is exposed to theft, security violation, and standardization difficulties [3].

Cloud computing technology is considered to be the new, most interesting and comprehensive solution in the IT world. Its main objective is to leverage Internet or Intranet for users to share resources [4]. The National Institute of Standards and Technology (NIST) defined it as: "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction"[5]. Cloud computing is a cost effective, automatically scalable, multitenant and securable platform that is managed by the cloud provider. Recently researchers have started to utilize Cloud computing services to solve many problems in healthcare IT adoption. But, not many researches entered the field of integrating EHR with the cloud services yet. A proposed system by Mohammed D., et al.[6] ,which represents a private Health Cloud

eXchange (HCX) system; this system outlines a distributed web based infrastructure for EHR sharing on the cloud among both local clients and third party healthcare information system. *NefeliPortal* is designed as a cloud EMS/PHR prototype architecture proposed by Koufi V., et al. [7]. Another paper presented *Artemis Cloud* computing framework by McGregor, C. [8]; for patients with critical care units (CCU) in rural and remote centers. It captures/process the real-time medical monitoring data with EHR data from the clinical information system. Some other researchers have implemented Cloud computing for Medical Image systems such as Yang C., et al.[9], they proposed MIFAS (Medical Image File Accessing System). There are more researches concentrating on patient's information security during exchange among the cloud's platform and other health institutions such as *EHR security reference model* by Zhang R.[10]. Healthcare and human life care comes in the first priority to get advantage of such technology. Therefore, this research is proposed to prove that the above challenges can be defeated by applying Cloud Computing technology to integrated EHR system. This paper is divided into three parts, the first part describes the proposed system components and its process, and then the discussion part explains the system impacts on healthcare institutions and discusses the system advantages. The paper ends with the conclusion.

2 Proposed Cloud system for EHR integration

A new healthcare Cloud system has been proposed for unification and integration of EHRs. The proposed system utilizes all features of Cloud computing combining them with EHR system features to gain one unified central system that controls electronic health records in the cloud infrastructure. And represent the solution for all hospitals in the region with an opportunity to use and share EHR. The Cloud system components are explained bellow, where different situations and scenarios for using the cloud and sharing EHR's are explained.

2.1 System components

The proposed public cloud infrastructure include: 1) a Central Database server that represents the clouds' IaaS data repository that communicates with the sharing hospitals through 2) a Unifier Interface Middleware (UIM). as an intermediary tool between the clouds central Database server and the sharing hospitals systems, 3) and the Cloud EHR Web Portal; that represent the cloud's SaaS for retrieving and displaying any required patient information.

2.1.1. Central Database server

The IaaS cloud Datacenter contains the Central Database server as the data repository for storing EHR's and retrieving patient information. The information is stored in XML format as a unified standard which can be stored and retrieved via query commands sent and resaved from the sharing hospitals Web Portal (web browser/EHR) system passing through the Unifier Interface Middleware (UIM). The Datacenter is managed by the Cloud Provider, and the Central Database applies virtualization techniques on its resources, where the hypervisor schedules the requests and handles the load balancing on each resource in the cloud datacenter.

2.1.2. Unifier Interface Middleware (UIM)

This part of the cloud provides an Interface that masks the heterogeneity of all sharing hospitals EHR standards, to facilitate the communication transactions between the Central Database and hospitals systems. It holds all types of EHR standards, so it recognizes any type it communicates with. It remains in the cloud infrastructure and communicates with the sharing hospitals via network connections. This is beneficial because rather than each hospital have to generate its own mask interface to benefit from the health cloud system; one interface will reside on the cloud and handle the heterogeneity from there. This interface handles two conditions;

2.1.2.1. Hospitals host their own EHR system locally

The UIM receives the Request message, and translates it into XML format. Then resend it to the cloud's Central Database.

2.1.2.2. Hospitals host their EHR's on the cloud (without local EHR system)

The UIM receives the Request message then resends it to the cloud's Central Database. The request message (Req.msg) has three parameters, shown as: *Req.msg (TN, CPID/LPID, NN)*:

- *TN*: Source Hospital ID composed of two parts: 1) *T*=EHR Type, either 0 if in cloud, or 1 if local, 2) *N*= hospital name 5 characters at most.
- *CPID/LPID*: this holds the patient ID number either in the Cloud (CPID) which is equivalent to the national number (NN), or the independent hospital local patient ID (LPID).
- *NN*: this holds patient national number; but this parameter is null when CPID exists.

The response message (Resp.msg) has four parameters, shown as *Resp.msg (CPID/LPID, NN, PCMH, PLMH)*:

- *CPID/LPID*: Patient ID (either in Cloud or Locally) in the responding hospital.
- *NN*: this holds patient national number, null if CPID exists.

- *PCMH*: this parameter holds Patient-Medical history stored in the cloud's EHR database, can take null value if patient do not have EHR in Cloud database. Only one EHR (PCMH) exists for each patient registered in hospitals with Cloud EHR.
- *PLMH*: this parameter holds Patient-Medical history stored in the local EHR system, can take null value if patient do not have EHR in local systems. PCMH and PLMH are collectively exhaustive. Each patient can have more than one local EHR, if he is registered in different hospitals with local EHR.
- *If Req.msg (0 #, CPID=NN) is true then the request message comes from a hospital that has its EHR on the cloud.* It matches the CPID to retrieve patients EHR information stored in the cloud. Where all hospitals that doesn't have a local EHR, they will have the same record for the same patient inside the cloud. So, only one online record for each patient with the visited hospitals names is stored in the cloud. The Central DB searches for a match for patient's National number since he might have EHR files stored in other hospitals with local EHR, if so: It resends a request message to the (match) found hospitals via UIM. The UIM reformat the request message according to the target hospital EHR standard format. The UIM send the reformatted request message and receives the response message, via network connection. After the UIM reformats the response message to XML format it sends it back to the clouds EHR. The central database combines the response message PLMH with the PCMH. Create a final response message in an XML format and send it to the requesting hospital.

2.1.3. Cloud EHR Web Portal

This is the third part of the cloud (top layer). This layer provides an application (SaaS) for EHR systems. The proposed Health Cloud system presents for end users a configurable EHR web portal for the Central Database. The web portal is responsible to issue send messages and receive response messages between the UIM and the hospital system. If a hospital has its own EHR system the web portal offers the user two tabs, either enter the hospital's local EHR system, or to the cloud Central Database. This web page provide the user with the ability to retrieve, update and receive EHR information from the cloud's Central Database EHR with limited access depending on the end user's privileges. The user can also, know from the retrieved information displayed on the web portal, if the requested EHR for a specific patient from a specific hospital exists inside the cloud or on the target hospital's local system. And can choose to view EHR information about the patient even from locally independent hospitals connected to the cloud.

2.2 System process

The system process starts when the cloud's Central Database receives the request message via the UIM, issued by end user via the web portal, It analyzes the Request message (Req.msg), and response (Resp.msg) in different ways according to:

- *If Req.msg (1 #, LPID, NN) is true then the request message is issued by an independent EHR system:* The UIM resends the request message in an XML format to the Central Database. The Database start searching for a mach for the NN from the request message to a CPID; if it finds a mach this means the patient have visited hospitals that have the clouds EHR, and then determine those hospitals. Then the Database start searching a mach for the NN from the request message to NN column for other hospitals connected with the cloud; if it finds a mach, this means the patient have visited hospitals that have their EHR locally. In this case the Central DB reformats a request messages to the matching independent hospitals, for gathering the patient's medical history. And send them via UIM that will reformulate the message type depending on the target hospital EHR standard type. Then the response messages will pass through UIM via network connection and reformed again into XML format and passed to the Central Database. The Cloud's Central Database will combine the existing (cloud's) EHR with the received medical history and form a complete report for the patient EHR in an XML Response message format. Finally the UIM will resend the final response message to the requesting hospital via network connection and a matching standard format. See figure1.

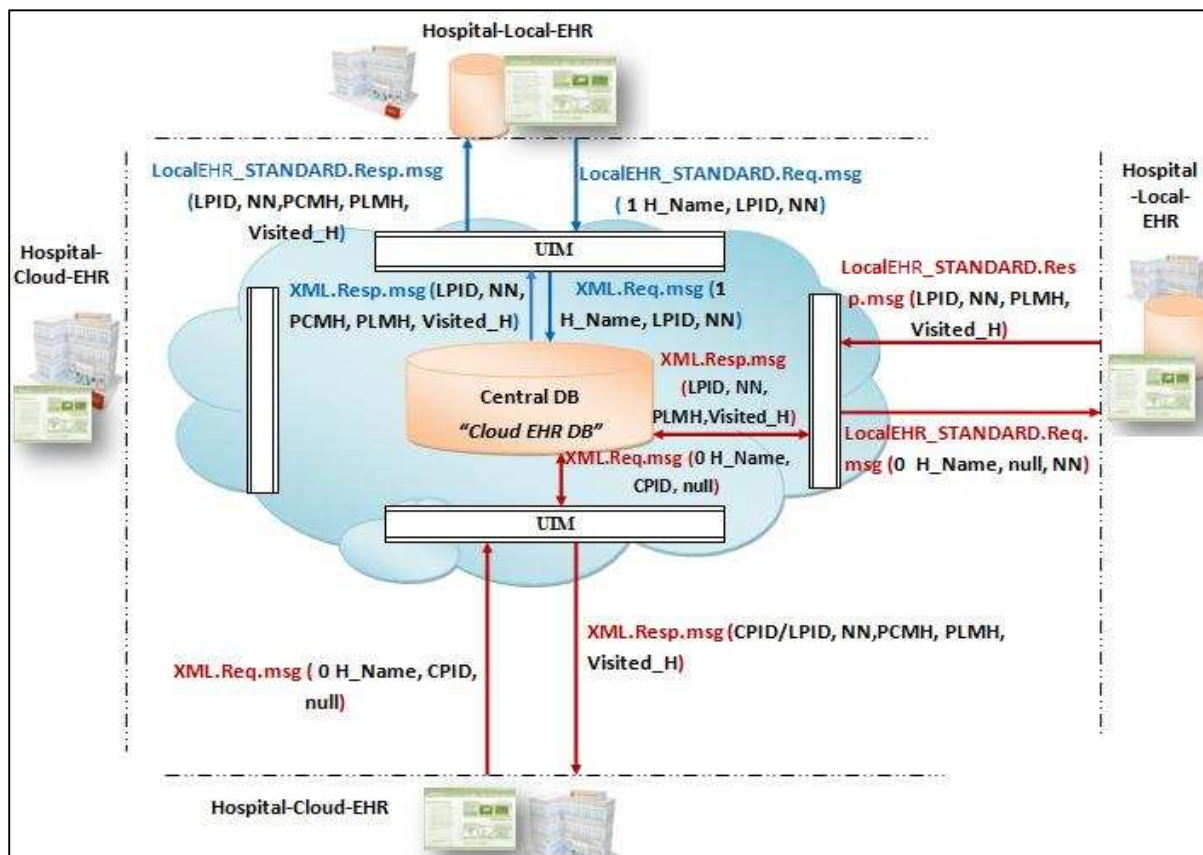


Figure 1 the proposed Healthcare Cloud system processes

3 Discussion

Cloud Computing can be applied to EHR system to facilitate EHR adoption for all types of healthcare institutions. Health Cloud features such as multi-tenancy, automatic scalability, securable connections and authorized data transactions managed by the cloud's provider; gives to many healthcare providers the ability to share a unified EHR system that handles as much users as possible with high performance. In fact the whole city is able to integrate into the cloud's EHR system without disk space, maintenance and security worries. Many other features of the health cloud such as pay as you go, solves high costs barriers for small healthcare institutions to adopt EHR technology ready from the cloud. Moreover, the proposed health cloud system showed that it is possible to integrate different kinds of EHR systems using the UIM tool that eliminates the burden of masking heterogeneity for healthcare institutions to share their local EHR system and share the clouds EHR in the same time. Thus, the proposed system provides a standardised unified environment for different EHR systems to communicate freely without any barriers.

The proposed system overcomes the challenges of implementing EHR systems for many hospitals such as maintenance complexities, staff training and high

cost. In all cases the proposed system has the following advantages: Present a comprehensive and successful healthcare service. It allows' many healthcare providers to communicate and easily share patients EHR information among the healthcare cloud. Moreover, It overcome the challenges of EHR system integration such as network security concerns and information standardization difficulties. And present a configurable and scalable EHR system in Cloud computing platform for healthcare providers. It also, maximizes healthcare services quality outcomes, by releasing them from technology problems. It offers the opportunity for any kind of healthcare institution especially; rural and small sized hospitals or clinics to use EHR and join the cloud. Finally, It release patients' from suffering to find and move to the specialized healthcare providers they need, facilitate healthcare delivery process and then offer patients more easy, reliable and corporative healthcare life.

4 Conclusion

This paper proposed a novel solution for healthcare institutions to use EHR systems and overcome its challenges. The proposed system is composed of three main components, and applies cloud computing technology on EHR system integration. It provides a ready EHR system for all

kinds of hospitals. Irrespective of the number or the size of hospitals that join the cloud; the system is capable to work in integrity and it will offer healthcare providers the ability to communicate in a controlled, scalable, safe and cost effective way under the cloud. Future work will focus on completing the implementation and on evaluating the system.

5 References

- [1] Spil, T.A.M, et al. (2010). Value, Participation and Quality of Electronic Health Records in the Netherlands. *IEEE Computer Society. System Sciences (HICSS), 2010 43rd Hawaii International Conference on January 2010, 1-10.*
- [2] AK, Jha. et al., (2009). Use of Electronic Health Records in U.S. Hospitals. *The New England Journal of Medicin.* (10). 1628-1638.
- [3] Sun, J and Fang, Y. (2010). Cross-Domain Data Sharing in Distributed Electronic Health Record Systems. *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS*, 21 (6), 754 - 764.
- [4] Zhang, L and Zhou,Q. (2009). CCOA: Cloud Computing Open Architecture. *IEEE International Conference on Web Services. Los Angeles, CA, July 2009. p 607-617.*
- [5] T. Sridhar. (2009). Cloud Computing - A Primer. *The Internet Protocol Journal.* 12 (3).
- [6] Mohammed,S. Servos,D and Fiaidhi,J. (2010). HCX: A Distributed OSGi Based Web Interaction System for Sharing Health Records in the Cloud. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 102-107.
- [7] Koufi,V. Malamateniou,F and Vassilacopoulos,G.. (2010). Ubiquitous Access to Cloud Emergency Medical Services. *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference, 1-4.*
- [8] McGregor, C.. (2011). A Cloud Computing Framework for Real-time Rural and Remote Service of Critical Care. *Computer-Based Medical Systems (CBMS), 2011 24th IEEE International Symposium, 1-6.*
- [9] Yang, C. Teng Chen,L., Chou,W and Chieh Wang, K. (2010). Implementation of a Medical Image File Accessing System on Cloud Computing. *Computational Science and Engineering (CSE), 2010 IEEE 13th International Conference on 11-13 Dec. 2010 Hong Kong, 321-326.*
- [10] Zhang,R. and Liu,L.. (2010). Security Models and Requirements for Healthcare Application Clouds. *2010 IEEE 3rd International Conference on Cloud Computing, 268-275.*

SESSION

HIGH PERFORMANCE METHODS, COMPUTATIONAL METHODS FOR FILTERING, NOISE CANCELLATION, AND SIGNAL AND IMAGE PROCESSING

Chair(s)

TBA

Localized Deconvolution: Characterizing NMR-based Metabolomics Spectroscopic Data using Localized High-throughput Deconvolution

Paul E. Anderson¹, Ajith H. Ranabahu², Deirdre A. Mahle³, Nicholas V. Reo⁴,
Michael L. Raymer², Amit P. Sheth², and Nicholas J. DelRaso³

¹Department of Computer Science, College of Charleston, Charleston, SC 29424

²Department of Computer Science, Wright State University, Dayton, OH 45435

³711 Human Performance Wing, Air Force Research Laboratory, Wright-Patterson AFB, OH

⁴Department of Biochemistry & Molecular Biology, Wright State University, Dayton, OH

Abstract—*The interpretation of nuclear magnetic resonance (NMR) experimental results for metabolomics studies requires intensive signal processing and multivariate data analysis techniques. Standard quantification techniques attempt to minimize effects from variations in peak positions caused by sample pH, ionic strength, and composition. These techniques fail to account for adjacent signals which can lead to drastic quantification errors. Attempts at full spectrum deconvolution have been limited in adoption and development due to the computational resources required. Herein, we develop a novel localized deconvolution algorithm for general purpose quantification of NMR-based metabolomics studies. Localized deconvolution decreases average absolute quantification error by 97% and average relative quantification error by 88%. When applied to a ¹H metabolomics study, the cross-validation metric, Q^2 , improved 16% by reducing within group variability. This increase in accuracy leads to additional computing costs that are overcome by translating the algorithm to the map-reduce design paradigm.*

Keywords: Metabolomics, quantification, map-reduce, deconvolution

Web: http://ds.cs.cofc.edu/index.php/Localized_Deconvolution

Contact: andersonp@cs.cofc.edu

1. Introduction

Metabolomics, the measurement of metabolite concentrations and fluxes in various biological systems, is one of the most comprehensive of all bionomics [1]. Unlike proteomics and genomics that assess intermediate products, metabolomics assesses the end product of cellular function, metabolites. Changes occurring at the level of genes and proteins (assessed by genomics and proteomics) may or may not influence a variety of cellular functions. But metabolomics, by contrast, assesses the end products of cellular metabolic function, such that the measured metabolite profile reflects the cellular metabolic status. For instance, a disease process or exposure to a xenobiotic may interfere at the genomic or proteomic level, while it will always manifest itself at the metabolomic level. Further, nuclear magnetic resonance

(NMR) spectroscopy of biofluids has been shown to be an effective method in metabolomics to identify variations in biological states [2], [3]. In contrast to various other proteomic, genomic, and metabolomic analyses, NMR spectroscopy is non-invasive, non-destructive, and requires little sample preparation [1].

Typically, NMR metabolic spectroscopic data are analyzed as follows: (1) standard post-instrumental processing of spectroscopic data, such as the Fourier transformation, phase adjustment, and baseline correction; (2) quantification of spectral signals commonly implemented via binning; (3) normalization and scaling; and (4) multivariate statistical modeling of data. Quantification of spectral signals, step (2), is a key step in the development of classification algorithms and biomarker identification (i.e., pattern recognition). A common method of quantification employed by the NMR community is known as binning or bucketing, which divides a NMR spectrum into several hundred regions. This technique is performed to (1) minimize effects from variations in peak positions caused by sample pH, ionic strength, and composition (Spraul et al. 1994); and (2) reduce the dimensionality for multivariate statistical analyses. The result is a data set with fewer features, thereby, increasing the tractability of pattern recognition techniques, such as principal component analysis (PCA) [4] and partial least squares discriminant analysis (PLS-DA) [5].

The standard quantification method is to divide a spectrum into several hundred non-overlapping regions or bins of equal size. This simple technique has been shown to be effective in the field of metabolomics [6], [7]. While standard quantification mitigates the effects from variations in peak positions, shifts occurring near the boundaries can result in dramatic quantitative changes in the adjacent bins due to the non-overlapping boundaries. This problem can be countered by incorporating a kernel-based binning method that weights the contribution of peaks by their distance from the center of the bin [8] or by dynamically determining the size and location of each bin [9], [10]; however, these techniques fail to remove irrelevant adjacent signals.

There are several alternatives to spectral binning that still provide data dimension reduction [11]. Examples of these

include PARS [12], direct quantification [13], peak alignment tools in HiRes [14], and targeted profiling [15]. These techniques identify peaks or specific peak patterns in the spectra that are conserved across spectra. After the patterns have been identified, they are quantified by determining the peak area or amplitude. The accuracy of these algorithms is dependent on the spectral resolution, the quality of the peak alignment, and the breadth of spectroscopic pattern databases. Since spectral resolution is dependent upon the magnetic field strength (i.e., instrument specific), the spectral patterns in complex mixtures (e.g., urine and plasma) are also field dependent. This adds another level of complexity to targeted profiling techniques that attempt to match spectral patterns against standard spectra acquired at a specific magnetic field.

Despite the development of these alternative quantification techniques, binning remains a common technique for the NMR community owing to high throughput quantification technique [16], [11]. The wide spread use of advanced quantification algorithms has been hindered by the additional computing resources and manual intervention required to incorporate them into general metabolomics workflows. Herein, we propose a novel localized deconvolution algorithm for NMR spectroscopic data that removes adjacent and convoluting signal for significantly improved full spectrum quantification that does not rely on the breadth of annotated spectral databases. By pursuing a localized strategy for deconvolution, the algorithm is suited for implementation in the map-reduce paradigm that will allow for web-scale high-throughput availability. We show this technique is superior to alternative high-throughput quantification techniques by comparing the improvement in quantification accuracy on complex ^1H NMR spectroscopic data and realistic synthetic spectra.

2. Approach

The variability and complexity inherent in ^1H NMR spectra of biofluids requires sensitive signal processing and pattern recognition techniques to discover novel patterns in the data. The technique of spectral quantification is a general signal processing technique that reduces the dimensionality of spectroscopic data by transforming full resolution spectra into a feature vector for subsequent pattern recognition. The goals of which are to retain pertinent information and mitigate quantitative effects of peak misalignment. Biomarker identification can then be defined as finding a set of features that describe a pattern between groups, thus the success of biomarker identification is directly related to the quality of the feature vectors. Here a biomarker is defined as a set of NMR signals that change relative to some reference (i.e., before and after exposure to a toxin). Such an experiment will have at least two groups (e.g., pre-dose and post-dose) for which spectroscopic data is compiled. A significant step prior to biomarker identification is spectral quantification,

our method, localized deconvolution, is comprised of three steps:

- 1) Solve the peak registration (correspondence) problem using an adaptive binning approach
- 2) Model the signals in each region using a Gauss-Lorentzian peak construct
- 3) Deconvolve the localized subproblem by removing adjacent and baseline signals

This technique is applied to a metabolomics study of toxicology for the identification of biomarkers associated with a kidney toxin (α -naphthylisothiocyanate) response.

3. Methods

3.1 Peak registration

The first step in localized deconvolution is to define the subproblems of interest, which are defined as regions containing a signal of interest across spectra. This problem, also known as the peak registration or correspondence problem, is solved by applying an adaptive binning technique: dynamic adaptive binning [9]. Peak registration is necessary to overcome the variability in signals between subjects (or samples). Our localized deconvolution technique leverages an adaptively binning technique to generate the regions of interest, which can subsequently be solved in parallel; however, our method can be easily adapted to other methods of registration, including peak alignment and targeted approaches.

Dynamic adaptive binning determines the optimal bin configuration of n observed peaks as measured by an objective function. This process is divided into two steps: (1) determining the location of the observed peaks in each spectra and (2) finding the optimal bin boundaries with respect to the objective function. The identification of the observed peaks in each spectrum is accomplished by identifying local maxima after smoothing via a wavelet transform [17], [18], [19], [20], [21]. After the observed peaks of each spectrum have been determined, the algorithm determines the optimal bin configuration using a dynamic programming strategy. A detailed description of dynamic adaptive binning and proofs verifying optimal substructure can be found in [9].

3.2 Model the signals

While peak registration provides a mechanism for matching corresponding signals between spectra, quantification is still impaired by adjacent signal and baseline distortions. This problem is mitigated by removing adjacent signals that affect the true value of the signal of interest. The observable NMR free induction decay (FID) signal is an exponential decaying sinusoid leading to an approximate Lorentzian peak shape after Fourier transformation. These individual signals, S , are modeled by a Gaussian-Lorentzian function that is defined by the standard deviation of the Gaussian (σ), the

center (x_c), the width at half height of the Lorentzian (Γ), and the magnitude (M):

$$S([M, \sigma, P, x_c], x) = P * L([M, \Gamma, x_c], x) + (1 - P) * G([M, \sigma, x_c], x) \quad (1)$$

$$L([M, \sigma, P, x_c], x) = \frac{M * \Gamma^2}{4(x - x_c)^2 + \Gamma^2} \quad (2)$$

$$G([M, \sigma, P, x_c], x) = \text{Mexp}(-(x - x_c)^2 / (2\sigma^2)) \quad (3)$$

where $\Gamma = 2 * \sqrt{2 * \ln(2\sigma)}$, and P is a real value between 0.0 and 1.0 that weights the contribution of the Lorentzian ($L(\dots)$) and Gaussian ($G(\dots)$) functions.

The mixture of the Gaussian and Lorentzian peaks is selected to provide a flexible peak shape. The relationship between the width at half height of the Lorentzian peak and the standard deviation of the Gaussian peak is fixed by assuming that both the height and the width at half height are the same for both peaks. This simplifies the model by avoiding a separate parameter for both the standard deviation and width at half height.

3.3 Deconvolve

Noise and baseline distortions arise from congested areas of the spectrum with multiple overlapping peaks, naturally broad signals from proteins or lipids, and the amplifier of a quadrature detection magnet system [22]. With the previously described model for the underlying signals, our algorithm removes unwanted signals from the region of interest. This deconvolution procedure divides each spectral subproblem into its constituent signals (baseline, noise, and individual signal). These predefined regions and subproblems are adapted from the results of dynamic adaptive binning. If a targeted or peak alignment approach is taken, the regions can be defined as fixed width regions containing the targeted or aligned peaks of interest.

The solution to each subproblem is obtained by breaking each region into signal of interest, adjacent signal, and baseline. The baseline and adjacent signals are then removed, leaving the signal of interest. This construction of subproblems allows the problem to be transformed into the map-reduce paradigm (described later). As part of this work, two alternative definitions of the subproblems were explored:

- 1) Region of interest
- 2) Region of interest with adjacent buffer regions

By including adjacent buffer regions, it is hypothesized that better estimates of adjacent signals are obtained, thus, improving the accuracy of the quantification. Solutions to subproblems for both definitions are constructed by combining a model of baseline and a set of Gauss-Lorentzian peaks:

$$\Theta(\beta, x) = \sum_{j=1}^N S([M_j, \sigma_j, P_j, x_{cj}], x) + \text{baseline}([b_1, \dots, b_k], x) \quad (4)$$

$$\beta = [M_j, \sigma_j, P_j, x_{cj}, b_1, \dots, b_k] \quad (5)$$

where $\Theta(\beta, x_i)$ is the model for each region with the model parameters, β . Further, N is the number of peaks in the subproblem, thus, M_j , σ_j , P_j , and x_{cj} refer to the height, standard deviation, fraction of Lorentzian, and the center of the j -th peak. $\text{baseline}(\dots)$ is a piecewise baseline linear function, where b_1, \dots, b_k are the heights of the piecewise segments.

The final locations of the peaks and their parameters (e.g., width, height) are determined algorithmically by solving the corresponding nonlinear curve-fitting problem. The parameters of the nonlinear curve-fitting problem are estimated by a subspace trust-region method based on the interior-reflective Newton method (Coleman and Li 1994, 1996). The parameters are adjusted to minimize the function:

$$1/2 \sum_i^m (\Theta(\beta, x_i) - y_i)^2, \quad (6)$$

where x_i and y_i are the chemical shift and intensity of the i -th point in the segment, m is the number of data points in segment, β is a vector of parameters, and Θ is the model of each subproblem that will be fit.

The nonlinear curve-fitting algorithm estimates the optimal model parameters using their initial values and bounds. The initial location, x_{cj} , of each peak is manually selected. The initial height, M_j , of each peak is defined as the difference between the maximum and minimum intensities in the region surrounding the peak. The initial value of the width at half height, Γ_j , is defined as double the distance (ppm) between the maximum intensity in the region and the location of the peak's half height (i.e., initial height divided by 2). The initial standard deviation, σ_j , can then be computed from the width at half height. The initial fraction Lorentzian, P_j , of each peak is defined as 0.5. The initial baseline heights, b_i , is defined as the minimum intensity in the segment. The lower and upper bounds for parameters are defined as:

$$\begin{aligned} 0 < M_j &\leq \text{MAX}_i, \\ 0 < \sigma_j &\leq |s_L - s_R|, \\ 0 &\leq P_j \leq 1.0, \\ \alpha_i &\leq x_{cj} \leq \omega_i, \\ 0 &\leq b_k \leq \text{MAX}_i, \end{aligned}$$

where MAX_i is the maximum height in the i -th segment, and s_L and s_R are the left and right boundaries of the segment. The boundaries for location of each peak, $[\alpha_j, \omega_j]$,

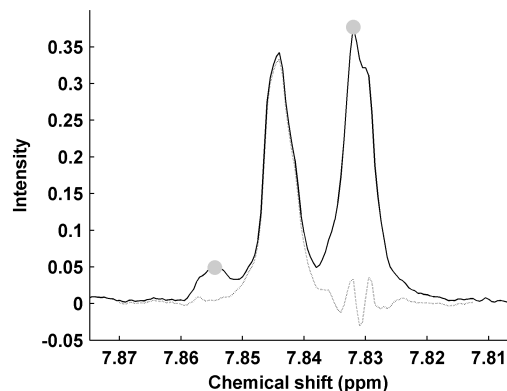


Fig. 1: Removal of adjacent signals (1st and 3rd peak) to target signal of interest (2nd peak) in overlapping regions

are defined as the locations corresponding to the minimum intensities between the current peak and the adjacent peaks. In the special cases of the first and last peaks of each segment, the segment boundary is used to define the region.

Through the solutions obtained for each subproblem, the frequency domain spectral data can be transformed into a feature vector by specifying a set of regions $\mathbf{R} = \{R_1, R_2, \dots, R_n\}$, where each region is identified by its chemical shift boundaries and the adjacent signals to remove from that region. The baseline is automatically removed from each region. By design, regions are allowed to overlap to filter out alternative sets of adjacent signals. This is demonstrated in Figure 1. The characterization of the metabolomics study for algorithm evaluation employs spectra binning to solve the correspondence problem; however, localized deconvolution can filter unwanted signals for the enhancement of targeted quantification, alignment algorithms, and other alternative quantification techniques.

3.4 Map-Reduce

A map-reduce architecture is employed to enable high-throughput spectral deconvolution. This architecture exposes cloud-based services using the web application framework Ruby on Rails. The algorithm is implemented as a Hadoop based map-reduce program using Hadoop streaming, a technique that allows one to use non Java based programs in the Hadoop architecture. This implementation uses a MATLAB implementation of the numerical optimization algorithm, in a similar fashion as experimented by [23], [24].

The Hadoop streaming mechanism processes data in lines. Hence the data format used as the input to the process is an independent deconvolution problem on each line. This is also important to maintain clear record boundaries for the record splitter. Given that this task is map centric, i.e. the critical process is performed during a map and reduce is merely a combine operation, the number of mappers is a sensitive operator. The map phase consists of solving the

aforementioned non-linear optimization subproblem. The reduce step is the recombination and ordering of these results. In order to expose the Hadoop functions in a convenient way to the biologists and also for better integration with existing workflow engines, a web service is implemented. The web service follows the REST paradigm and can be accessed by an HTTP POST operation. The web service is deliberately made into an asynchronous service due to the longer processing time for larger jobs. The processing time varies depending on the complexity of the spectra, and therefore, could not be incorporated into a synchronous web service.

3.5 Cluster Setup

The Hadoop cluster consists of 15 dedicated server computers, each having 16GB of RAM and Quad core AMD processor and connected via Gigabit Ethernet. The Hadoop software version is 0.20.1. The cluster was configured to have a total map task capacity of 120 and reduce task capacity of 90. Jobs were submitted in groups of 5, 10, 15, and 20 (e.g., 5 spectra at a time).

3.6 Synthetic Data

Both empirical and synthetic spectroscopic data are employed to show the application of localized deconvolution. The synthetic spectroscopic data sets are based on urine ^1H spectra and were developed by characterizing the salient distributions in empirical spectroscopic data (Anderson et al., 2009). These synthetic data sets enable the use of exacting performance metrics because the true location and size of each peak is known *a priori*.

A synthetic data set of 20 complex ^1H spectra was generated, and it was analyzed by two direct measures of the spectral quantification accuracy for each algorithm: absolute quantification error (*AQE*) and relative quantification error (*RQE*):

$$AQE = \frac{100}{N * M} \sum_{b=1}^M \sum_{s=1}^N \left| \frac{predicted_{b,s} - true_{b,s}}{true_{b,s}} \right| \quad (7)$$

$$RQE = \frac{100}{M} \sum_{b=1}^M \left| \frac{std(predicted_b) - std(true_b)}{std(true_b)} \right| \quad (8)$$

where $predicted_{b,s}$ is the localized deconvolution results for bin b and spectrum s , $true_{b,s}$ is the true deconvolution results, M is the total number of bins, N is the total number of spectra, and $std(predicted_b)$ is the standard deviation of the set of all localized deconvolution results for bin b , and $std(true_b)$ is the standard deviation of the set of all true deconvolution results for bin b .

3.7 Experimental Data

In addition to comparing spectral binning algorithms on synthetic data sets, this manuscript demonstrates the application of high-throughput localized deconvolution on empirical

Table 1: Mean/median absolute and relative quantification error for standard binning (Standard), localized deconvolution with positive baseline constraint (Region (+)), localized deconvolution with additional buffer and positive baseline constraint (Region & Buffer (+)), localized deconvolution (Region (+/-)), and localized deconvolution with additional buffer (Region & Buffer (+/-))

	Absolute Quantification Error		Relative Quantification Error	
	MEAN	MEDIAN	MEAN	MEDIAN
Standard	1405	125	409	45
Region (+)	36	25	50	25
Region & Buffer (+)	50	19	178	21
Region (+/-)	37	24	48	24
Region & Buffer (+/-)	55	18	172	21

data from a ^1H NMR-based experiment to monitor rat urinary metabolites after exposure to α -naphthylisothiocyanate (ANIT) [16]. A subset of this data set was used to compare the quantification algorithms. Specifically, an ANIT dose of 20 mg/kg at 2 days post-exposure was selected, and the performance of the algorithms were analyzed by studying the results of a standard supervised learning procedure, Orthogonal Projection onto Latent Structures (O-PLS) [25].

The O-PLS model was evaluated on its predictive ability, using the Q^2 (coefficient of prediction) metric. Q^2 was calculated as follows:

$$Q^2 = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

where $PRESS$ is the Predicted REsidual Sum of Squares calculated as the residual e between the predicted and actual Y during leave-one-out cross-validation, SSY is the Sum of Squares for y , \bar{y} is the y mean across all samples, and y_i is the y value for sample i . As Q^2 approaches 1, the more predictive capability the model exhibits. A Q^2 value less than 0 shows the model has no predictive power.

4. Results and Discussion

Standard high throughput quantification techniques, such as uniform binning or bucketing, have shown to be effective in reducing the dimensionality and mitigating spectral misalignment; however, these techniques often introduce erroneous quantification errors due to overlapping and adjacent signals. To illustrate the advantages of localized deconvolution, we analyzed synthetic and empirical data. The absolute and relative accuracy of quantification was measured on realistic ^1H synthetic spectroscopic data, which were modeled after a traditional urine NMR-based metabolomics study. These results are summarized in Table 1. The difference in performance by including a buffer region and constraining the baseline to positive offsets are shown in Figures 2(a) and 2(b).

As determined by a one-way ANOVA ($\alpha = 0.05$ assumed for all subsequent statistical tests), the means absolute quantification error for all quantification methods are signifi-

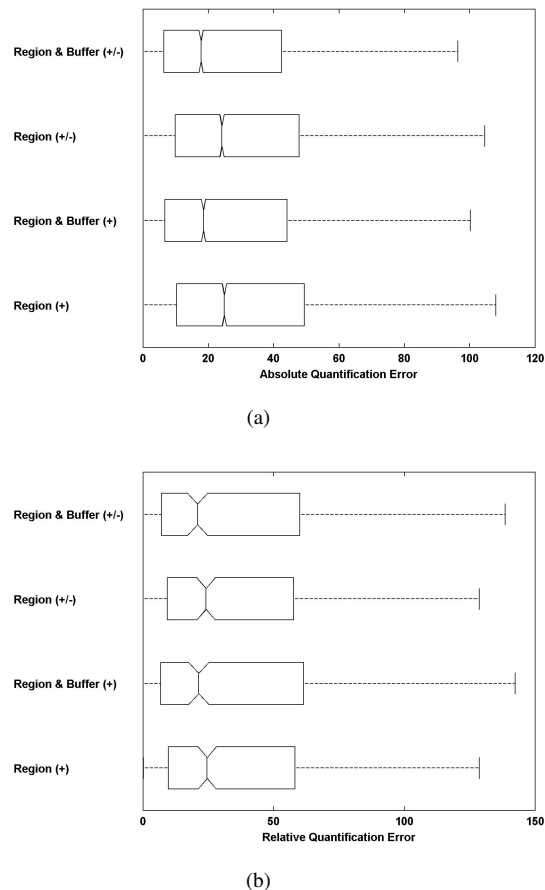


Fig. 2: Box and whisker plot of the absolute quantification error (a) and the relative quantification error (b)

cantly different. Comparing pairs of methods shows that the standard quantification mean absolute quantification error is significantly different than all localized deconvolution methods using the Tukey-Kramer multiple test correction. To evaluate the median absolute error, the Kruskal-Wallis test was applied to the performance data; the results of which showed that there is a difference between quantification methods as measured by the median absolute quantification error. Specifically, the standard quantification median absolute quantification error is significantly different from all localized deconvolution methods. The mean relative quantification error is significantly different for all methods (one-way ANOVA). The standard quantification mean relative quantification error is significantly different from all localized deconvolution methods (Tukey-Kramer multiple test correction).

Among the four different versions of localized deconvolution, a one-way ANOVA showed that the means of the absolute quantification error are significantly different, and the mean absolute quantification error of Region & Buffer (+/-) is significantly different from the means of Region (+)

and Region (+/-). Using the Kruskal-Wallis test, the medians are significantly different, and specifically, the medians of Region & Buffer (+) and Region & Buffer (+/-) are significantly different from the average rank of Region (+) and Region (+/-). A Tukey-Kramer correction was applied to correct for multiple tests. The one-way ANOVA on the mean relative quantification error and Kruskal-Wallis test on the median relative quantification error failed to reject their null hypotheses. i.e., there is not a significant difference among the localized deconvolution methods when examining relative quantification error.

These significant results demonstrate the error in approximating the underlying peak signals with standard binning. If two peaks are adjacent in a spectrum, the degree to which they influence each other will be proportional to their intensity and proximity. Adjacent peaks that are drastically smaller will be heavily influenced by the larger adjacent peaks. Quantifying these smaller peaks is of particular interest to the metabolomics community, as the magnitude of the peak does not determine its relevance in any given study. By modeling each peak individually while simultaneously providing high throughput quantification, localized deconvolution significantly improves the absolute and relative quantification accuracy in NMR-based metabolomics.

In addition to demonstrating the improvement gained through localized deconvolution on synthetic data, we analyzed its effect on quantifying a study of toxicity, as measured by subsequent pattern recognition methods. Specifically, we observed an improvement of 16% in the cross-validated measure Q^2 during the application of a standard supervised learning method, orthogonal projection onto latent structures (O-PLS). The Q^2 metric improved from 0.7569 to 0.8782 after applying localized deconvolution (Region (+/-)). The improved Q^2 metric can be attributed to removing within group variability. Figure 3 shows this improvement in the projected space used to separate the two groups (48 hrs, 20 mg/kg and 0 hrs, Control). The x-axis is representative of the signal responsible for the difference in the groups. The y-axis is signal uncorrelated to the difference in the groups. The tightening of the within group variability on the x-axis leads to the improvement of the Q^2 metric.

The adoption of a general purpose high-throughput quantification method by the metabolomics community is dependent on its ease of applicability. This can be broken into two parts: speed and flexibility. By providing access via RESTful web interface, we are providing a resource that can be incorporated in scientific workflows and other quantification methods. Using a map-reduce framework allows us to parallelize the deconvolution procedure and run the process at a rapid rate. The running time is dependent on the number of mappers, which is shown in Figure 4. On a moderately sized cluster with 15 nodes, it requires approximately 4 minutes to complete a detailed deconvolution of five congested ^1H spectra from the data using 20 mappers.

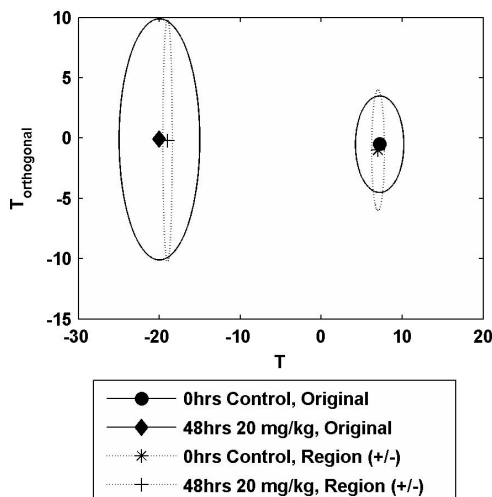


Fig. 3: O-PLS results showing the separation between 0 hrs, Control and 48 hours after a 20 mg/kg dose using 10 fold cross-validation

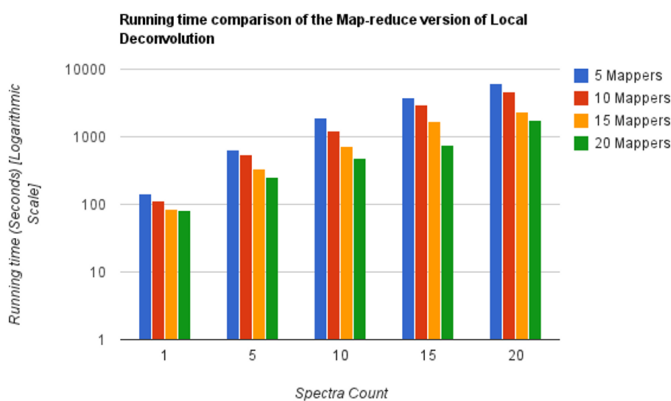


Fig. 4: The running time required to quantify different ^1H spectra as a function of the number of mappers

In our implementation, we set the default number of mappers at 20 since it seemed to provide reasonable running times for the typical file sizes encountered in our experimental set up; however, for larger files, higher number of mappers definitely makes an improvement and can be set accordingly by passing the relevant parameter.

5. Conclusion

In conclusion, we have shown that localized deconvolution is a robust method to process highly congested spectra that improves accuracy over standard high-throughput quantification methods. Our algorithm is naturally decomposed into concurrent tasks which are implemented in a map-reduce paradigm with a Web-service interface, thus, providing a scalable and accessible tool for the metabolomics community.

Our experiments have shown that the removal of adjacent, convoluting, and irrelevant signals results in significantly improved absolute and relative quantification, as demonstrated on realistic synthetic data. The performance metrics also demonstrate that including a buffer region does not improve overall accuracy, and allowing the baseline to be positive or negative results in the best accuracy. However, it was observed that specific spectral configurations did benefit from including a buffer region. Developing an algorithm to take advantage of the strengths of both methods is currently in process.

The advantages of our method were also observed on an experimental metabolomics data set of organ toxicity. Specifically, the within group scatter was reduced by localized deconvolution, resulting in an improved cross-validation score (Q^2); however, this increase in accuracy leads to additional computing costs. Such issues can easily be overcome by parallelizing the process with map-reduce and making use of cheaply available cloud resources. While our method provides a significant improvement over standard binning methods, alternative techniques that rely on annotated spectral databases, such as targeted and direct quantification methods, can also improve their accuracy by filtering and removing obfuscating signals with localized deconvolution.

6. Acknowledgement

We would like to acknowledge the Kno.e.sis Cloud Computing Collaboratory for providing computing resources: <http://knoesis.wright.edu/aboutus/infrastructure/cloud>.

References

- [1] N. V. Reo, "NMR-based metabolomics," *Drug and Chemical Toxicology*, vol. 25, no. 4, pp. 375–382, 2002.
- [2] J. C. Lindon, E. Holmes, and J. K. Nicholson, "Pattern recognition methods and applications in biomedical magnetic resonance," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 39, no. 1, p. 1, 2001.
- [3] J. K. Nicholson, J. C. Lindon, and E. Holmes, "Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," *Xenobiotica*, vol. 29, no. 11, p. 1181, 1999.
- [4] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.
- [5] H. Martens and T. Naes, *Multivariate Calibration*. London: Wiley, 1989.
- [6] S. C. Connor, R. A. Gray, M. P. Hodson, N. M. Clayton, J. N. Haselden, I. P. Chessell, and C. Bountra, "An NMR-based metabolic profiling study of inflammatory pain using the rat FCA model," *Metabolomics*, vol. 3, no. 1, pp. 29–39, 2007.
- [7] T. L. Whitehead, B. Monzavi-Karbassi, and T. Kieber-Emmons, "1H-NMR metabonomics analysis of sera differentiates between mammary tumor-bearing mice and healthy controls," *Metabolomics*, vol. 1, no. 3, pp. 269–278, 2005.
- [8] P. E. Anderson, N. V. Reo, N. J. DelRaso, T. E. Doom, and M. L. Raymer, "Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics," *Metabolomics*, vol. 4, no. 3, pp. 261–272, 2008.
- [9] P. Anderson, D. Mahle, T. Doom, N. Reo, N. DelRaso, and M. Raymer, "Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data," *Metabolomics*, pp. 1–12, 2011. [Online]. Available: <http://www.springerlink.com/index/C5NP143U061K0585.pdf>
- [10] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson, "Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform," *Chemometrics & Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 144–154, 2007.
- [11] K. M. Åberg, E. Alm, and R. J. O. Torgrip, "The correspondence problem for metabonomics datasets," *Analytical and Bioanalytical Chemistry*, vol. 394, pp. 151–162, 2009.
- [12] J. Forshed, R. J. Torgrip, K. M. Åberg, B. Karlberg, J. Lindberg, and S. P. Jacobsson, "A comparison of methods for alignment of NMR peaks in the context of cluster analysis," *J Pharm Biomed Anal*, vol. 38, no. 5, pp. 824–832, 2005.
- [13] D. J. Crockford, H. C. Keun, L. M. Smith, E. Holmes, and J. K. Nicholson, "Curve-Fitting Method for Direct Quantitation of Compounds in Complex Biological Mixtures Using 1H NMR," *Application in Metabonomic Toxicology Studies*, *Analytical Chemistry*, vol. 77, no. 14, pp. 4556–4562, 2005.
- [14] Q. Zhao, R. Stoyanova, S. Du, P. Sajda, and T. R. Brown, "HiRes - a tool for comprehensive assessment and interpretation of metabolomic data," *Bioinformatics*, vol. 22, no. 20, pp. 2562–2564, 2006.
- [15] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky, "Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data," *Analytical Chemistry*, vol. 78, no. 13, pp. 4430–4442, 2006.
- [16] D. Mahle, P. Anderson, and N. DelRaso, "A generalized model for metabolomic analyses: application to dose and time dependent toxicity," *Metabolomics*, 2011. [Online]. Available: <http://www.springerlink.com/index/H7861V6218327H10.pdf>
- [17] B. K. Alsberg, A. M. Woodward, and D. B. Kell, "An introduction to wavelet transforms for chemometricians: A time-frequency approach," *Chemometrics & Intelligent Laboratory Systems*, vol. 37, no. 2, p. 215, 1997.
- [18] H. F. Cancino-De-Greiff, R. Ramos-Garcia, and J. V. Lorenzo-Ginori, "Signal de-noising in magnetic resonance spectroscopy using wavelet transforms," *Concepts in Magnetic Resonance*, vol. 14, no. 6, pp. 388–401, 2002.
- [19] K. Kaczmarek, B. Walczak, S. de Jong, and B. G. Vandeginste, "Preprocessing of two-dimensional gel electrophoresis images," *Proteomics*, vol. 4, no. 8, p. 2377, 2004.
- [20] C. Perrin, B. Walczak, and D. L. Massart, "The Use of Wavelets for Signal Denoising in Capillary Electrophoresis," *Anal. Chem.*, vol. 73, no. 20, pp. 4903–4917, 2001. [Online]. Available: http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ac010416a
- [21] X. G. Shao, A. K. Leung, and F. T. Chau, "Wavelet: a new trend in chemistry," *Accounts of Chemical Research*, vol. 36, no. 4, p. 276, 2003.
- [22] H. Grage and M. Akke, "A statistical analysis of NMR spectrometer noise," *Journal of Magnetic Resonance*, vol. 162, no. 1, pp. 176–188, 2003.
- [23] K. Gunaratna, P. Anderson, A. Ranabahu, and A. Sheth, "A study in Hadoop streaming with MATLAB for NMR data processing," in *Proceeding of 2nd IEEE International Conference on Cloud Computing (Cloudcom)*, Indianapolis, IN, 2010.
- [24] A. Manjunatha, P. Anderson, A. Ranabahu, and A. Sheth, "Identifying and Implementing the Underlying Operators for Nuclear Magnetic Resonance based Metabolomics Data Analysis," in *Proceedings of 3rd International Conference on Bioinformatics and Computational Biology (BICoB)*, New Orleans, LA, 2011.
- [25] J. Trygg and S. Wold, "O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter," *Journal of Chemometrics*, vol. 17, no. 1, pp. 53–64, 2003.

Simulating Anaesthetic Effects on a Network of Spiking Neurons with Graphics Processing Units

A. Leist, C.J. Scogings and K.A. Hawick

Computer Science, Institute for Information and Mathematical Sciences,
Massey University, North Shore 102-904, Auckland, New Zealand

Email: { a.leist, c.scogings, k.a.hawick }@massey.ac.nz

Tel: +64 9 414 0800 Fax: +64 9 441 8181

Abstract—*The structure and emergent behaviours of neuronal networks remain important unknowns but can be investigated by computer simulation of biologically plausible networks of microscopically simple individual neurons. We describe a software model developed to simulate in excess of 10^6 individual Izhikevich neurons with connectivities of in excess of 100 connections per neuron. We simulate the effect of adjusting some of the microscopic neuronal parameters and observe emergent oscillatory phenomena that relate to the introduction of anaesthetic drugs on the collective neuronal system. We report some preliminary computational performance results and comment on the feasibility of simulating realistic sized collective networks of cortical spiking neurons.*

Keywords: GPU; CUDA; spiking cortical neurons; anaesthesia

1. Introduction

The problem of understanding neuronal processes and structures in the brain [1], [2] is a long standing one. Of particular interest are those emergent collective properties that are thought to arise from the complex network structure [3] and nature of the brain rather than necessarily from microscopic details of individual neurons. One process of particular interest is the manner in which cortical neural activity [4] rises and falls in states of consciousness. Campbell and others have suggested an intriguing way to study this through simulating the action of an anaesthetic drug [5] on individual neurons [6], [7], [8] linked together in an artificial network structure.

Although there are many outstanding questions and unknowns concerning real brain structure we have constructed a simulated neural network structure on the assumption that there are likely some emergent properties that we may observe due to the sheer size of a suitably simulated network of many interacting individual neurons. In this paper we discuss some preliminary simulation software development work in scoping the computational feasibility of simulating large ensembles of individual neurons [9] that are arranged in structures and with connectivities that are at least plausible if not biologically justified in detail [10].

In particular we describe our use of massively data parallel computing techniques and graphical processing units (GPUs) with many individual cores, to simulate around 10^6 individual neurons arranged in regular and small-world interconnected networks [11], [12], [13]. We focus on the use of the Izhikevich neural model [14] of cortical spiking neurons [15] for the work reported here. Our simulation software apparatus could be readily adapted to use other neuronal models such as the Hodgkin-Huxley neuronal model [16].

Our article is structured as follows: In Section 2 we summarise the properties of the individual neuronal model we use. The simulated network structures are described in Section 3 and details of our data-parallel Compute Unified Device Architecture (CUDA) implementations for GPUs are given in Section 4. We discuss some of the emergent properties and features of our simulated system in Section 6, in which we also offer some tentative conclusions and suggested areas for further work.

2. Neuronal Model

We use the Izhikevich model [14] to simulate the spiking and bursting behaviour observed in cortical neurons [17]. Although more biophysically meaningful models do exist, including the well known Hodgkin-Huxley [16] model, the computational demands of these models are significantly higher and the size of the simulated neuronal networks is thus more limited. The aim of this paper is to lay the computational foundations for further studies of large-scale neuronal networks, specifically in terms of the spatial and temporal effects of anaesthetic drugs on the neural interactions. Size matters for simulations of such complex systems, as some macroscopic behavioural patterns only emerge for large systems with many microscopic interactions or when system properties are analysed over several length-scales. This is not to say that the quality of the model is not relevant, of course, as a system that does not exhibit realistic behaviour is essentially useless. However, in [15], Izhikevich compares a number of commonly used models and shows that the model proposed in [14] is computationally much cheaper than the Hodgkin-Huxley model, but nevertheless capable of reproducing realistic spiking and bursting behaviour.

The model uses four parameters, which can be adjusted to produce different spike patterns, such as those observed for real excitatory and inhibitory neurons. These spikes – which are also called action potentials – produce an electrochemical impulse that is transmitted to connected cells. Action potentials created by excitatory cells depolarise the membrane potentials v of their neighbouring cells and, thus, decrease the distance to their spike thresholds, making them more likely to “fire” an action potential of their own. Spikes created by inhibitory cells, on the other hand, hyperpolarise the membrane potentials and increase the distance to the threshold. Neurons are connected through transmitting fibres called *axons* and receiving fibres called *dendrites*. The *synapse* is the axon-dendrite junction. A postsynaptic neuron receives a postsynaptic potential after a suitable delay δ from the time the presynaptic action potential has been

generated. But the postsynaptic potential does not apply all at once, it rather diminishes exponentially over a period of time as suggested in [18]. Our implementation uses separate washout tables $W_{e,i}$ for excitatory and inhibitory neurons, which define the gradual washout as $W(t) = Ae^{-t/\tau}$, where t is time measured in simulation steps, $A_{e,i}$ regulates the voltage amplitude and $\tau_{e,i}$ defines the exponential washout rate. The washout table is chosen to be of length 3τ , which allows the action potential to diminish to about 5% of its base value before it stops having any effect.

In addition to the incoming currents from action potentials generated by presynaptic neurons, every neuron also receives an input current I . This is used to simulate currents received from sources external to the cortex, for example other parts of the brain – like the thalamus – or any other part of the nervous system. I is calculated individually for every neuron and at every time step as $I = I_{e,i} + I_{noise} + I_{boost}$, where $I_{e,i}$ is the base current for all excitatory (I_e) or inhibitory (I_i) cells. $I_{noise} = n_{e,i} \times r_{norm}$ is random noise computed from a base noise value $n_{e,i}$, times a normally distributed random variable. $I_{e,i}$ and I_{noise} are used to simulate a constant source of activity that drives the cortex. I_{boost} , on the other hand, is meant as a temporary boost, like an external shock delivered to the neuronal network. It is applied to a fraction of all cells selected at random during each simulation step that the boost is active. Although I_{noise} can be negative, the sum of these inputs I is not allowed to fall below zero.

The effects of anaesthetic drugs are modelled using parameters $\lambda_{A(e,i)}$, $\lambda_{\tau(e,i)}$ and $\lambda_{I(e,i)}$, which are defined individually for excitatory and inhibitory neurons. Different combinations of these values can be used to simulate different types of anaesthetics. The λ values can be updated between simulation steps. They modify the corresponding base values to get the effective values as follows:

For excitatory neurons:	For inhibitory neurons:
$A = A_e / \lambda_{Ae}$	$A = A_i \lambda_{Ai}$
$\tau = \tau_e / \lambda_{\tau e}$	$\tau = \tau_i \lambda_{\tau i}$
$I = I_e - \lambda_{Ie} + 1$	$I = I_i + \lambda_{Ii} - 1$

Thus, values of $\lambda > 1$ simulate drugs that have a dampening effect when applied to excitatory neurons and a strengthening effect when applied to inhibitory neurons. No drugs are administered when $\lambda = 1$.

3. Network Model & Data Structure

A 2-dimensional lattice is used to assign a unique global ID to each neuron – which can be calculated from its (x, y) -coordinates – and to restrict connections between neurons to a maximum distance r that is defined by the user. The one-way nerve connections are generated at random using a normal distribution that is centered on the postsynaptic neuron, that is, the end of the arc in graph terminology. The standard deviation of the distribution is set to $\sigma = r/3$ and the distance is strictly restricted to $\leq r$. This results in a network structure where most connections are relatively short, with a decreasing number of longer distance connections. While the number of outgoing connections varies between neurons, every neuron has the same number of incoming connections. This, together with the fact that arcs are stored

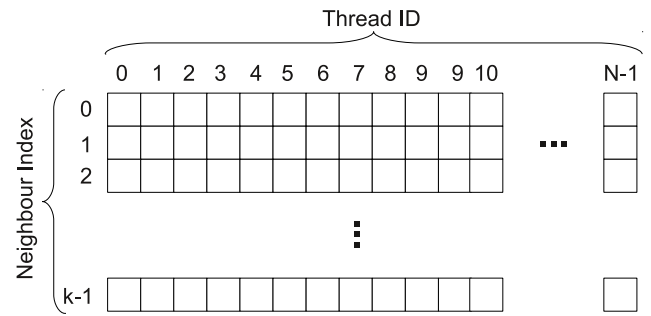


Fig. 1: The one-way connections between neurons are stored in a 1D array of length $N \times k$ – here illustrated as a 2D array with N columns and k rows – where N is the system size and k is the in-degree. Each one of these arcs is only identified by the global neuron ID of its source node. The destination node is determined by the position in the array. As neurons are processed in the order of the thread IDs given in Figure 2, the source IDs are stored at index $n_{idx} \times N + tid$, where n_{idx} is the n 's neighbour of the neuron that is processed by the thread with ID tid . This ensures fully coalesced memory transactions when accessing this array.

in the adjacency-list of the postsynaptic cell as described in the caption of Figure 1, is an important optimisation that significantly improves the utilisation of the GPU's memory bandwidth. It makes it possible to use an information pull-model for the transmission of action potentials. This process is described in more detail in Section 4. Note that even though the neurons are addressable by their coordinates in the grid, the graph itself is far from regular, due to the way neighbours are selected. It is possible to change the graph structure without any modifications to the algorithm, as long as the invariant of having the same number of incoming connections per neuron is maintained. This flexibility allows for plenty of future experimentation and a small-world network [19], [12] may be a particularly interesting candidate.

The lattice structure is also used to determine which CUDA thread gets to process a particular neuron when it is time to evolve the simulation by another step. This mapping is illustrated in Figure 2. It is chosen to maximise the data locality for threads within the same thread block when querying information from neighbouring cells, taking advantage of the texture cache available on CUDA GPUs¹ and the knowledge that cells located close to each other on the lattice are more likely to share some of their neighbours than cells separated by a larger distance.

Whether a neuron is excitatory or inhibitory as well as the exact parameter values that determine its spike patterns are all determined when the graph is generated using the approach suggested in [14]. This leads to an approximately 4:1 ratio of excitatory to inhibitory cells. The type information needs to be available when the state of a neuron is queried, because postsynaptic neurons need to know what effect an incoming action potential has on their membrane potential. To do this efficiently, the type and the current firing state of each neuron are bit-packed into a single byte. All these bytes are stored in a two-dimensional array D_d that is addressed using a cell's

¹See [20], [21] for details about the CUDA architecture.

0	1	2	3	16	17	18	19
4	5	6	7	20	21	22	23
8	9	10	11	24	25	26	27
12	13	14	15	28	29	30	31
32	33	34	35	48	49	50	51
36	37	38	39	52	53	54	55
40	41	42	43	56	57	58	59
44	45	46	47	60	61	62	63

Fig. 2: The mapping of CUDA threads to neurons. The numbers represent the ID of the CUDA thread (tid) that processes a particular neuron. The actual implementation uses blocks of size 16×16 instead of the 4×4 blocks shown here. Independent from the tid , the 2D lattice is used to assign a global ID to each neuron. This global ID is implicitly defined by the position in the grid using row-major ordering from the top left to the bottom right. Note that the lattice structure does not reflect the actual neuronal network, it is only used to assign a unique global ID to each neuron and to restrict the neighbour selection to a given radius r .

(x, y) -coordinates. 2D texture fetches are used to retrieve the data when iterating over the adjacency-lists to identify any incoming action potentials. As only the two least-significant bits are used, the data could be compressed even more by storing the information of four neurons in a single byte. However, this is not done in the current implementation, as it is expected that future versions will make use of this space to store additional state information.

While all elements in D_d are read many times during each simulation step, once for every nerve connection originating from the respective neuron, the following arrays are all used to store data that is only read and updated by the thread that processes the cell it belongs to:

- V_d and U_d record the current values of the Izhikevich variables v (membrane potential) and u (membrane recovery).
- Δ_d stores the synaptic delay δ for each neuron. A minimum delay of $\delta_{min} = 5$ and a maximum delay of $\delta_{max} = 15$ milliseconds are used in the current implementation. The delay is defined on the postsynaptic neuron and not on the link itself. This simplification reduces the memory requirements for the delay terms from $\mathcal{O}(Nk)$ to $\mathcal{O}(N)$. The values are randomly initialised within the given range.
- EV_d is the extra volts array. It records the effective input voltage from action potentials – both excitatory and inhibitory – generated by all presynaptic neurons over the last $w + \delta$ simulation steps, where w is the current washout table length. Because $w = 3\tau$ and τ can be modified by $\lambda_{\tau(e,i)}$, a maximum washout table

length W_{max} is defined at compile time. Based on this, $EV_{max} = W_{max} + \delta_{max}$ space is allocated for every neuron.

- TV_d records the type values that define the spike dynamics of individual neurons.

As each neuron only requires its own data, these arrays are indexed using the thread ID tid and transactions are fully coalesced. The only exception to this are transfers to and from EV_d during `Phase1` of the simulation, which are only partially coalesced. The reason being that the index used to access EV_d during this phase depends on the neural delay δ , which is initialised randomly for each cell.

Next, memory for T instances of the CUDA implementation of the 64-bit random number generator `Ran` from Numerical Recipes [22] is allocated. T is the number of CUDA threads used to process the system. The optimal value for T depends on the execution hardware, but it should be a large power of two smaller or equal to the system size, which is always a power of two itself. Every thread thus processes an equal number of neurons. $T = 2^{19}$ is used for the performance measurements, except when $N < 2^{19}$, in which case $T = N$.

Finally, arrays V_{sum} and F_{sum} are used to record partial sums of the membrane potentials and firing rates of all excitatory neurons. For reasons explained in the following section, these arrays are of length $32 \times (\text{number of thread blocks})$.

4. CUDA Implementation

The simulation is split into two distinct phases and each of these phases is implemented as a CUDA kernel. Every simulation step executes both phases, advancing the simulation by one millisecond of model time. The control flow and the secondary tasks processed by the host system are described by the pseudo-code in Algorithm 1. The main task of `Phase1` is to update the membrane potentials v using the equations proposed by Izhikevich [14] for all neurons based on the various input currents reaching each cell at the current time step. This includes the external currents modelled by I as well as the inputs from all presynaptic neurons that have generated an action potential during time steps $[-(\delta + w), -(1 + \delta)]$ from the current time as recorded in EV_d . When a neuron's membrane potential reaches 30mV, then it fires a new action potential. This event is recorded by setting the respective bit in array D_d .

The only data that needs to be moved between the host and the device at every time step – not counting kernel parameters – are arrays V_{sum} and F_{sum} . The average membrane potential of all excitatory neurons is used to plot a pseudo-EEG and to compute the power spectral density. The firing rate is plotted as an additional visual indicator of the current cortical activity. Making the data immediately available to the host makes it possible to monitor the simulation as it is running and to store the generated data for later reuse. As mentioned before, both of these arrays are of length $32 \times (\text{number of thread blocks})$ and not of length N . The reason for this is that kernel `Phase1` performs a parallel reduction to compute the sum of the respective values for all neurons processed by the threads that belong to the same thread block. The reduction is done in the fast on-chip shared memory and the process is only stopped when the number of elements in the

Algorithm 1 This host function is called to evolve the simulation by STEPS simulation steps. The generated values v_{avg} and f_{avg} are the average membrane potential and the average firing rate of all excitatory neurons at the current time step. The former can be used to plot a pseudo-EEG and to compute the power spectral density.

determine the current washout table length w

for $s \leftarrow 1$ to STEPS **do**

do in parallel on the device using T threads: call kernel Phase1(EV_{idx})

wait until Phase1 is completed

increment EV_{idx} , the index into the extra volts array EV_d //wraps around when it reaches the end of the array

copy V_{sum} from device memory to host memory //asynchronous, can overlap with kernel Phase2

copy F_{sum} from device memory to host memory //asynchronous, can overlap with kernel Phase2

do in parallel on the device using T threads: call kernel Phase2(EV_{idx}, w)

wait until V_{sum} and F_{sum} have been copied to host memory

$v_{avg} \leftarrow f_{avg} \leftarrow 0$ //the average voltage v and firing rate f of all excitatory neurons

for $i \leftarrow 0$ to $(32 * [\text{number of thread blocks}])$ **do**

$v_{avg} \leftarrow v_{avg} + V_{sum}[i]$

$f_{avg} \leftarrow f_{avg} + F_{sum}[i]$

end for

$v_{avg} \leftarrow v_{avg}/n_{type0}$ // n_{type0} is the number of excitatory neurons

$f_{avg} \leftarrow f_{avg}/n_{type0}$

wait until Phase2 is completed

end for

input reaches the warp size of 32 threads, at which point the first warp in the thread block writes the partial sums to V_{sum} and F_{sum} respectively. These arrays are then copied to host memory, where the CPU can sequentially perform the remaining summation. The memory copies and CPU processing can be overlapped with the execution of kernel Phase2 and, therefore, do not add to the overall runtime.

In Phase2, every neuron queries the current state of all its neighbours using texture fetches from array D_d as discussed in the previous section. To be able to do this, the global IDs of the presynaptic neurons – which can be used to compute the textures coordinates – are looked up from each neuron's adjacency-list using fully coalesced data transfers as explained in Figure 1. The data from D_d is then used to determine the number of new excitatory and inhibitory action potentials generated by all neighbours. Then, the next w values of the extra volts array EV_d are updated according to the number of inputs, using the values provided in the washout tables $W_{e,i}$ to compute the resulting excitatory and inhibitory effects over time. The fact that all threads update the next w elements of their neuron's extra volts array, starting from the same offset into EV_d , is very important. It means that all w reads and w writes per neuron performed during this phase are fully coalesced. This easily makes up for the single read and write per cell that is only partially coalesced in Phase1. The washout tables are stored in constant memory and can be accessed very quickly. The entire process of using the extra volts array is visualised in Figure 3.

As the performance results given in the next section show, the system size is mainly limited by the on-board memory of the GPUs used to run the simulation. In order to be able to simulate much larger systems, or to reduce the execution time, a multi-GPU implementation has been developed. It can be executed either on a single host machine with multiple GPUs or on a cluster of machines with one or more GPUs each. OpenMPI is used for the communication between cluster nodes. Figure 4 describes how the neuronal network is split into multiple components and how the communication time between GPUs can be at least partially hidden by computation.

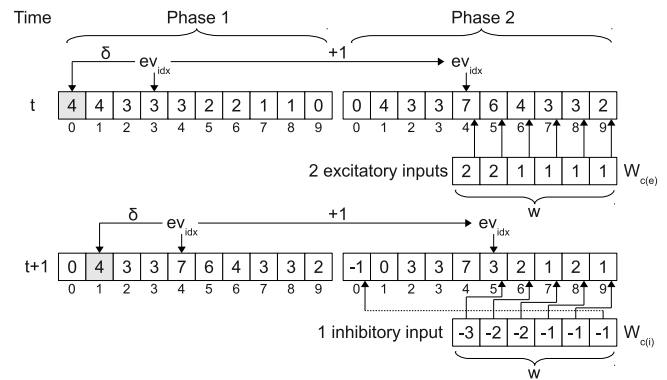


Fig. 3: This diagram illustrates how the extra volts array EV_d for a single neuron is indexed and modified over the two phases of the simulation. Note that the actual implementation interleaves the arrays of all neurons and uses a stride of N between indices to facilitate coalesced memory transactions. The neuron's delay is $\delta = 3$ and the current washout table length is $w = 6$ in this example. At time step t , kernel Phase1 reads the input value from index $ev_{idx} - \delta = 3 - 3 = 0$, computes the new membrane potential and resets the value in EV_d to 0. Index ev_{idx} is incremented before kernel Phase2 is called. This kernel first queries all neighbours and finds that two of them are firing an excitatory action potential. It then proceeds to add $2 \times$ the values from the excitatory washout table $W_{c(e)}$ to the corresponding w elements of the extra volts array, beginning with index $ev_{idx} = 4$. The next simulation step $t+1$ repeats this procedure, but finds that the neuron is now receiving a single inhibitory input. Phase2 thus adds $1 \times$ the values from the inhibitory washout table $W_{c(i)}$ to the correct values in EV_d .

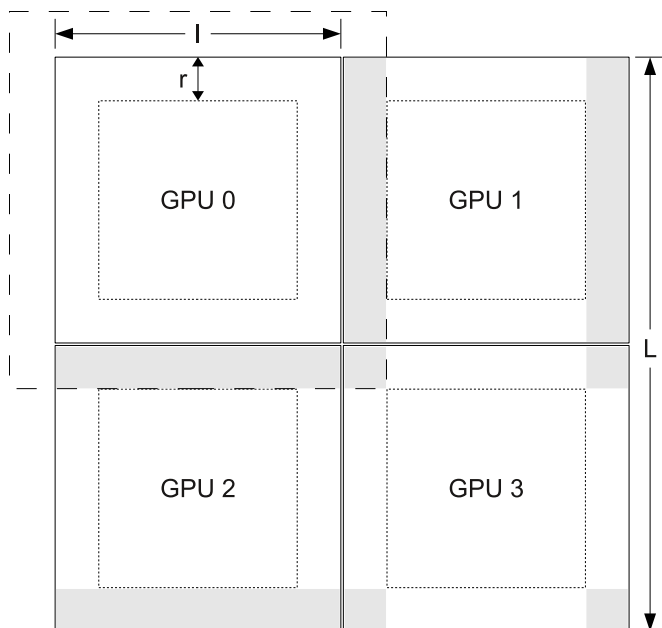


Fig. 4: This figure illustrates how the neurons are divided up to be processed in a multi-GPU setup. The maximum neighbour distance r determines the size of the border region that needs to be exchanged between devices processing neighbouring components before Phase2 can be completed. The neurons located in the core region, however, can be processed independently from all other devices. To exploit this, each GPU concurrently processes its core region and performs the data exchange to obtain the information for its local border on devices that support this feature.

Please refer to [23] for more details about the CUDA implementations of the different simulation phases and the modifications necessary for the cluster implementation. The reference provides detailed code listings and additional information about the model and its configuration options. It also describes the signal processing techniques that are used to obtain a power spectrum from the generated pseudo-EEG.

5. Performance Results

This section shows how the CUDA implementation performs when executed in batch-mode. In this mode, the results are not visualised in real-time. Instead, the average membrane potentials and firing rates are written to a file. While the graphical user interface is very useful when testing the effects of certain parameter combinations on the model, more extensive parameter value range scans are generally performed in batch-mode, with an automatic extraction of relevant metrics following the simulation run. A configuration file can be used to specify the exact settings for each time step. Most importantly, the values for each of the λ parameters can be modified to define the beginning and end of periods during which a particular drug effect is being simulated. This mode also gives a more accurate measure of the actual performance of the simulation code itself.

A number of different GPUs are used to compare the execution speed on two generations of CUDA devices. The GTX260 provides 896 MB of device memory and 216 CUDA cores and represents the GT200 series of GPUs. All other

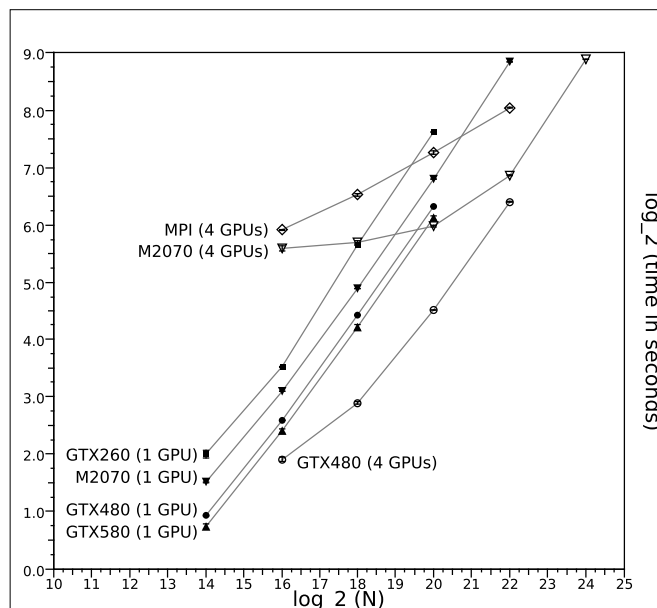


Fig. 5: The execution times for 10,000 simulation steps with system sizes ranging from $N = 2^{14}$ to $N = 2^{24}$. The in-degree is set to $k = 100$, which adds up to a maximum of $\approx 1.68 \times 10^9$ neural connections in the largest system. The maximum neighbour distance $r = 64$.

devices are based on the Fermi-architecture. The GTX480 and GTX580 provide 1536 MB of memory and have a total of 480 and 512 CUDA cores respectively. The clock speeds and memory bandwidth of the GTX580 are approximately 8–10% higher than those of the GTX480. The professional M2070 offers a full 6 GB of device memory, 448 CUDA cores and uses a slightly more moderate clock speed and bandwidth. All results presented in this section are averaged over 10 independent simulation runs. Error bars representing the standard deviations are smaller than the symbol size.

Figure 5 shows the results for a range of system sizes with fixed in-degree $k = 100$ for all neurons. The maximum neighbour distance $r = 64$. Not every system size can be simulated on all devices due to the different amount of memory available on each GPU. The multi-GPU implementations require a system size of $N \geq 256^2$, as the local dimension length l has to be at least $2r$. Unsurprisingly, the GTX580 is the fastest GPU. It is capable of computing one second of simulated time (1000 time steps) in the cortical model with $N = 262,144$ neurons and over 26 million neural connections in about 1.85 seconds of real time. Although the M2070 is slower than the GTX580, its large DRAM makes it possible to process systems of up to 4.2 million neurons on a single device.

The setup using four GTX480 GPUs, all installed in separate x16 PCIe slots of the same host system, shows how well the multi-GPU implementation scales given a large enough system size. They complete the simulation 3.5 times faster than a single GTX480 when running the largest system supported by both configurations. Both the MPI implementation and the node with four M2070s, which have to share a single x16 PCIe bus, scale very well, but have a much higher communication overhead. They perform best when the system

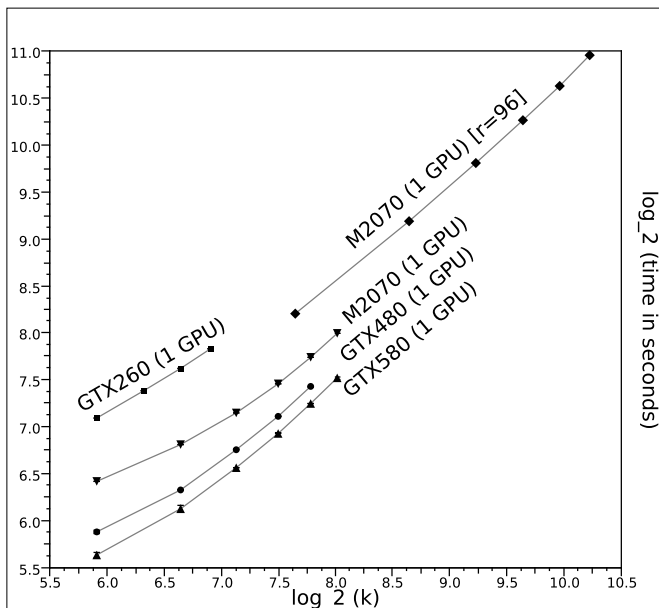


Fig. 6: The execution times for 10,000 simulation steps with in-degrees ranging from $k = 60$ to $k = 1200$. The system size is $N = 2^{20}$ and the maximum neighbour distance is $r = 64$, except for the second result set for the M2070 ($k = 200$ to 1200), which uses $r = 96$.

size is large, as the ratio of neurons located in the border regions to those located in the core regions decreases, which enables the devices to overlap more of the data transfer times with computation. The large amount of device memory on the four M2070s makes it possible to process a system of over 16 million neurons with a total of over 1.6 billion neural connections.

Figure 6 shows the execution times for various in-degrees, with a constant system size of $N = 2^{20}$ neurons and a maximum neighbour distance $r = 64$ where not explicitly marked otherwise. No results are given for multi-GPU configurations, as they can only increase the degree at the expense of the sub-system size processed by individual devices. The results offer no surprises, except that the GTX480 runs out of memory when $k = 260$, whereas the GTX580, which is supposed to have the same amount of device memory, completes the simulation successfully. For the large degrees of up to $k = 1200$ that are possible with the M2070, a value of $r = 96$ is used to increase the size of the pool of possible neighbours. This also shows the effect of the neighbour distance on the performance, as a larger value of r means that the texture fetches are less likely to result in a cache hit.

6. Discussion & Conclusions

We have proposed a data-parallel implementation of a neural network model that is based on Izhikevich type neurons. The model is designed with the intention to simulate and analyse neural processes involved in anaesthesia. We are interested in large scale simulations with millions of neurons and many more neural connections to facilitate emergent behaviour that may not be visible at smaller scales. The implementation described in this article merely lays the

algorithmic foundation for further studies, as a significant computational effort is still required to find parameter value combinations that work well together and produce realistic neural activity patterns. This is particularly difficult, as many of the system parameters are correlated with each other, which leads to disproportionately strong reactions to relatively small parameter changes.

The implementation of the model also demonstrates how some inherently irregular problems can be optimised for the data-parallel architecture of modern GPUs. This is especially true when it is possible to design the model with this architecture in mind. One such example discussed here is the use of an information pull model for the transfer of action potentials along neural pathways and the decision to use the same in-degree for every neuron. Although this is not biologically realistic, we believe that it does not have a significant effect on the behaviour of the model, as the out-degrees remain randomly distributed. These decisions dramatically improve the simulation performance on GPUs, as they reduce the divergence of threads and increase the utilisation of the memory bandwidth.

As demonstrated, the proposed implementation can be used to simulate systems with more than four million neurons and one hundred times that many neural connections on a single CUDA capable graphics accelerator from the year 2011. A multi-GPU implementation that divides the computational workload and memory requirements between several devices has also been discussed. Although the simulation has only been tested with up to four devices, the implementation can scale to significantly larger compute installations.

The model offers a number of opportunities for future studies, such as the use of different graph structures – in particular small-world network based layouts – and the modelling of the delay terms directly on the neural connections. An extension of the model to include other regions of the mammalian brain would also be of interest.

Acknowledgements

We would like to acknowledge Dr. Douglas Campbell from Auckland Hospital, whose knowledge of the neuronal structure and chemical processes governing the mammalian cortex as well as the different phases observed in patients undergoing anaesthesia were invaluable to this project, and to thank him for suggesting it.

References

- [1] D. Lindley, "Computational neuroscientists are learning that the brain is like a computer, except when it isn't." *Communications of the ACM*, vol. 53, pp. 13–15, 2010.
- [2] L. F. Abbott, "Theoretical neuroscience rising," *Neuron*, vol. 60, pp. 489–495, 2008.
- [3] Y. Fregnac, M. Rudolph, A. P. Davison, and A. Destexhe, *Biological Networks*. World Scientific, 2007, ch. Complexity in Neuronal Networks, pp. 291–340.
- [4] R. Ananthanarayanan, S. K. Esser, H. D. Simon, and D. S. Modha, "The cat is out of the bag: cortical simulations with 10^9 neurons, 10^{13} synapses," in *Proc. Conf. on High Performance Computing, Networking, Storage, and Analysis, Portland, OR*, 14–19 Nov 2009.
- [5] N. P. Franks, "General anaesthesia: from molecular targets to neuronal pathways of sleep and arousal," *Nature*, vol. 9, pp. 370–386, May 2008.
- [6] D. Campbell, "Use of simulated anaesthesia effect of drugs on individual neurons," Private Communication, April 2009.

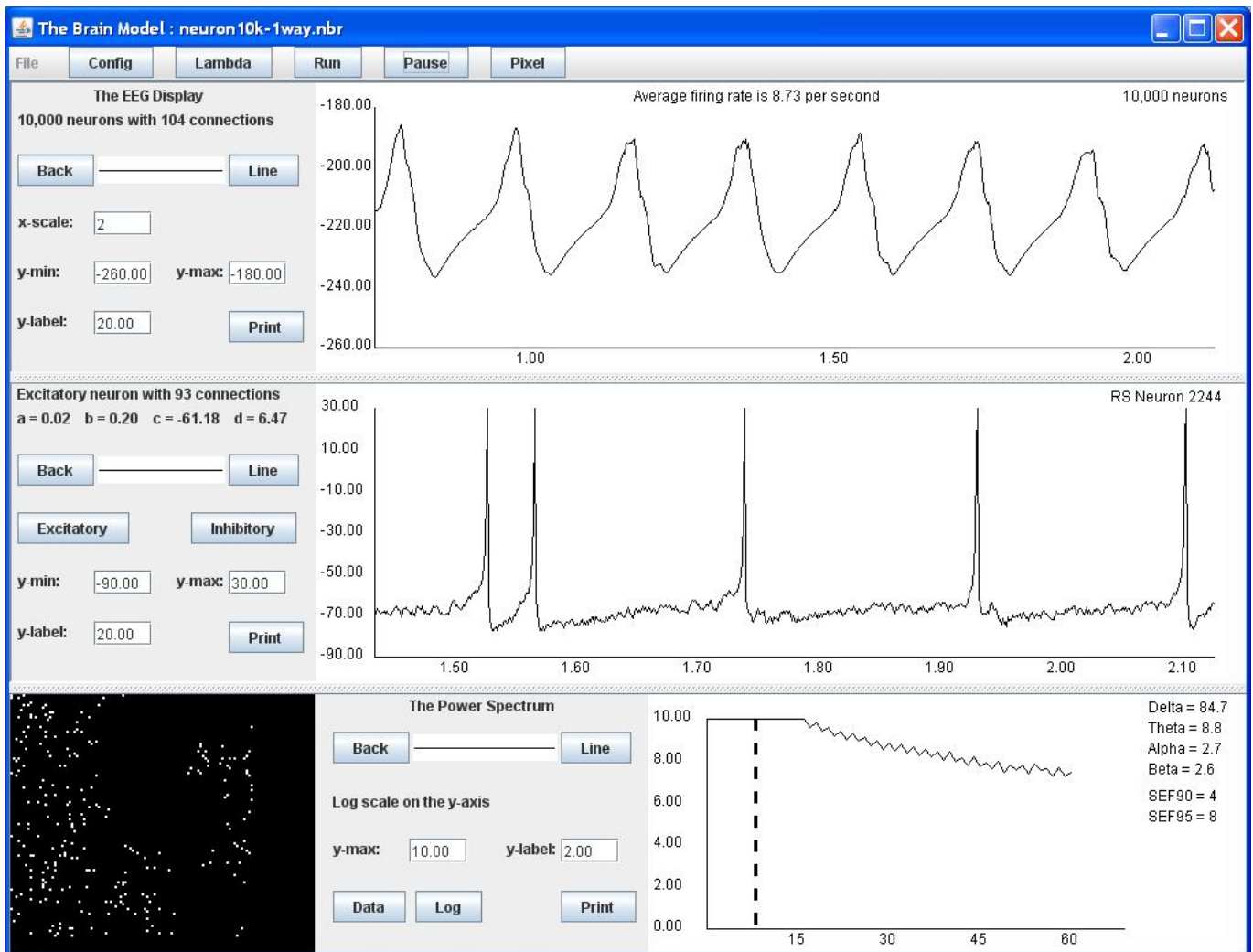


Fig. 7: This image shows one of the two GUIs developed to visualise the output generated by the simulation. While this version is written in Java, a C/OpenGL version is used to interface directly with the CUDA implementation of the algorithms. The plots show the pseudo-EEG (top), the membrane potential of a random neuron (middle) and the power spectrum (bottom). The field in the bottom left corner illustrates the action potentials generated by a 200×200 subset of the neurons in real-time. This field can be expanded to show all neurons, which has proven to be useful to visually observe firing patterns that may be of interest to the modeller.

- [7] D. A. Steyn-Ross, M. L. Steyn-Ross, L. C. Wilcocks, and J. W. Sleight, "Toward a theory of the general-anesthetic-induced phase transition of the cerebral cortex. ii. numerical simulations, spectral entropy, and correlation times," *Phys. Rev. E*, vol. 64, pp. 011918–1–12, 2001.
- [8] J. A. Talavera, S. K. Esser, F. Amzica, S. Hill, and J. F. Antognini, "Modeling the gabaergic action of etomidate on the thalamocortical system," *Anesth. Analg.*, vol. 108, pp. 160–167, 2009.
- [9] H. R. Wilson, "Simplified dynamics of human and mammalian neocortical neurons," *J. Theor. Biol.*, vol. 200, pp. 375–388, 1999.
- [10] F. Kepes, Ed., *Biological Networks*, ser. Complex Systems and Interdisciplinary Science. World Scientific, 2007, vol. 3, no. ISBN 978-981-270-695-9.
- [11] J. Kleinberg, "Small-world phenomena and the dynamics of information," in *Proc. Advances in Neural Information Processing Systems (NIPS) 14*, 2001.
- [12] D. S. Bassett and E. Bullmore, "Small-world brain networks," *The Neuroscientist*, vol. 12, pp. 512–523, 2006.
- [13] O. Sporns and C. J. Honey, "Small worlds inside big brains," *Proc. Nat. Acad. Sci.*, vol. 103, pp. 19219–19220, 2006.
- [14] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, November 2003.
- [15] —, "Which model to use for cortical spiking neurons?" *IEEE Trans. on Neural networks*, vol. 15, no. 5, pp. 1063–1070, September 2004.
- [16] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, pp. 500–544, 1952.
- [17] E. M. Izhikevich, "Neural Excitability, Spiking and Bursting," *International Journal of Bifurcation and Chaos*, vol. 10, no. 6, pp. 1171–1266, June 2000.
- [18] M. L. Steyn-Ross, D. A. Steyn-Ross, and J. W. Sleight, "Modelling general anaesthesia as a first-order phase transition in the cortex," *Progress in Biophysics & Molecular Biology*, vol. 85, pp. 369–385, 2004.
- [19] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.
- [20] *NVIDIA CUDA C Programming Guide Version 4.1*, NVIDIA® Corporation, 2011, <http://www.nvidia.com/> (last accessed April 2012).
- [21] A. Leist, D. Playne, and K. Hawick, "Exploiting Graphical Processing Units for Data-Parallel Scientific Applications," *Concurrency and Computation: Practice and Experience*, vol. 21, pp. 2400–2437, December 2009, CSTN-065.
- [22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes - The Art of Scientific Computing*, 3rd ed. Cambridge, 2007, ISBN 978-0-521-88407-5.
- [23] A. Leist, "Experiences in Data-Parallel Simulation and Analysis of Complex Systems with Irregular Graph Structures," Ph.D. dissertation, Massey University, Auckland, New Zealand, November 2011. [Online]. Available: <http://hdl.handle.net/10179/2992>

Improving Performance of a TFD-based Spectral Estimation Method in Doppler Ultrasound Blood Flow Measurement

F. García-Nocetti, J. Solano González, E. Rubio Acosta

Universidad Nacional Autónoma de México, IIMAS, México D.F., 04510, México

Abstract - A fundamental problem in Doppler Ultrasound blood flow measurement is the computation of the signal instantaneous frequency. The Cohen class of Time-Frequency Distributions (TFD) has efficiently determined a very close estimation of the instantaneous frequency for quasi-stationary signals such as those associated to arterial blood flow. Nevertheless, its computation has an $O(N^3)$ complexity, where N is the sample length. This imposes a great limitation when working with real-time systems. Previous work has proposed simplified expressions with comparable order of complexity. In this work a study is conducted to observe the response of different distributions when truncating the TFD's autocorrelation function. It also studies the relationship with the precision obtained in frequency estimation when considering different distribution kernels. SNR and sample length are considered in order to define a truncation procedure for minimizing RMS error. A real Doppler Ultrasound signal taken from a carotid artery is used for the performance evaluation.

Keywords: Time-Frequency Distributions, Signal Analysis, Doppler ultrasound, Blood flow measurement.

1 Introduction

A classical method for spectral estimation is the so-called Fourier Transform. However, its use is limited to stationary signals, giving as a result a poor frequency resolution when estimating non-stationary ones. Other types of spectral estimators, called time-frequency distributions, have been developed [2]. Unlike conventional methods, these distributions are not limited to the use of stationary signals. Despite of this important advantage, the number of calculations involved in obtaining the spectral estimation increases substantially compared to the traditional methods. Therefore, it is desirable to simplify the formulation of the distributions in such a way that the computations involved can be reduced without any loss in the spectral resolution. Simplified expressions that calculate the time frequency distributions have been previously introduced [1], [3], [11] and [12].

Also, previous works have suggested that a controlled truncation of the time frequency distributions' autocorrelation function does not significantly affect the accuracy when

estimating spectral parameters such as the pseudo instantaneous mean frequency and the RMS mean bandwidth [6], [7] and [8]. On the contrary, it further diminishes the amount of calculations involved. This strategy may be usefully in order to achieve efficient real time algorithms suitable to be implemented in high performance DSP architectures. This work accomplishes those studies and extends them with a performance evaluation using a real Doppler Ultrasound signal taken from the Carotid artery [6] and [7].

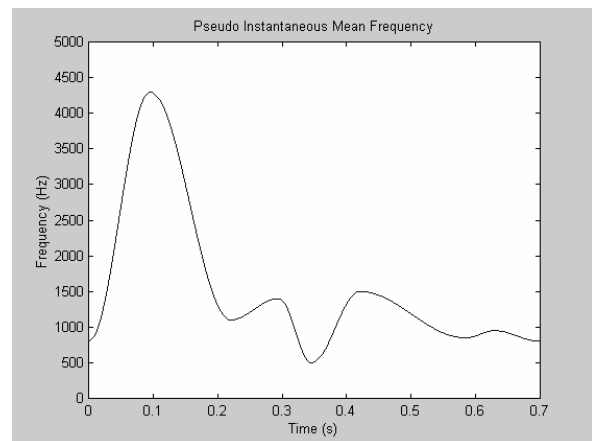


Figure 1. Signal's pseudo instantaneous mean frequency (PIMF) wave form of the simulated Doppler ultrasonic quasi-stationary signal that represents a typical blood flow in the Carotid artery.

2 Time-frequency distributions

The time-frequency distributions (TFD) of the Cohen class considered in this work are the Bessel, the Born Jordan and the Choi Williams distributions [2]. The discrete Bessel TFD when it is evaluated at discrete time zero and optimized is:

$$DBD(0, k, TI) = -2|x(0)|^2 + 4 \operatorname{Re} \left[\sum_{\tau=0}^{N-1} W(\tau) W^*(-\tau) e^{-j\frac{2\pi k \tau}{N}} \right] \cdot \sum_{\mu=\max\{-TI, -2\alpha| \tau|, -N+1+|\tau|\}}^{\min\{TI, 2\alpha| \tau|, N-1-|\tau|\}} \left(\frac{1}{\pi \alpha |\tau|} \sqrt{1 - \left(\frac{\mu}{2\alpha \tau} \right)^2} \right) x(\mu + \tau) x^*(\mu - \tau) \quad (1)$$

where k is the discrete frequency taking integer values from 0 to $N-1$, α is a scaling factor taking the half of any natural value, and $W(n)$ is a Hanning window of length $2N-1$ [5]. Note that $TI = N-1$; the exactly meaning of TI parameter is explained in section 3. The discrete Born Jordan TFD when it is evaluated at discrete time zero and optimized is:

$$DBJD(0, k, TI) = -2|x(0)|^2 + 4 \operatorname{Re} \left[\sum_{\tau=0}^{N-1} W(\tau) W^*(-\tau) e^{-j\frac{2\pi k\tau}{N}} \right] \cdot \sum_{\mu=\max\{-TI, -2\alpha|\tau|, -N+1+|\tau|\}}^{\min\{TI, 2\alpha|\tau|, N-1-|\tau|\}} \left(\frac{1}{4\alpha|\tau|} \right) x(\mu+\tau) x^*(\mu-\tau) \quad (2)$$

where k is the discrete frequency taking integer values from 0 to $N-1$, α is a scaling factor taking the half of any natural value, and $W(n)$ is a Hanning window of length $2N-1$ [2]. Note that $TI = N-1$. The discrete Choi-Williams TFD when it is evaluated at discrete time zero and optimized is:

$$DCWD(0, k, TI) = -2|x(0)|^2 + 4 \operatorname{Re} \left[\sum_{\tau=0}^{N-1} W(\tau) W^*(-\tau) e^{-j\frac{2\pi k\tau}{N}} \right] \cdot \sum_{\mu=\max\{-TI, -N+1+|\tau|\}}^{\min\{TI, N-1-|\tau|\}} \left(\sqrt{\frac{1}{4\pi\tau^2/\sigma}} e^{-\frac{\mu^2}{4\tau^2/\sigma}} \right) x(\mu+\tau) x^*(\mu-\tau) \quad (3)$$

where k is the discrete frequency taking integer values from 0 to $N-1$, σ is a scaling factor taking any positive real value, and $W(n)$ is a Hanning window of length $2N-1$ [3]. Note that $TI = N-1$.

3 Truncation Procedure

The inner summation respect to index μ in equations (1), (2) and (3) is the generalized time-index autocorrelation function of the TFD. Observe that the autocorrelation function has a factor that vanishes as index μ increases. As a consequence, a controlled truncation in the index μ results in a controlled decrement in the accuracy of TFD calculation. That controlled decrement in the accuracy of TFD calculation provokes a controlled increment in the spectral estimation errors but a decrement in the amount of calculations involved.

Such a truncation index (TI) has already been imposed in the inner summation respect to index μ in the equations (1), (2) and (3). The admissible values of TI are from 1 to $N-2$. Note that a TI values greater or equal than $N-1$ has a non-truncation effect.

Optimal scaling factors are considered in calculations [7]. These are presented in table 1.

Win. Length	Bessel	Born Jordan	Choi Williams
L = 63	$\alpha = 2$	$\alpha = 1$	$\sigma = 4$
L = 127	$\alpha = 2.5$	$\alpha = 1$	$\sigma = 5$
L = 255	$\alpha = 2.5$	$\alpha = 1$	$\sigma = 6$

Table 1. Optimal scaling factors of TFD's.

4 Doppler ultrasound signal simulation

In order to characterize the pseudo instantaneous mean frequency (PIMF) and the RMS mean bandwidth (RMSMB) error estimations when the TFD are used, it has been proposed the utilization of a simulated Doppler ultrasonic quasi-stationary signal that represents a typical blood flow in the Carotid artery. Its characteristics are well documented [8], [9] and [10].

Briefly, the signal's duration is 0.7s., indeed, it is the signal's mean period; it has a constant RMSMB of 100Hz and its PIMF wave form is shown in figure 1. The simulation procedure is accurate described in [6]. In this work, a sampling rate $f_s=19200Hz$ is considered, i.e. $T=13440$ samples are taken. 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 -10 dB, -20 dB, -30 dB and -40 dB are considered (the minus signal will be omitted); also, noiseless case is treated.

5 Spectral estimation

The spectral estimation of both the RMSMB and the PIMF is worked out as in [4] and [6]. Their procedures have a common part. First, a signal piece of length L is taken from the n^{th} to the $(n+L-1)^{\text{th}}$ elements of the whole signal, it will be called the n^{th} signal window. In this work, L can be 63, 127 and 255, and $L=2N-1$. The signal window's elements are numbered in the discrete time domain from $1-N$ to $N-1$. Second, the analytic signal of this signal window is calculated. The analytic signal's elements are also numbered in the discrete time domain from $1-N$ to $N-1$. Third, the TFD of this analytic signal is calculated using equation (1), (2) or (3), depending on the study case, and considering prescribed truncation indexes TI and optimal scaling factors. The TFD's elements are numbered in the discrete frequency domain from 0 to $N-1$. Note that the components corresponding to negative frequencies, which are numbered from $N/2$ to $N-1$, all are equal to zero. Finally, the pseudo instantaneous power distribution (PIPD) of this TFD is calculated. Its elements are also numbered in the discrete frequency domain from 0 to $N-1$. Observe that the components corresponding to negative frequencies, which are numbered from $N/2$ to $N-1$, all are also equal to zero. The PIPD is defined as:

$$PIPD(0, k) = \begin{cases} TFD(0, k) & TFD(0, k) \geq 0 \\ 0 & TFD(0, k) < 0 \end{cases} \quad (4)$$

In case of the PIMF calculation, the pseudo instantaneous mean frequency associated to the n^{th} window signal is stated by:

$$PIMF(n) = \frac{\sum_{k=0}^{N/2-1} f_k \cdot PIPD(0, k)}{\sum_{k=0}^{N/2-1} PIPD(0, k)} \quad (5)$$

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 RMSMB calculation, the RMS mean bandwidth associated to the n^{th} window signal is stated by:

$$RMSMB(n) = \sqrt{\frac{\sum_{k=0}^{N/2-1} (PIMF(n) - f_k)^2 \cdot PIPD(0, k)}{\sum_{k=0}^{N/2-1} PIPD(0, k)}} \quad (6)$$

with the same considerations as in equation (5).

6 Error estimation

Typically, in any spectral estimation, the error has two independent components [6]. 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:

$$error_{RMS} = \sqrt{bias^2 + std^2} \quad (7)$$

In case of calculating the error estimation of the PIMF, it can be done with:

$$bias = \frac{1}{m} \sum_{n=0}^{m-1} (PIMF_{estimated}(n) - PIMF_{theoretic}(n)) \quad (8)$$

$$std^2 = \frac{1}{m} \sum_{n=0}^{m-1} (PIMF_{estimated}(n) - PIMF_{theoretic}(n))^2 \quad (9)$$

where m is the total amount of fully overlapped signal windows of length L in the whole signal of length T , in

consequence, $m = T-L+1$. Whereas, in case of calculating the error estimation of the RMSMB, it can be done with:

$$bias = \frac{1}{m} \sum_{n=0}^{m-1} (RMSMB_{estimated}(n) - RMSMB_{theoretic}(n)) \quad (10)$$

$$std^2 = \frac{1}{m} \sum_{n=0}^{m-1} (RMSMB_{estimated}(n) - RMSMB_{theoretic}(n))^2 \quad (11)$$

with same considerations as in equations (8) and (9).

7 Results

Figures 2, 3 and 4 show the detailed results obtained for the Bessel, Born Jordan and Choi Williams TFD, respectively. Each figure shows a set of graphs which relates the increment of jointly RMSMB and PIMF estimation error with the truncation index of the generalized time-index autocorrelation function of the considered TFD. Note that the calculations are

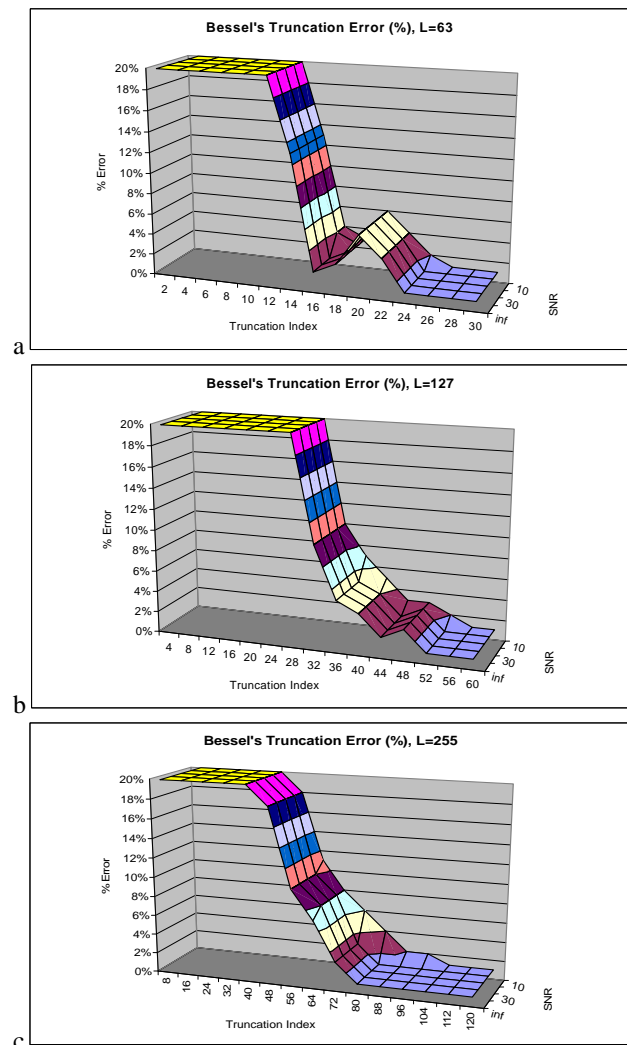


Figure 2. Bessel TFD Increment of jointly RMSMB and PIMF estimation error vs. Truncation index for SNR dynamical range (10-inf dB, 20-inf dB, 30-inf dB, 40-inf dB, inf dB), and window lengths of a) 63, b) 127, c) 255. Optimal scaling factors are considered.

made involving the TFD's optimal scaling factors. Each graph takes in account several SNR dynamical ranges (10-inf dB, 20-inf dB, 30-inf dB, 40-inf dB, inf dB), and several window lengths (63, 127, 255). The increment of estimation error is referred to that obtained when no truncation of the generalized time-index autocorrelation function is involved.

Table 2 shows the truncation index (TI) that correspond to a jointly RMSMB and PIMF spectral estimation error increment of 5%, 3% and 1% for a SNR dynamical range of 30-inf dB. Note that the admissible values of TI are from 1 to $N-2$, where the window length is $L=2N-1$.

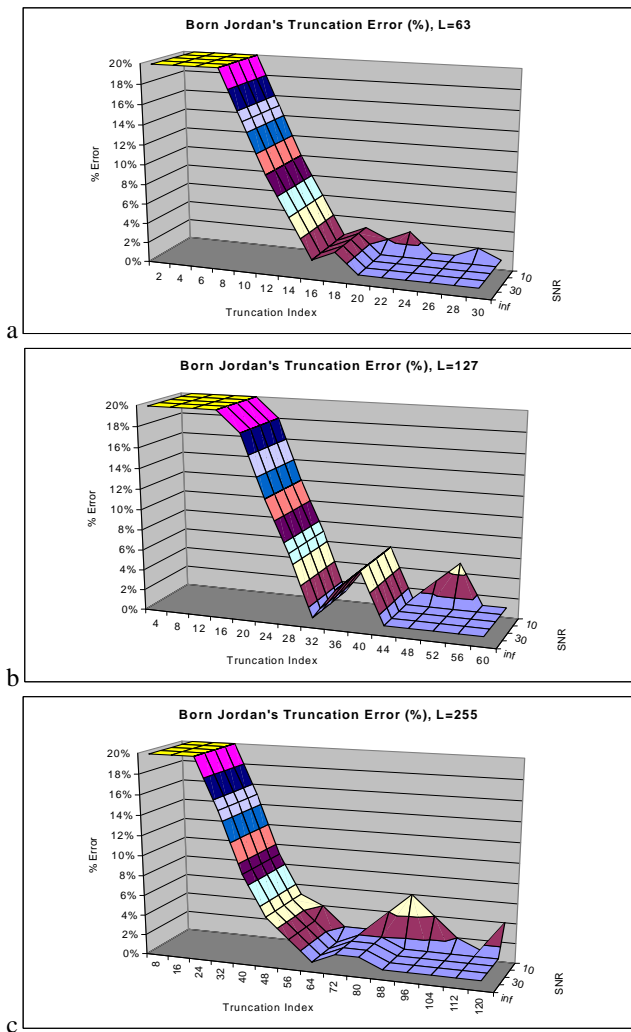


Figure 3. Born Jordan TFD Increment of jointly RMSMB and PIMF estimation error vs. Truncation index for SNR dynamical range (10-inf dB, 20-inf dB, 30-inf dB, 40-inf dB, inf dB), and window lengths of a) 63, b) 127, c) 255. Optimal scaling factors are considered.

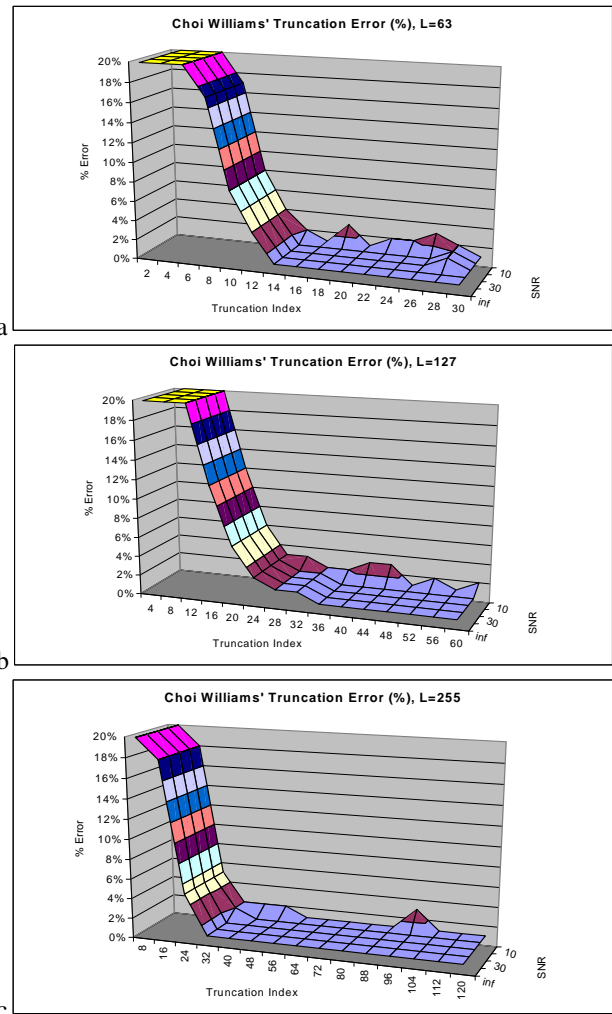


Figure 4. Choi Williams TFD Increment of jointly RMSMB and PIMF estimation error vs. Truncation index for SNR dynamical range (10-inf dB, 20-inf dB, 30-inf dB, 40-inf dB, inf dB), and window lengths of a) 63, b) 127, c) 255. Optimal scaling factors are considered.

Error	Win.length	Bessel	Born Jordan	Choi Williams
1%	L = 63	TI=26	TI=20	TI=14
1%	L = 127	TI=52	TI=44	TI=36
1%	L = 255	TI=80	TI=88	TI=40
3%	L = 63	TI=24	TI=16	TI=14
3%	L = 127	TI=44	TI=44	TI=24
3%	L = 255	TI=72	TI=56	TI=32
5%	L = 63	TI=22	TI=16	TI=12
5%	L = 127	TI=36	TI=44	TI=24
5%	L = 255	TI=72	TI=48	TI=24

Table 2. Truncation index corresponding to a SNR dynamical range of 30-inf dB. Jointly RMSMB and PIMF estimation error increment of 1%, 3% and 5%.

8 Analysis of a real Doppler ultrasound signal

This section analyses a real Doppler ultrasound signal measured in the laboratory. Again, correspond to a signal that models the carotid artery blood flow mean velocity. Figure 5 shows its PIPD using the Born Jordan distribution.

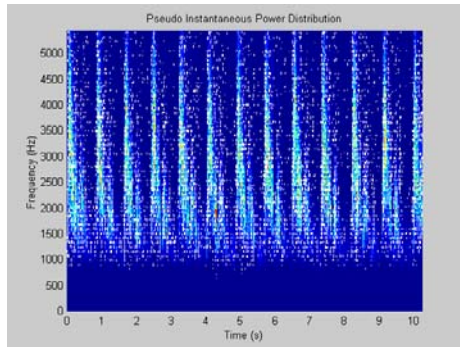


Figure 5. PIPD corresponding to a Doppler ultrasonic signal measured in laboratory (Carotid artery blood flow mean velocity). Born Jordan distribution with $\alpha = 1$, and $L = 127$ are used.

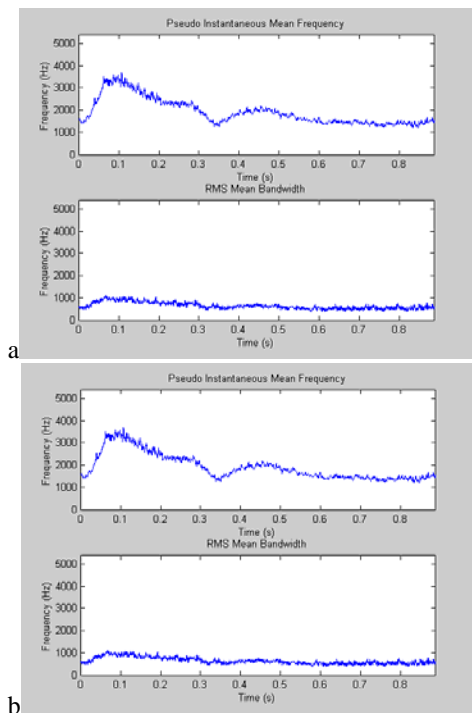


Figure 6. Averaged PIMF and RMSMB per cardiac cycle corresponding to a Doppler ultrasonic signal measured in laboratory (Carotid artery blood flow mean velocity). Born Jordan distribution with $\alpha = 1$, $L = 127$, a) no truncation and b) a truncating index $TI = 44$.

Figure 6.a shows the PIMF and the RMSMB, both averaged per cardiac cycle. An optimal scaling factor $\alpha = 1$, a

window length $L = 127$ and no truncation ($TI = 63$) are used. Finally, figure 6.b shows the PIMF and the RMSMB but using a truncating index $TI = 44$. Similar waveforms are experimentally obtained using Bessel and Choi Williams distributions. Table 3 shows the RMS deviations obtained using truncation respect to the spectral estimations done without truncation.

	L	TI	PIMF	MBRMS
Bessel	63	26	0.00	0.00
		24	0.00	0.00
	127	22	2.43	3.85
		52	0.00	0.00
	255	44	3.06	3.21
		36	6.69	5.78
Born Jordan	63	80	3.08	2.77
		72	4.09	3.65
	127	20	0.00	0.00
		16	5.69	6.46
	255	44	0.00	0.00
		88	0.00	0.00
Choi Williams	63	56	3.32	2.76
		48	4.48	3.69
	127	24	0.77	0.58
		32	1.03	0.76
	255	40	0.60	0.46
		24	1.77	1.32

Table 3. RMS deviations (Hz) using truncation respect to spectral estimations with no truncation.

9 Conclusions

A controlled truncation in the index of the generalized time-index autocorrelation function results in a controlled decrement in the accuracy of TFD calculation. Such a controlled reduction in the accuracy of TFD calculation produces a controlled increment in the spectral estimation errors but an important reduction in the amount of calculations is involved, this being the main motivation of this study.

Three simplified expressions including an autocorrelation truncating index that calculate some TFD have been considered: the Bessel (1), the Born Jordan (2) and the Choi Williams (3) distributions. The optimal parameters of time frequency distributions (TFD) have been used (table 1). The case study considered is a simulated Doppler ultrasonic quasi-stationary signal that represents a typical blood flow in the Carotid artery. Figures 2, 3 and 4 show the detailed results obtained. Those consist on the characterization of the increment PIMF and RMSBW estimation error as a function of the truncation index, the SNR

and the sample window length. Table 2 depicted the truncation index (TI) that corresponded to a jointly RMSMB and PIMF spectral estimation error increment of 5%, 3% and 1% for a SNR dynamical range of 30-inf dB, respectively. Note that the truncation of the autocorrelation function is more convenient for the Choi Williams distribution.

Finally, in section 8, a real Doppler ultrasound signal measured in laboratory is used for the TFD performance evaluation. The results corresponding to the Born Jordan distribution ($\alpha=1$, $L=127$) with and without truncation are shown in figure 6. Similar waveforms are experimentally obtained using Bessel and Choi Williams distributions. Table 3 shows the RMS deviations obtained using truncation respect to the spectral estimations done without truncation. Results are being applied to the development of a real-time spectrum analyzer for Doppler blood flow instrumentation [13].

10 Acknowledgements

The authors acknowledge project DGAPA-UNAM-PAPIIT (IN114710), project Consorciado CYTED (P506PIC0295) by the financial support. Also we want to acknowledge to M. Fuentes, J.A. Contreras, for their technical support in the development of this work.

11 References

- [1] Boashash, B. and Black, P.J. "An efficient real-time implementation of the Wigner-Ville distribution". IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-35 (11). 1611-1618, November 1987.
- [2] Cohen, L. "Time-Frequency Distributions –A Review". Proceedings of the IEEE, 77 (7), 941-981, July 1989.
- [3] Choi, H. and Williams, W.J. "Improved time-frequency representation of multicomponent signals using exponential kernels". IEEE Transactions on Acoustics, Speech and Signal Processing, 37(6), 862-871, June 1989.
- [4] Fan, L. and Evans, D.H. "Extracting instantaneous mean frequency information from Doppler signals using the Wigner distribution function". Ultrasound in Med. & Biol., 20(5), 429-443, May 1994.
- [5] Guo, Z., Durand, L.G. and Lee, H.C. "The time-frequency distributions of nonstationary signals based on a Bessel kernel". IEEE Transactions on Signal Processing, 42(7). 1700-1707, July 1994.
- [6] Cardoso, J. G. Ruano and P. Fish. "Nonstationary Broadening Reduction in Pulsed Doppler Spectrum Measurements Using Time-Frequency Estimators". IEEE Transactions on Biomedical Engineering, 43(12), 176-1186, December 1996.
- [7] García-Nocetti F., Solano J. and Rubio E. "Precision enhancement of Doppler ultrasound spectral estimation by finding TFD optimal parameters. Forum Acusticum Sevilla. Special Issue of the Revista de Acústica, 33, Sevilla, Spain, September 2002.
- [8] García-Nocetti F., Solano J., Rubio E. and Moreno, E. "High Performance Computing of Time Frequency Distributions for Doppler Ultrasound Signal Analysis". Proceedings of the 15th IFAC World Congress, T-Fr-A17-5, Barcelona, Spain, July 2002.
- [9] F. García, E. Moreno, J. Solano, M. Barragán, Sotomayor, M. Fuentes and P. Acevedo, "Design of a continuous wave blood flow bi-directional Doppler system". Ultrasonics Journal, 44, e307-e312, December 2006.
- [10] J. Solano, M. Vazquez, E. Rubio, I. Sanchez, M. Fuentes and F. García-Nocetti. "Doppler ultrasound signal spectral response in the measurement of blood flow turbulence caused by stenosis", Physics Procedia, 3(1), 605-613, January 2010.
- [11] García-Nocetti, D. F., Solano-González, J., Rubio-Acosta, E. and Moreno-Hernández, E. "Towards the Simplified Computation of Time-Frequency Distributions for Signal Analysis", Proc. of the 6th IFAC Workshop on Algorithms and Architectures for Real-Time Control AARTC'2000, 161-166, Palma de Mayorca, España, Mayo 2000.
- [12] García-Nocetti, D.F., Solano-González, J., Rubio-Acosta, E. and Moreno-Hernández, E. "Fast Computation of Time-Frequency Distributions Using a Parallel DSP-based System for Signal Analysis", IFAC Conference on New Technologies for Computer Control 2001 NTCC-2001, 187-192, Hong Kong, China, Noviembre 2001
- [13] J. Solano, M. Fuentes, A. Villar, J. Prohias, and F. García-Nocetti. "Doppler Ultrasound Blood Flow Measurement System for Assessing Coronary Revascularization". Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology (BIOCAMP'11), WORLDCOMP'11, Julio 18-21, 2011. Las Vegas Nevada, USA. 429-433. CSREA Press 2011. ISBN: 1-60132-169-4.

REAL TIME SURFACE TRACKING METHOD FOR IMAGE ENHANCEMENT OF OPTICAL COHERENCE TOMOGRAPHY

C. G. Song¹, D. H. Shin¹, Y. K. Oh¹, J. Kang²

¹ Div. of Electronic Eng., Chonbuk Natl. Univ., Korea

² Dept. of Electrical & Computer Eng., Johns Hopkins Univ., U.S.A

ABSTRACT

We developed a surface topological compensation algorithm to extend imaging range in common-path Fourier-domain optical coherence tomography configuration. A surface detection algorithm based on a Savitzky-Golay filter of A-scan data and thresholding was applied to real-time depth tracking. The algorithm output controlled a motorized stage to adjust the probe position according to the sample's topological variance in real-time. OCT images obtained using our algorithm showed a significantly extended imaging range, consequently, the devised algorithm demonstrated potential for improving endoscopic OCT.

Index Terms— OCT, surface tracking, topological variance, motorized stage, Savitzky-Golay filter

1. INTRODUCTION

Optical coherence tomography (OCT), which is one of the various optical imaging modalities, is a novel imaging technology that provides high-resolution, subsurface depth profiling, and cross-sectional imaging in vivo with relatively simple optical arrangements and an inexpensive light source in a non-invasive manner. The concept of OCT and its application were first introduced by Fujimoto et al. in 1991 [1]. It has several benefits for the non-invasive, high-resolution and fast-acquisition tomography of the internal microstructure in biological systems and materials. First of all, it can provide much higher-resolution images (2-10 μm) than conventional imaging techniques, such as ultrasound (over 500 μm), MRI and CT (over 100 μm), although its depth information is limited to a range of approximately 2-3 mm in turbid tissue [2]. Also, OCT has a faster scanning speed for acquisition and relatively wider dynamic range [3]. Moreover, the entire system is simple and portable and, thus, has the potential to enable OCT catheters to be incorporated into endoscopic instruments or bedside devices. Finally, since OCT is based on optics, it can be combined with other spectroscopic techniques to assess the optical and biochemical aspects of the tissue being imaged.

Common-path OCT (CP-OCT) was proposed by Vakhtin et al. in 2003 [4]. In the CP-OCT configuration, the beam paths, which the sample signals backscattered from the sample and reference signals reflected from the reference plane follow, are commonly shared, thereby eliminating the need for the reference arm in the interferometer. This modality can minimize the effect caused by the mismatch of the polarization and dispersion states between the optical elements in the interferometer and the sensitivity to vibration, and enhance the scanning speed, simplicity and system robustness. Consequently, this configuration has the potential to be used as a microsurgical tool. Some researchers have reported the feasibility of an endoscopic CP-OCT implementation based on the common-path modality [5].

However, unfortunately, most OCT systems generally suffer from a limited imaging depth range of only 1-3 mm, depending on the tissue type and, thus, this limitation restricts their clinical applications when the sample's topological variance is larger than the imaging depth range [6]. To overcome these limitations, some techniques such as the adaptive ranging technique based on depth tracking have been proposed in previous papers [7]. In these methods, the coherence gate offset and range on the reference arm are adaptively adjusted by means of an active tracker consisting of various optical lenses and a galvanometer. However, these techniques require the supplementary alignment of the various optical lenses or components and synchronization control and, thus, the composition and control procedure of the OCT system might become more onerous and complicated. Also, they compensate for the topological variance and motion by adjusting the optical pathlength on the reference arm and, therefore, this strategy might be inappropriate for the CP-FD-OCT system constructed in this study, since CP-OCT uses the common beam path of the sample and reference signal instead of using the reference mirror used in the conventional OCT composition. Recently, Zhang et al. [8] reported a CP-FD-OCT system providing a surface topology and motion compensation technique in the axial direction by means of a 1-D erosion-based edge-searching algorithm, which makes use of the relatively

simple signal processing of the A-scan data instead of the alignment of complex optical components.

To assess the feasibility of the system described in this paper, an active compensation algorithm of the topological variance by means of a sample surface detection algorithm using a Savitzky-Golay smoothing filter and feedback control for adjusting continuously the position of the motorized stage was developed in the present study. This algorithm makes it possible to image a deeper range along the z-axis by keeping the distance between the end of the probe and the sample's surface constant, as compared to the conventional scanning strategies.

2. ACTIVE TRACKING WITH COMPENSATION ALGORITHM

Figure 1 shows a flow chart of the active topological variance compensation algorithm during B-mode scanning in CP-FD-OCT, while the distance from the sample's surface exceeds the OCT imaging depth range or when the probe is too close to the sample.

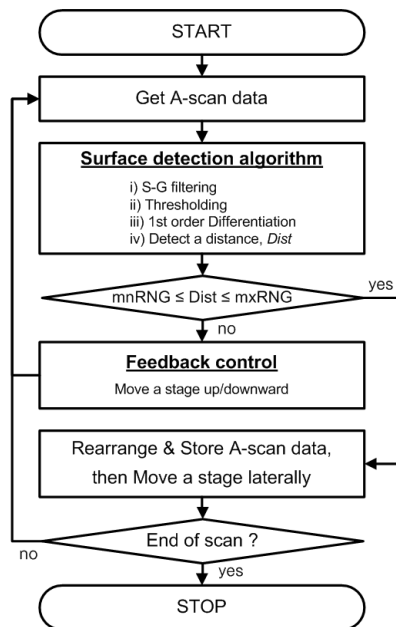


Fig. 1. Flow chart for active surface tracking algorithm.

In 'Step-1', the A-scan data, $a(z)$ is obtained from the probe (N is the total length of $a(z)$, as shown in Figure 2(a).

In 'Step-2', the distance ($Dist$) between the end of the probe and the sample's surface is determined, as follows; i) $a(z)$ is smoothed by a 3rd-order Savitzky-Golay filter (its window length is 9), as shown in Figure 2(b). The main

advantage of the Savitzky-Golay filter used in this algorithm is that it can preserve the unique features of the distribution, such as the relative maxima, minima and width, which are usually flattened by other adjacent averaging techniques, such as a moving average or low-pass filter, as well as effectively reducing the unnecessary speckle noise [9]. This attribute is quite useful for the more accurate detection of the edges from the A-scan data and, thus, over- or under-estimation of the distance can be effectively diminished compared to the other smoothing methods. ii) the smoothed A-scan data, $a_{sm}(z)$, is processed using a certain threshold level ($thre$) to avoid the noise effect, as follows (Figure 2(c));

$$a_{thre} = \begin{cases} a_{sm}(z), & a_{sm}(z) > thre \\ thre, & others \end{cases} \quad (1)$$

iii) $Dist$ is given by the first increment point of the differential of the post-thresholding data, as shown in Figure 2(d).

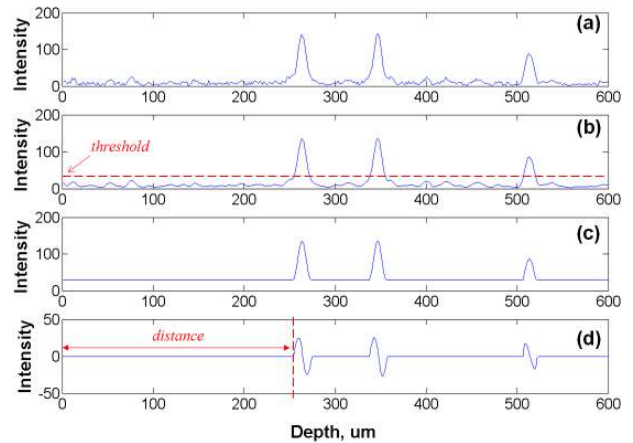


Fig. 2. Active surface tracking algorithm. (a) Raw A-scan data, (b) A-scan data after Savitzky-Golay smoothing filter, (c) Thresholding of A-scan, and (d) First increment point detection for edge location.

In 'Step-3', the discrepancy ($Diff$) between the preset ($setDist$) and measured ($Dist$) distances is calculated, as follows;

$$Diff = Dist - setDist \quad (2)$$

In 'Step-4', by using the $Diff$ value obtained in 'Step-3', the control system sends the feedback control signal to the motorized stage. If the absolute value of $Diff$ is outside of the preset acceptable range ($AcptRng$), the stage is moved either upward for a positive value of $Diff$ or downward for a negative value of $Diff$. Subsequently, 'Step-1' is performed again until $Diff$ is within $AcptRng$. On the other hand, if it is within $AcptRng$, the measured $a(z)$ is rearranged and stored in memory. During recording, the values of $a(z)$ are

repeatedly obtained while maintaining a constant distance between the end of the probe and the sample's surface and, thus, this $a(z)$ can be rearranged by considering the practically moved height of the stage. The variable, dZ , is used for counting the relative displacement at the current position compared to that at the start of the B-scan ($dZ=0$) on the z-axis. For example, a positive value of dZ indicates that the probe has moved closer to the sample, so it implies that the practical depth of the OCT image might be relatively larger than that of the measured $a(z)$, whereas a negative dZ means that the practical depth of the OCT image is relatively smaller than that of the measured $a(z)$.

In 'Step-5', the stage is moved laterally for one step, and 'Step-1' is performed again until the moved position of the stage is the end of the scan on the x-axis.

3. EXPERIMENTAL SETUP

To obtain high-resolution OCT images with an extended range of imaging depths, a motorized-stage-based OCT system was developed. It consists of a high-resolution spectrometer, actively controllable motorized-stage, actuators and control modules, as well as basic compositions such as a light source, 50/50 coupler, and single mode fiber-optic probe. Figure 3 shows the block diagram of the developed CP-FD-OCT system.

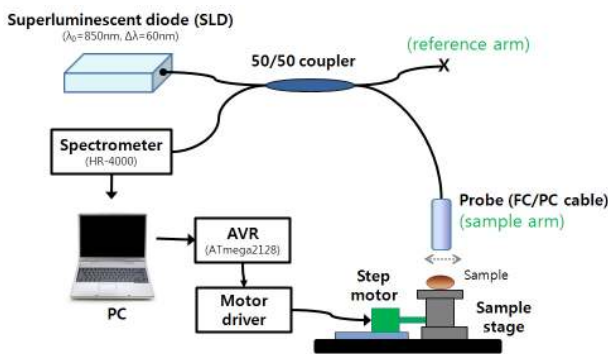


Fig. 3. Block diagram of the developed CP-FD-OCT system.

A superluminescent diode (SLD) (SLD-351, Superlum Diode Ltd., Ireland) with a central wavelength of 860 nm and spectral full-width at half maximum (FWHM) of ~60 nm was used as the light source. A 50/50 coupler (FC850-40-50-APC, Thorlabs Inc., U.S.) was used as the beam splitter, and only one branch on the right side was used as the common path for the signal and reference. The single mode fiber-optic probe constructed in this study was fixed on a standing vise, with A-scan (z-axis) and B-scan (x-axis). The two axes of the x and z directions were driven by a

motorized stage (M-561D-XYZ, Newport Corp., U.S.) with two separate step motors (SE-SM243, N.T.C., Korea) installed on its lateral side. The reference signal came from the Fresnel reflection at the fiber probe end and the sample signal and the reference were received by a high-speed spectrometer (HR-4000, Ocean Optics, U.S.) with a charge-coupled device detector array with 3648 pixels covering a range of 700-900 nm. This system make it possible to extend the imaging range, since the position of the probe can be adjusted actively and simultaneously according to the sample's topological variance, whereas the time needed for image acquisition is relatively longer.

4. RESULTS

The performance of the active topology compensation algorithm was tested under static conditions using an onion sample with several layers of highly curved surfaces. At first, a B-scan 2-D OCT image was obtained by the conventional fixed-stage method, as shown in Figure 4(a). The 860 nm CP-FD-OCT provided effective imaging in the range below 500 nm and the structure of some of the layers was very clear within this range. However, the OCT image fades away as the probe is moved further away from the sample's surface, due to the limited depth range.

Figure 4(b) shows an improved OCT image obtained using the active topological variance compensation algorithm. By using our algorithm, the probe could actively track the sample surface variance and, consequently, the effective imaging depth was extended to the probe's free-moving range. Also, the sub-layers of the sample could be monitored more clearly, even if the distance between the probe and sample's surface was outside of the limited imaging range.

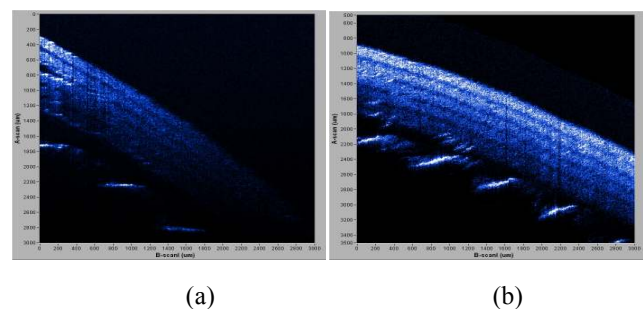


Fig. 4. Image of an onion sample obtained by (a) the conventional static stage on the z-axis with limited imaging depth and (b) active topological variance compensation algorithm with extended imaging depth.

5. CONCLUSION

We devised an active surface tracking algorithm to extend the image range of OCT scanning for CP-FD-OCT configuration. Consequently, the OCT images obtained using the motorized-stage-based system showed a significantly extended imaging range through real-time accurate depth tracking. These results demonstrate that our OCT system and algorithms have good potential to resolve several of the limitations of conventional OCT systems.

6. ACKNOWLEDGMENT

This work was supported by the Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning(KETEP) grant funded by the Korea government Ministry of Knowledge Economy(No.20104010100660) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MEST 2011-0030781, 2010-0021864)

7. REFERENCES

- [1] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, J. G. Fujimoto, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178-1181, 1991.
- [2] S. A. Boppart, B. E. Bouma, C. Pitris, J. F. Southern, M. E. Brezinski, J. G. Fujimoto, "In vivo cellular optical coherence tomography imaging," *Nat. Med.*, vol. 4, no. 7, pp. 861-865, 1998.
- [3] G. J. Tearney, M. E. Brezinski, B. E. Bouma, S. A. Boppart, C. Pitris, J. F. Southern, J. G. Fujimoto, "In vivo endoscopic optical biopsy with optical coherence tomography," *Science*, vol. 276, no. 5321, pp. 2037-2039, 1997.
- [4] A. B. Vakhnin, D. J. Kane, W. R. Wood, K. A. Peterson, "Common-path interferometer for frequency-domain optical coherence tomography," *Appl. Optics*, vol. 42, no. 34, pp. 6953-6958, 2003.
- [5] U. Sharma, N. M. Fried, J. Kang, "All-fiber common-path optical coherence tomography: Sensitive optimization and system analysis," *IEEE T. Sel. Top. Quant.*, vol. 11, no. 4, pp. 799-805, 2005.
- [6] A. Low, G. Tearney, B. Bouma, I. Jang, "Technology insight: Optical coherence tomography - Current status and future development," *Nat. Clin. Pract. Card.*, vol. 3, no. 3, pp. 154-162, 2006.
- [7] N. Iftimia, B. Bouma, J. F. de Boer, B. Park, B. Cense, G. Tearney, "Adaptive ranging for optical coherence tomography," *Opt. Express*, vol. 12, no. 17, pp. 4025-4034, 2004.
- [8] K. Zhang, W. Wang, J. Han, J. U. Kang, "A surface topology and motion compensation system for microsurgery guidance and

intervention based on common-path optical coherence tomography," *IEEE T. Biomed. Eng.*, vol. 56, no. 9, pp. 2318-2321, 2009.

[9] J. Luo, K. Ying, J. Bai, "Savitzky-Golay smoothing and differentiation filter for even number data," *Signal Process.*, vol. 85, no. 7, pp. 1429-1434, 2005.

Software Development Kit to Verify Quality Iris Images

Isaac Mateos, Gualberto Aguilar, Gina Gallegos

Sección de Estudios de Posgrado e Investigación Culhuacan, Instituto Politécnico Nacional,
México D.F., México

Abstract - *This paper proposes an algorithm for the development of an SDK to verify quality on iris images based on ISO/IEC 19794-6:2005, standard that is being mainly used by several manufacturers of biometric systems based on iris recognition. For the development of this algorithm an assessment is made for each parameter recommended in the standard, with the aim of determine a total quality score of iris images and decide if they have unacceptable, low, medium or high quality, selecting from this way the good ones and thus increase the efficiency of iris recognition systems. The proposed algorithm has been tested with images from a own dataset collection. It is still a need to adopt a method to determine an overall value to the fusion of all the individual feature values.*

Keywords: Biometrics, iris image quality evaluation.

1 Introduction

Currently, iris recognition has become the most reliable biometric system performance in terms of verification and identification of people. Today there are several systems dedicated to iris recognition, however the performance of these systems is affected due to poor quality iris images. The main problems that affect the system are the false accept (FMR) and the false reject (FNMR). If one can detect low-quality biometric samples, the information can be used to initiate the acquisition of new data and improve system performance. For best performance of the development of the SDK, the proposed algorithm refers to the ISO 19794-6 standard, which makes recommendations about the features which must be met by the iris biometric images to determine if you have a suitable quality for specific purposes.

The main goal of standardization is to enable harmonized interpretation of quality scores and can differentiate them from the different vendors, algorithms and versions, enabling in this manner a competitive multi-vendor marketplace. As result of the measurement, the same quality measure can be used to selectively improve an operational biometric database by replacing low-quality biometric samples with high-quality samples of the same biometric.

1.1 ISO/IEC 19794-6:2005

The International Organization for Standardization (ISO) has created the ISO/IEC 19794-6:2005. [1] in support to the necessity of iris images quality samples, which recommends the assessment of essential characteristics of iris images, giving an overall value between 0 and 100, with 100 being the highest quality and 0 de poorest quality . However, this issue is not specifically defined and is still ongoing research.

1.2 IREX -IQCE

The Iris Exchange (IREX) [2] was initiated at National Institute of Standards and Technology (NIST) in support of an expanded marketplace of iris-based applications based on standardized interoperable iris imagery, mainly in support of the ISO/IEC 19794-6 standard. Iris Quality Calibration and Evaluation (IQCE) [3] aims to evaluate the effectiveness of image quality assessment algorithms (IQAAs) that produce a scalar overall image quality in predicting the recognition accuracy of particular comparison algorithms (from the supplier of the IQAA), and of other algorithms.

1.3 Current SDK's

Some of the leading vendors in the iris biometric recognition system in the marketplace with their own SDK are LG, NEUROTECHNOLOGY, CROSSMATCH, AWARE, MORPHO, IRITECH, IRISID, KYNEN, L1, and exists different performance between their SDK results, of course they utilize their own algorithms for measurement of the features of the images, and compare some of these SDK with the current development one, compare some of these SDK's may give us information about which one has better performance and create a competitive environment for best results.

2 Proposed algorithm

For the development of the SDK, different algorithms must be adopted for measuring the individual characteristics that indicates the standard, and also for the segmentation of the iris in the images. There are several methods to achieve this purpose however is a challenge to select the appropriate one that complies with speed and accuracy for each feature. In this work we started with the identification and location of

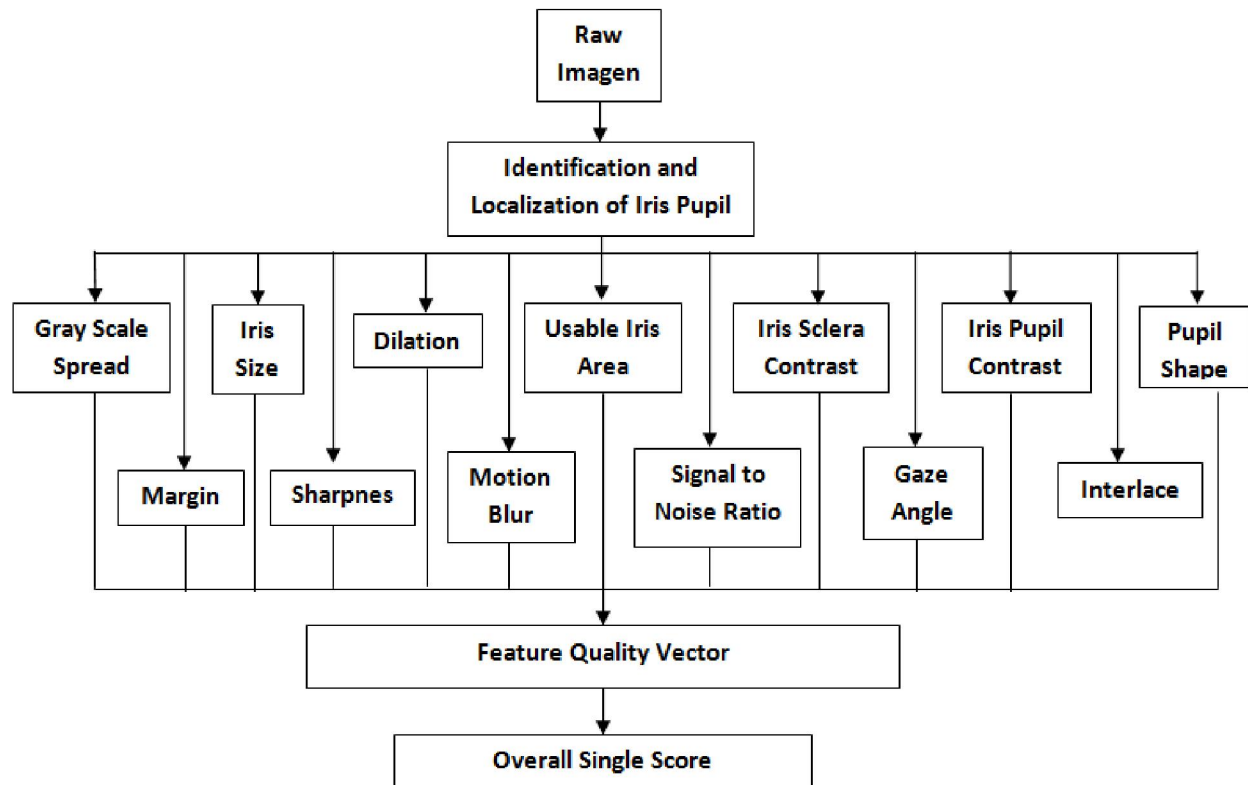


Figure 1. Proposed Algorithm to the SDK

the pupil, because if you find an image that of is not an iris, the SDK automatically rule it out, thus avoiding the image processing passes through the measurement of each feature. The image processing is described in the figure 1.

As shown in the algorithm there are many individual characteristics to measure, so it is necessary to give a brief description of the most important ones according to IQCE [2] as following:

Usable iris area is defined as the percentage of iris that is not occluded by eyelash, eyelid, specular reflections and ambient specular reflections.

Iris pupil contrast is a measure of the image characteristics at the boundary between the iris region and the pupil.

Pupil shape is a measure of regularity in pupil-iris boundary.

Iris sclera contrast is a measure of the image characteristics at the boundary between the iris region and the sclera.

Gaze angle is the deviation of the optical axis of the subject's iris from the optical axis of the camera.

Sharpness, defined as the absence of defocus blur, can result from many sources, but in general, defocus occurs when the object is outside the depth of field of the camera.

Dilation is defined as the ratio of the pupil radius to iris radius.

An image with a high *Gray scale spread* (good quality) is a properly exposed image, with a wide, well distributed spread of intensity values.

Iris shape is defined as the shape of iris-sclera boundary.

Iris size is defined as the number of pixels across the iris radius, when the iris boundary is modeled by a circle.

Motion blur is defined as the blur cause by motion of the camera or the iris, or both.

Once performed the measurement of the characteristics indicated in the algorithm, the system must create a feature vector containing the measurements of each characteristic separately and finally an overall score value should be given according to the ISO/IEC 19794-6:2005 standard [1] to determine if the image has a poor, low, medium or high quality.

3 Experimental results

Tests have been performed with 60 biometric iris images from an own dataset collection, all images were acquired using an LG IRIS ID iCAM TD100 [11]. The iris images are 480x640 in resolution.

3.1 Pupil Identification and Localization

First to locate the iris pupil, is used the method described by Lili Pan [4], using a binarization by selecting an appropriate threshold and finding the center where the true value of pixel intensity is minimal. As shown in Figure 2.

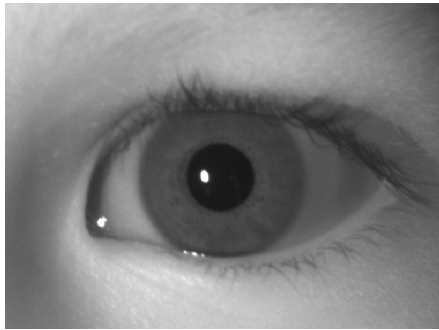


Figure 2. Binarized image to pupil detection

3.2 Gray Scale Spread Measurement

Once identified that it is a true image of the iris, the measurement of the first feature 'Gray Scale Spread'. An image with a high GRAY SCALE SPREAD (good quality) is a properly exposed image, with a wide, well distributed spread of intensity values [3]. This is accomplished by performing a histogram as shown in Figure 3. Then the image is cropped for fast image segmentation of the pupil and iris as shown in Figure 4.

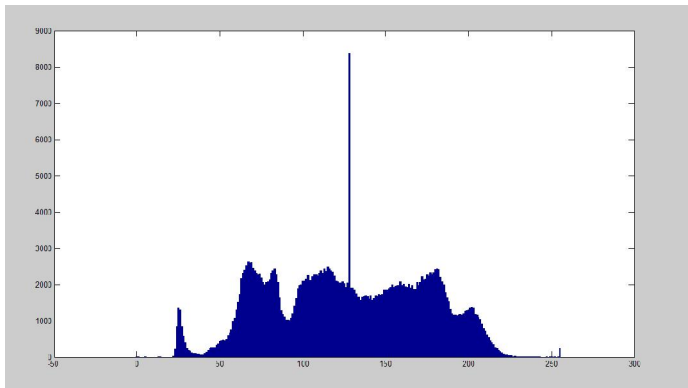


Figure 3. Histogram wide spread image values

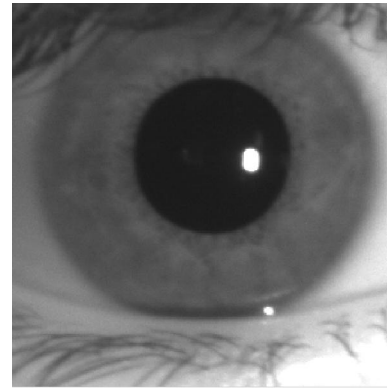


Figure 4. Cropped image

3.3 Contrast Measurement

The next step is to measure the contrast level of pupil and sclera, performing the measurement along the diameter of the iris, obtaining a graph as shown in Figure 5.

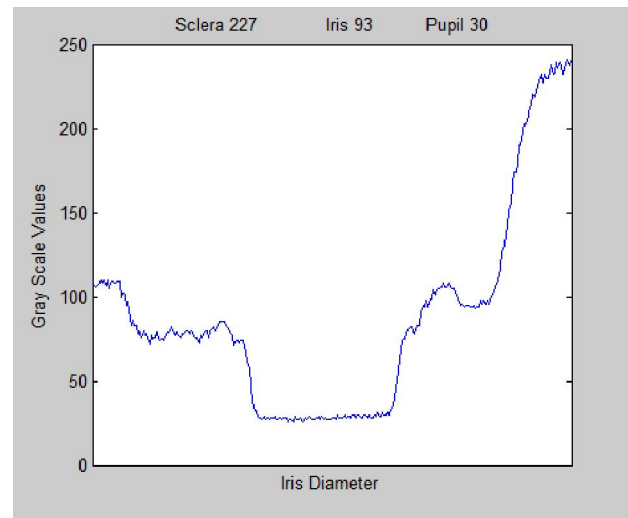


Figure 5. Contrast in Sclera, Iris and Pupil

3.4 Pupil and iris Shape

Continuing using the algorithm of Lili Pan [4], noting the graph of Figure 5 shows that the intensity values between pupil, iris and sclera vary drastically, based on this fact can be detected edges, result can be seen in Figure 6.

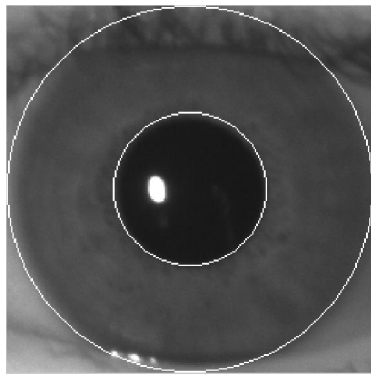


Figure 6. Iris and Pupil Segmentation

3.5 Iris Size and Dilation Measurement

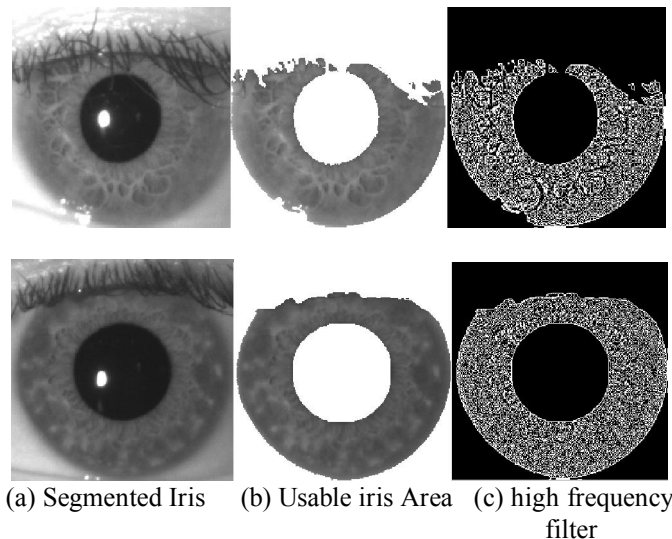
IRIS SIZE is defined as the number of pixels across the iris radius and DILATION as the ratio of the pupil radius to iris radius [3]. As a result of segmentation by edge detection, we can easily measure the diameter of the pupil and iris, and thus obtain the value of the dilatation.

3.6 Usable Iris Area Measurement

USABLE IRIS AREA is defined as the percentage of iris that is not occluded by eyelash, eyelid, specular reflections and ambient specular reflections [3]. Occluded images are another big problem in iris image quality assessment and possibly one of the most difficult features to measure, because it is not possible to adopt a special algorithm. Lili Pan recommends compute the number of pixels in the iris area, then set two gray level thresholds one for detect eyelash and other for detect eyelid [4]. Which throws high error, because in many images the iris has both dark and light parts that may be confused with eyelashes and eyelids. Chunlei Shi recommends selecting the upper rectangle area of the pupil as the ROI, we regard the average gray value of it as the judgment criterion, and then get rid of the occlusion images [6]. By a combination of these methods, we obtain the results observed in Figure 7 (b).

3.7 Sharpness Measurement

SHARPNESS, defined as the absence of defocus blur, mainly affects FNMR and FMR, images with low sharpness inflate FMR. [3]. For the measurement of this feature, using the high frequency power of the image to evaluate the degree of focus is a common method in previous research on image focus assessment [5,6,7,8]. Measuring with a high frequency filter on the ROI (Usable iris Area) as shown in Figure 7(c).



(a) Segmented Iris (b) Usable iris Area (c) high frequency filter

Figure 7. Measurement of Visible Area and Sharpness

4 Future work

It is necessary to adopt algorithms to assess missing features as Motion Blur, Signal to Noise Ratio, Gaze Angle, Interlacing. Papers [5,7,10] talk about different algorithms for the measurement of these features so they should be tested to check which have the best performance. Once all features have been measured, we obtain a Feature Quality Vector, as shown in Figure 1, and the final step is to obtain an Overall Single Score. Not all features have the same weight of importance in the iris recognition system, thus giving an overall score is a problem, a bad quality in a feature with little weight of importance should not greatly affect the overall quality score. The papers [7,9,10] propose algorithms to obtain a score fusion, also this work intends to use a neuronal network, so continuous research and testing should be done to choose the best algorithm and finally succeed in developing a robust SDK with the best performance. Future tests will be made on dataset collection of standard biometric iris images.

5 Conclusions

Iris Usable Area is the feature most important to weight the iris recognition system and is therefore a key factor for the overall score, this because although the image count with good lighting and good sharpness if the iris is obstructed by eyelashes or eyelids, the necessary features for the recognition system never would be obtained for a good performance. so in this way, give priority in importance to all the features measured, where the second most weight of importance is the Contrast, followed by Sharpness, after this the Dilation of the pupil, the Gaze Angle, Interlace, Gray Scale, Spread, Iris Size and finally the, Motion Blur and Signal to Noise Ratio.

6 References

- [1] INCITS/ISO/IEC 19794-6:2005 — Information technology — Biometric data interchange formats — Part 6: Iris image.
- [2] National Institute of Standards and Technology NIST., IREX Iris Exchange., <http://www.nist.gov/itl/iad/ig/irex.cfm>
- [3] Elham Tabassi, Patrick Grother, Wayne Salamon., IREX II – IQCE Iris Quality Calibration and Evaluation 2011., Concept, Evaluation Plan and API 6 Version 4.1 Image Group, Information Access Division, Information Technology Laboratory, National Institute of Standards and Technology September 28, 2011.
- [4] Lili Pan, Mei Xie., Research on Iris Image Preprocessing Algorithm., School of Electronic Engineering University of Electronic Science and Technology of China, Chengdu, China, Proceedings of ISCIT2005.
- [5] Z. Wei, T. Tan, Z. Sun and J. Cui., Robust and fast assessment of iris image quality., Lecture Notes in Computer Science , v 3832 LNCS, p 464-471, 2006, Advances in Biometrics - International Conference, ICB 2006, Proceedings.
- [6] C. Shi and L. Jin., A fast and efficient multiple step algorithm of iris image quality assessment., Proceedings of the 2010 2nd International Conference on Future Computer and Communication, ICFCC 2010, v 2, p V2589-V2593, 2010, Proceedings of the 2010 2nd International Conference on Future Computer and Communication, ICFCC 2010.
- [7] N. Kalka, J. Zuo, N. Schmid and B. Cukic., Image quality assessment for iris biometric., Proceedings of SPIE - The International Society for Optical Engineering, v 6202, 2006, Biometric Technology for Human Identification III.

SESSION

**COMPUTATIONAL METHODS FOR DRUG
TARGETS AND SCREENING, AND
PHARMACOINFORMATICS**

Chair(s)

TBA

Quantitative Analyses of Kinase Inhibitor Selectivity Using Very Small Size Panels

Quoc-Nam Tran, PhD.[†]
Lamar University, USA

Abstract—Kinases are known to regulate the majority of cellular pathways, especially those involved in signal transduction. By modification of substrate activity, protein kinases also control many other cellular processes, including metabolism, transcription, cell cycle progression, cytoskeletal rearrangement, and cell movement, apoptosis, and differentiation. Because protein kinases have profound effects on a cell, and because of their central role in cellular processes, the number of kinases with potential as drug targets is significant. Furthermore, because kinases share common evolutionary backgrounds, they also share structural attributes, making it difficult for drugs to tell apart paralogs of clinical importance from off-target kinases. Thus, multi-target kinase inhibitors (KIs) tend to have undesired cross-reactivities with potentially lethal or debilitating side effects.

In this paper, we present methods for analyses of kinase inhibitor specificity and promiscuity that provide affirmative answers for the following two major questions: (a) What is the smallest subset of kinases that any compound needs to be screened against to obtain an accurate indication of specificity or promiscuity? (b) How do we find this small inferential set if it exists.

I. INTRODUCTION

Protein kinases are enzymes that modify other proteins by chemically adding phosphate groups to them. This process, called phosphorylation, usually results in a functional change of the target protein (also known as the substrate) by changing enzyme activity, cellular location, or association characteristics with other proteins. Kinases are known to regulate the majority of cellular pathways, especially those involved in signal transduction. By modification of substrate activity, protein kinases also control many other cellular processes, including metabolism, transcription, cell cycle progression, cytoskeletal rearrangement, and cell movement, apoptosis, and differentiation. Protein phosphorylation

also plays a critical role in intercellular communication during development, in physiological responses, and in homeostasis and in the functioning of the nervous and immune systems. Because protein kinases have profound effects on a cell, and because of their central role in cellular processes, the number of kinases with potential as drug targets is significant. Kinases have been implicated as drug targets not only in the treatment of cancer, but also in a number of non-oncology indications, including central nervous system disorders, autoimmune disease, post-transplant immunosuppression, osteoporosis, and metabolic disorders. Kinase inhibitors are molecules that bind to enzymes and decrease their activity. Since blocking an enzyme's activity can kill a pathogen or correct a metabolic imbalance, many drugs are kinase inhibitors.

To fully explore and exploit this opportunity of targeting kinases as drug targets, potent and selective inhibitors are required for a multitude of kinases, both as tool compounds for target validation and as leads for drug development. Kinase-inhibitor discovery has been a mainly linear process that addresses one kinase at a time and requires significant investment of time and resources for each target. In this resource-intensive and time-consuming process, an inhibitor is screened against a large size panel of kinases, typically 500 or more, to identify hits that often have weak or modest potency. Kinase selectivity is typically assessed on only a subset of the screening hits, and is monitored only at the end of the process. This strategy has significant drawbacks. First, targets are addressed one at a time, and the entire process has to be repeated for each new target of interest. Second, decisions about which targets to pursue are based on biology alone, with minimal knowledge about the availability or quality of hits against the designated target in the available chemical library.

As the heterogeneous nature of cancer is delineated, the focus of molecular therapy is shifting progressively

[†]E-mail: qntran@lamar.edu

towards multi-target drugs [1], [2], [3]. In the treatment of tumors, scientists are advocating molecular therapies based on a multi-pronged attacks [4], [5], [6], [7], [8]. For example, drug-based interference with several signaling pathways provides a multi-pronged attack that is proving more effective than single-pronged “magic bullet” attacks in hampering development and progression of malignancy. Such therapeutic agents typically target the kinases, thus blocking or interfering with signaling pathways that control cell fate and proliferation.

Small molecule kinase inhibitors are a new class of drug with a tendency to inhibit multiple targets. This new class of drug will grow remarkably as the large number of compounds currently in preclinical and clinical development progress towards the market. However, because kinases share common evolutionary backgrounds, they also share structural attributes, making it difficult for drugs to tell apart paralogs of clinical importance from off-target kinases. Thus, multi-target kinase inhibitors tend to have undesired cross-reactivities with potentially lethal or debilitating side effects. The issue of multi-target therapy has led to the requirement of analyzing the promiscuity or specificity of kinase inhibitors. A pressing issue exists of which type of clinical impact can only be achieved with a promiscuous drug, and conversely, which clinical effect lends itself to drug specificity.

Of central clinical importance in this regard is the issue of whether the desired clinical impact is likely to promote side effects or may be achieved by drugs endowed with high specificity. Currently, compounds must be screened against a large number of kinases in order to obtain an accurate indication of specificity and promiscuity. As this is a time consuming process, methods for determining specificity and promiscuity by screening compounds against fewer kinases are highly desirable.

In the subsequent sections, we will first define a new selectivity scale which is finer than the known scales in the literature by using three different binding affinity thresholds. We then use machine learning techniques [9], [10], [11] to classify the inhibitors based on their selectivity. Third, we use statistical analysis to reduce the set of kinases from 317 to 85 without losing drug screening information. Finally, we use data mining techniques to select very small subsets of kinases that provides the most accurate predictive model for drug inhibitors.

II. METHODS

A. Weighted Selectivity Scores

A first step in determining a kinase subset involves the use of data from competition binding assays where kinase inhibitors are evaluated against a panel of protein kinases. For each interaction, a quantitative dissociation constant (K_d) is needed. We develop a quantitative description, called weighted selectivity score, for assessing kinome-wide compound reactivities by suitably coarse-graining the binding-affinity vector. The reasons for introducing the weighted selectivity scores are two-fold: (a) even though ligand-kinase interaction maps provide a useful graphic overview of how compounds interact with the kinome, these maps provide only a qualitative overall measure of selectivity; (b) selectivity scores calculated by counting the number of binding interactions with less than a threshold constant such as $3\mu\text{M}$ divided by the number of kinases tested [12] do not represent all the information at our disposal about the strength of the binding interactions. In our approach, we adopt the three main binding thresholds of 100nM, $1\mu\text{M}$, and $3\mu\text{M}$. Kinases found to bind with a dissociation constant less than 100nM will be given an individual weight of 1.0. Similarly, kinases found to bind with a dissociation constant greater than 100nM but less than $1\mu\text{M}$, and kinases found to bind with a dissociation constant greater than $1\mu\text{M}$ but less than $3\mu\text{M}$ will be given an individual weight of 0.75 and 0.6, respectively. While inferential accuracy may be maintained within some latitude in the selection of individual weights, our choices responded to the need to maintain the coherence of multiple bindings vis a vis the compound score. For instance, two bindings with dissociation constant less than 100nM should yield a higher compound score than two bindings with dissociation constant less than $1\mu\text{M}$ and three bindings with dissociation constant less than $3\mu\text{M}$. The weighted selectivity score $\sum_{i=1}^3 n_i \cdot c_i$, where n_i is the number of dissociation constants within threshold i and c_i is the corresponding weight, is an unbiased measure that enables quantitative comparisons between compounds and the detailed differentiation and analysis of interaction patterns. Scores ranged from 3.5 for GW-2580 to 146.70 for Sunitinib.

B. Selectivity Classes Maximizing Inferential Accuracy

Once a weighted selectivity score for a particular kinase inhibitor has been determined, the kinase inhibitor may then be classified into one of three selectivity classes, representing specificity (class S), promiscuity

(class P), or neither specificity nor promiscuity (class N), based upon the weighted selectivity score. The selectivity scores of kinase inhibitors are assumed to be generated by a sequence of probability distributions in which each distribution generates one class. Since the distribution parameters such as mean, standard deviation, and probability for the three selectivity classes are not known, an Expectation-Maximization (EM) algorithm [9], [10], [11] may be used to find these unknown probability distributions.

To estimate the range of weighted selectivity scores for the incomplete data using the EM algorithm, we start with an initial guess of the mean and standard deviation $\mu_S^{(0)}, \delta_S^{(0)}$ for selective compounds and an initial guess of the mean and standard deviation $\mu_N^{(0)}, \delta_N^{(0)}$ for non-selective compounds (Step 1). Next, we calculate the probability w_i for each compound x_i to be selective using the corresponding probability distributions (Step 2). Then, the new guessing values are calculated as $\mu_S^{(k+1)} = \frac{\sum_{i=1}^{10} w_i \cdot x_i}{\sum_{i=1}^{10} w_i}$, $\delta_S^{(k+1)} = \frac{\sum_{i=1}^{10} w_i \cdot (x_i - \mu_S^{(k)})^2}{\sum_{i=1}^{10} w_i}$, $\mu_N^{(k+1)} = \frac{\sum_{i=1}^{10} (1-w_i) \cdot x_i}{\sum_{i=1}^{10} (1-w_i)}$, and $\delta_N^{(k+1)} = \frac{\sum_{i=1}^{10} w_i \cdot (x_i - \mu_N^{(k)})^2}{\sum_{i=1}^{10} w_i}$ (Step 3). After some repetitions, the algorithm converges to a local maximization of the log probability of the observed data (Step 4). In this example, we only use the weighted selectivity scores of 10 compounds. The probability distributions show that the range for selective compounds is [0,40). Similarly, the range of weighted selectivity scores for non-specificity can be partitioned into a range that represents promiscuity, and a range that represents neither of the two.

C. Modified Lorenz curves and the Gini coefficients

Once a kinase inhibitor has been placed into a selectivity class, a kinase inferential bases may be determined. A straightforward approach that evaluates all possible subsets of kinases and finds the smallest one with the highest predictive accuracy would be an impossible task even for a computer, as there are currently 2^{317} (or approximately 2.7×10^{95}) subsets to evaluate. Accordingly, as recognized by the methods of the present disclosure, a better approach for finding a target universe is to determine which kinases are crucial in deciding the selectivity of inhibitors. It has been found that not all kinases are equally crucial in deciding the selectivity of inhibitors. Randomly chosen subsets of kinases will not give an accurate measure of selectivity. Figures 1 and 2 show the average accuracy for predicting the selectivity status of a compound

when 10 randomly selected kinases and 300 randomly selected kinases are used.

Furthermore, naive use of machine learning techniques for predicting the selectivity of a kinase inhibitor also may yield an unsatisfactory result because (a) an inhibitor still has to be screened against almost the whole set of kinases and (b) the accuracy for the prediction is not high.

To improve the accuracy, a Gini-based method for ranking the kinases due to its ability to overcome biases may be used. We consider a simple example of the dataset D of affinities with respect to a kinase A where D has d elements and three classes. The values were discretized into three ranges. When this kinase is evaluated by the current methodologies, for example by calculating the Gini index $gini_A(D) = \sum_{i=1}^m \frac{|R_i|}{d} \cdot gini(R_i)$, the first two rows (called partitions) contribute equally to the Gini index because $gini(R_i) = 1 - \sum_{j=1}^n p_{i,j}^2$ where $|\cdot|$ is the cardinality and $p_{i,j} = \frac{|C_{i,j}|}{|R_i|}$ is the relative frequency of class C_j in partition R_i . That said, when one just considers the probability distribution without taking into account the order of the classes, the first two partitions will be considered the same.

Clearly, the two partitions should not be considered the same because partition R_1 says that 75% of drug inhibitors with affinity values within this range are classified into Class C_3 while partition R_2 says that 75% of drug inhibitors with affinity values within this range are classified into Class C_2 . Hence in order to have a robust kinase selection method, one has to differentiate the partitions with different class orders because they have different amount of information. To solve this problem, we modified the well known Lorenz curves, a common measure in economics to gauge the inequalities in income and wealth. In [13], [14], we illustrated how modified Lorenz curves and modified Gini coefficients are calculated. The Equality Line (Eq) is defined based on the percentages of elements in $|C_1|, |C_{1..2}| = |C_1| + |C_2|, \dots, |C_{1..n}| = \sum_{i=1}^n |C_i|$ at x -coordinates $0, 1/n, 2/n, \dots, 1$, where n is the number of classes and $|C_1| \leq |C_2| \leq \dots \leq |C_n|$. The Lorenz polygon $L(R_j)$ of a partition, say R_j , is defined based on the percentage of elements in $|C_1^j|, |C_1^j| + |C_2^j|, \dots, \sum_{i=1}^n |C_i^j|$ at x -coordinates $0, 1/n, 2/n, \dots, 1$. The Gini coefficient of a partition, say R_j , is defined as $(\int_0^1 L(R_j) \cdot dx - \int_0^1 Eq \cdot dx) / \int_0^1 Eq \cdot dx$. One can easily see that the partitions with different class orders are now differentiated.

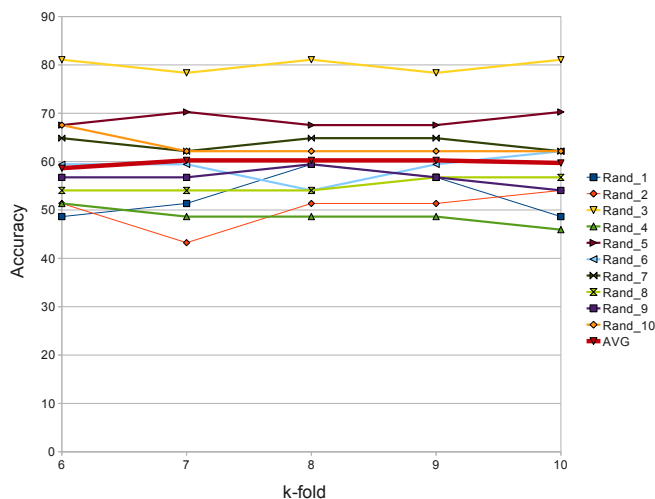


Fig. 1. Accuracy of predicting the specificity or promiscuity when 10 random subsets (Rand_i, $i=1..10$) each containing 10 randomly chosen kinases were used. The average accuracy (AVG) is $\sim 60\%$.

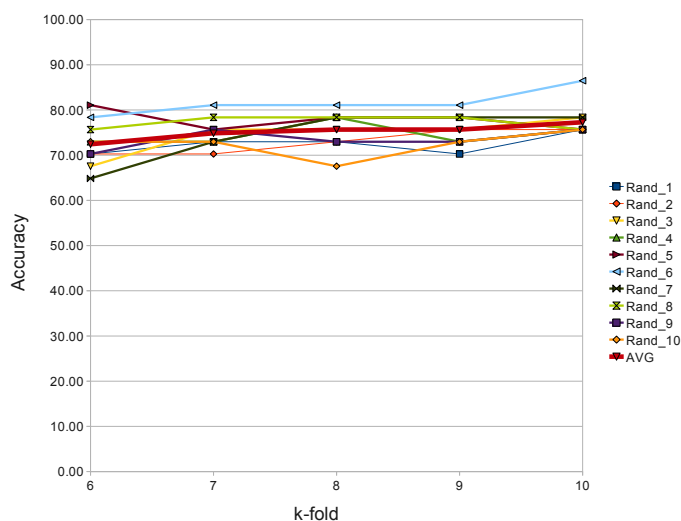


Fig. 2. Accuracy of predicting the specificity or promiscuity when 10 random subsets (Rand_i, $i=1..10$) each containing 300 randomly chosen kinases were used, the average accuracy (AVG) is $\sim 75\%$.

D. Kinase Bases

This section is devoted to systematically finding small inferential bases. Our method reveals that not all kinases are equally important in inferring the selectivity of inhibitor, which explains why screening of randomly chosen subsets of kinases [12] cannot give an accurate measure of selectivity. Furthermore, our experiments show that a direct use of traditional machine learning techniques for predicting the selectivity of a kinase inhibitor will give an unsatisfactory result because the accuracy for the prediction would not be high. As an

illustration, the highest accuracy of 78.4% was obtained when using all 317 kinases with the available machine learning techniques in Weka [15] in that 2 out of 9 promiscuous inhibitors were falsely predicted as neither promiscuous nor specific, 1 out of 9 specific inhibitor was falsely predicted as non promiscuous nor specific, and 5 out of 19 non promiscuous nor specific inhibitors were predicted specific.

It is our intention to first find a small target universe, which is the inferential basis for promiscuity or specificity for kinase inhibitors before using machine

learning techniques for predicting the selectivity. But a question arises: *how small a target universe can be?*

As we mentioned before, it is infeasible to test all combinatorial possibilities from 317 kinases for finding the smallest subset of kinases that any compound needs to be screened against to obtain an accurate indication of specificity or promiscuity because such a task amounts to the building and testing of 2.7×10^{95} predictive models. It also requires many thousand years to finish.

To reduce the size of the kinase subsets and to improve the prediction accuracy at the same time, we use the modified Lorenz curves and the Gini coefficients [16], [13], [14], which take into account the order of the classes and the order of affinity values, for selecting relevant kinases.

To further reduce the size of the kinase subsets we use different subset search techniques such as Correlation-based Feature Selection (CFS) [17] to create smaller kinase subsets. Since exhaustive search is infeasible, other searching strategies must be employed to identify the optimal kinase subsets. We use best-first and greedy search methods in the forward and backward directions as explained below. Greedy search considers local changes to the current subset through the addition or removal of kinases. For a given 'parent' set, a greedy search examines all possible 'child' subsets through either the addition or removal of kinases. The child subset that shows the highest goodness measure then replaces the parent subset, and the process is repeated. The process terminates when no more improvement can be made. Best-first search is similar to greedy search in that it creates new subsets based on the addition or removal of kinases to the current subset. However, it has the ability to backtrack along the subset selection path to explore different possibilities when the current path no longer shows improvement in terms of inferential power. To prevent the search from backtracking through all possibilities in the kinase space, a limit is placed on the number of non-improving subsets that are considered. Subsets of kinases that are highly correlated with the class while having low inter-correlation are preferred. In our evaluation we chose a limit of five. These subset search techniques resulted in a subset of 11 kinases that provides the most accurate indication of specificity or promiscuity.

E. Building and Verifying the Predictive Models

To build a predictive model for specificity and promiscuity, we exploited Bayesian networks [18], [19],

[20], [21]. The networks are structured as a combination of a directed acyclic graph of nodes and links, and a set of conditional probability tables. Nodes represent kinases or classes, while links between nodes represent the relationship between them. Conditional probability tables determine the strength of the links. There is one probability table for each node that defines the probability distribution for the node given its parent nodes. If a node has no parents the probability distribution is unconditional. If a node has one or more parents the probability distribution is a conditional distribution, where the probability of each feature value depends on the values of the parents. From our experiments, predictive models using Bayesian networks give a better accuracy for our kinase selectivity prediction.

To test the accuracy of our kinase selectivity models, we use k -fold cross validation, which is a common method for estimating the error of a model on some benchmark medical data sets [22]. The reason for using this testing approach is that when a model is built from training data, the error on the training data is a rather optimistic estimate of the error rates the model will achieve on unseen data. The aim of building a model is usually to apply the model to new, partially screened compounds—we expect the model to generalize to data other than the training data on which it was built. Another reason for using this testing approach is that the available kinase-inhibitor data sets are small and no test data set is available. It is well-known that k -fold cross-validation is very useful for this type of data sets [22].

For a reliable evaluation of the accuracy, we test the classification algorithm for $k = 6..10$. For each value of k , the data set D is randomly divided into k subsets D_1, D_2, \dots, D_k . Each time we leave out one of the subsets $D_i, i = 1..k$ to be used as a test data set for cross validation. The remaining subset $\cup_{j \neq i} D_j$ is used to build the model. The cross validation costs computed for each of the k test samples are then averaged to give the k -fold estimate of the cross validation costs. To ease the effects of the random partitions on the data set, this whole process is repeated 10 times and the results are then averaged again to give the estimated accuracy of the comparing algorithms.

III. RESULTS & DISCUSSION

A. Input screening data

To build predictive models and test our algorithms, we use the comprehensive data from a previous

competition-binding assay [12], where 38 kinase inhibitors were screened against a panel of 287 distinct human protein kinases, three lipid kinases and 27 disease-relevant mutant variants. The kinases in the assay represent 55% of the predicted human protein kinome. The compounds tested included 21 tyrosine kinase inhibitors, 15 serine-threonine kinase inhibitors and 1 lipid kinase inhibitor. We excluded staurosporine from our selectivity analysis due to its obvious promiscuity and lack of therapeutic value, but will use it for validating the accuracy of our results. Each compound was screened against the panel of 317 kinases at a single concentration of 10 μ M to identify candidate kinase targets, and for each interaction observed in this primary screening, a dissociation constant (K_d) was quantitatively determined.

B. Very Small Sets of Targets

Using the methods described above, 85 highest ranking kinases were selected with respect to the LorenzGini indexes from the original set of 317 kinases. The best-first and greedy searches were then used to find an optimal subset of 11 kinases: RET, SLK, FGFR2, FGR, FLT3(D835H), GAK, JAK2(Kin.Dom.2AJH1-catalytic), KIT(D816V), MAP4K3, ABL1(E255K), and CIT. After evaluating all possibilities of bases for this small subset of 11 kinases, a small inferential basis with five kinases was found: AMPK-alpha1, FGR, FLT3(D835H), LOK, GAK. The affinity interactions of kinase inhibitors with these five kinases gave a robust measure of specificity or promiscuity with 100% accuracy. All testing inhibitors were predicted correctly whether it is specific, promiscuous or none of those. While this small inferential basis of five kinases predicted the specificity and promiscuity of kinase inhibitors with 100% accuracy, random sets of the same number of kinases gave an accuracy of approximately 52%.

Other small inferential bases of 5 kinases were found using the above methods, which included SLK, FGR, FLT3 (D835H), GAK and JAK2 and SLK, FGFR2, FLT3 (D835H), GAK and JAK2. The affinity interactions of kinase inhibitors with these two set of five kinases also gave a robust measure of specificity or promiscuity with 97.3% accuracy. Again, all inhibitors that were specific were correctly predicted, and all inhibitors that were promiscuous were correctly predicted. The only false prediction was a non-specific, non-promiscuous inhibitor that was falsely predicted as promiscuous.

Using the above methods, inferential bases with four, three and two kinases were also determined. A two kinase basis consisted of GAK, and MAP4K5 gives a robust measure of specificity or promiscuity with 89.2% accuracy. All inhibitors that were specific were correctly predicted, and all inhibitors that were promiscuous were correctly predicted. Only 3 out of 19 inhibitors that were neither specific nor promiscuous were falsely predicted as specific. Only 1 out of 19 inhibitors that was neither specific nor promiscuous was falsely predicted as promiscuous.

Strikingly, the small inferential basis of two kinases GAK, and MAP4K5 gives a kinome-wide measure of specificity or promiscuity that is more robust than what a random subset of 300 kinases can give. This attests to the importance of our results.

Furthermore, the inferential bases have some very special features. For example, if a kinase inhibitor hits all kinases SLK, FGFR2, FLT3 (D835H) and GAK of the inferential basis, it is promiscuous. If it misses all kinases of the inferential basis, it is specific.

C. Discussion

The identification of kinases such as SLK that can only be targeted with promiscuous ligands poses a major challenge to structural biologists. This challenge may be summarized by the following question: Why is the affinity for this kinase driven exclusively by targeting highly conserved structural features? Direct examination of the kinase-inhibitor complex with PDB accession 2UV2 reveals two backbone-drug hydrogen bonds which surely promote promiscuity. These bonds involve backbone proton donors and acceptors (amides and carbonyls) of the nucleotide-binding loop (residues 109 & 111) whose spatial orientation is highly conserved across the kinase super-family [23]. What remains to be proven in this case is that SLK can only be targeted by forming such intermolecular bonds. To the best of our understanding this is an unsolved problem. Thus, the dearth of structural information and our inability to identify the structural culprits of promiscuity at this juncture make the machine-learning approach described in this work a most valuable predictive tool for molecular therapy.

IV. CONCLUDING REMARKS

We presented a novel method enabling a highly accurate prediction of a single attribute of a kinase inhibitor, its promiscuity or specificity. The inference

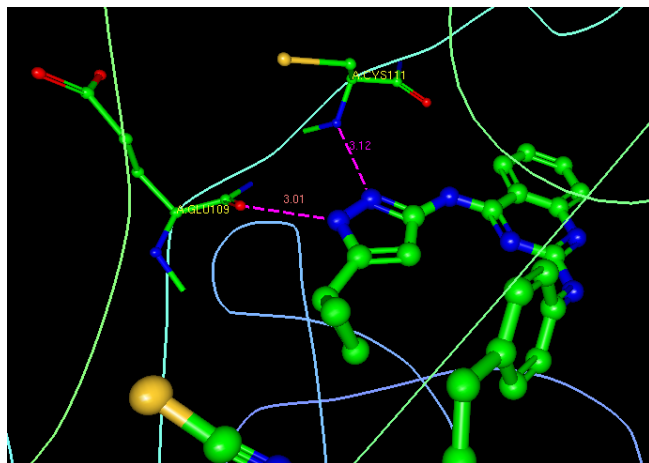


Fig. 3. Examination of the kinase-inhibitor complex with PDB accession 2UV2 for the SLK protein kinase reveals two backbone-drug hydrogen bonds which surely promote promiscuity. Pictured from RCSB PDB Ligand Explorer 3.9 (powered by MBT) running on 2UV2.

is based on a screening against a very small (2 to 5) set of target kinases. The target kinases have some very special features: if a kinase inhibitor hits all target kinases then it is promiscuous. If the kinase inhibitor misses all kinases of the inferential basis, it is specific. By dividing the experimental affinity fingerprinting of 37 inhibitors against 317 kinases into every possible testing set of 4 inhibitors and a training set of 33 inhibitors, we obtained an accurate prediction in all kinase inhibitors. This level of performance is reflective of 100% accuracy for an optimal inference base of 5 kinase targets. Smaller bases yield only slightly less accuracy. The method is build on a Bayesian Model and uses as input a five-entry affinity vector. The method is expandable whenever more kinases can be screened. Its full implementation can be provided in Supplementary Material. The main conclusion of this study is that the choice of a multi-pronged or highly specific molecular therapeutic agent is strictly conditioned by the desired clinical impact on a very small set of targets. The type of clinical impact determines whether promiscuity or specificity is the relevant and inescapable therapeutic constraint.

This work highlight the important of systematic search algorithm and undermines random screening as a tool of pharmaceutical informatics relevant.

ACKNOWLEDGMENT

This work is supported in parts by NSF award #0917257.

REFERENCES

- [1] Vogelstein, B. & Kinzler, K. Cancer genes and the pathways they control. *Nature Medicine* **10**, 789–799 (2004).
- [2] Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006). URL <http://www.sciencemag.org/content/314/5797/268.abstract>.
- [3] Whibley, C., Pharoah, P. D. P. & Hollstein, M. p53 polymorphisms: cancer implications. *Nat Rev Cancer* **9**, 95–107 (2009).
- [4] Frantz, S. Drug discovery: playing dirty. *Nature* **437**, 942–943 (2005).
- [5] Keith, C., Borisy, A. & Stockwell, B. Multicomponent therapeutics for networked systems. *Nat. Rev. Drug. Discov.* **4**, 71–78 (2005).
- [6] Mencher, S. K. & Wang, L. G. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.* **5**, 3–9 (2005).
- [7] Hopkins, A., Mason, J. & Overington, J. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127–136 (2006).
- [8] Hampton, T. "promiscuous" anticancer drugs that hit multiple targets may thwart resistance. *JAMA* **292**, 419–422 (2004).
- [9] Baum, L. E., Petrie, T., Soules, G. & Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Stat.* **41**, 164–171 (1970).
- [10] Dempster, A., Laird, N. & Rubin, D. Maximum likelihood from incomplete data via em algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–38 (1977).
- [11] Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm? *Nature Biotechnology* **26**, 897–899 (2008).
- [12] Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat Biotech* **26**, 127–132 (2008). URL <http://dx.doi.org/10.1038/nbt1358>.
- [13] Tran, Q.-N. Mining medical databases with modified Gini index classification. In *Proceedings of IEEE-ITNG 2008 Conference* (IEEE, Las Vegas, Nevada, 2008).
- [14] Tran, Q.-N. Microarray data mining: A new algorithm for gene selection using Gini ratios. In *Proceedings of IEEE-ITNG 2010 Conference* (Las Vegas, Nevada, 2010).
- [15] <http://www.cs.waikato.ac.nz/ml/weka> (2009).
- [16] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees* (Wadsworth & Brooks/Cole Advanced Books & Software, 1984). Monterey, CA.
- [17] Hall, M. A. *Correlation-based Feature Subset Selection* (Hamilton, NZ, 1998).
- [18] Cooper, G. F. & Herskovits, E. A bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9**, 309–347 (1992).
- [19] Heckerman, D., Geiger, D. & Chickering, D. M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20**, 197–243 (1995).
- [20] Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian network classifiers. *Machine Learning* **29**, 131–163 (1997).
- [21] Russell, S. J. & Norvig, P. *Artificial Intelligence - A Modern Approach* (3. internat. ed.) (Pearson Education, 2010).
- [22] Witten, I. H. & Frank, E. *Data Mining: Practical machine learning tools and techniques* (Morgan Kaufmann, 2008), 2nd edn.
- [23] Pike, A. C. W. *et al.* Activation segment dimerization: a mechanism for kinase autophosphorylation of non-consensus sites. *EMBO J.* **24**, 704–714 (2008).

Co-evolutionary Multi agent Pharmacoinformatics

Tagelsir Mohamed Gasmelseid

Department of Information Systems, College of Computer Sciences & IT, King Faisal University, Al Ahsaa, Saudi Arabia

Abstract - *The expansion of drug-related problems motivated healthcare organizations to use Pharmacoinformatics in pursuit of improving communications, signaling, analyzing and reporting of adverse drug reactions and facilitating scenario-based interventions. Despite their benefits, the use of such systems is limited and is not delivering a real value to healthcare professionals in different environments. This project aimed at the development of a reference multi-agent Pharmacoinformatics model that can be used to improve the quality of pharmaceutical care provided and the management of hospitals. The model reflects three main modules: a data capture and update module, diagnosis module and a pharmaceutical care and drug monitoring module. The study also reflected on the need to adopt co-evolutionary concepts which are related to system thinking and sociomateriality considerations.*

Keywords: software agents, pharmaceutical care; bioinformatics; adverse drug events; interactions

1 Introduction

The diversity of diet, traditions of societies, differences of diseases and prescribing practices and the growing use of Herbal medicines have been associated with a wide range of drug-induced complications originating from unexpected adverse effects. Therefore, healthcare organizations started to develop different methods, procedures, processes and systems to identify, analyze and manage adverse drug reactions (ADRs). Over the last couple of years there has been a growing interest in using information systems (such as Pharmacoinformatics) to increase information accessibility to healthcare providers, enhance outcomes and improve convenience for patients. Pharmacoinformatics is concerned with the use of information systems for the improvement of pharmacy decision making. It assists in the assessment and management of therapeutic outcomes in patients as well as in detecting, signaling, evaluating, and solving potential and actual drug-related problems (including adverse drug events or drug interactions) [1]. They aim at improving the capacity of clinical practitioners to efficiently acquire and develop new treatment strategies through the facilitation of information exchange, supporting the detection and management of adverse drug events and enabling supply chain management process. But in real practice, the realization of such applications is still limited. Pharmacoinformatics

applications are used as sub-modules and tend to be limited to stock control, monitoring drug availability and issuance at outpatient and ward pharmacies. Even for stock control, there seems to be no emphasis on the use of electronic ordering and procurement processes for which no standard operating procedures exist. The analysis of drug therapies and management of prescription inconsistencies are not supported. There is no support for signaling and detecting adverse drug events manifested in patients and recorded by healthcare professional. Pharmacoinformatics therefore are not used at early stages of the medication process for the analysis of drug therapies, cross-checking prescriptions, alerting physicians and other medical professionals about the non-existence of drug prescription pre-requisites such as general lab analysis and recommending change of drug regimens and the use of free-use drugs wherever applies. There is no link between hospitals and Pharmacovigilance centers (PV) to enable the reporting and tracking of adverse events. The difficulty of restructuring hospital-wide processes has also limited the capacity of healthcare organizations to use Pharmacoinformatics in a patient-oriented format (to improve the knowledge of patients about drugs, the dysfunctional consequences of adverse drug events and their capacity and willingness to report such events).

On the other hand, Pharmacovigilance Centers justify their existence by claiming that "pharmaceutical care systems in public hospitals are weak enough to provide relevant real time information about ADRs". Despite the fact that PVs are using specialized software (such as Empirica Trace), they have also been facing considerable difficulties in collecting and analyzing ADR signals using spontaneous reporting procedures widely used by them. Such inability to address ADRs will continue to grow if the concerned pharmaceutical care authorities continued to adopt fragmented thinking with regards to their information systems. Because of the "structural" and "functional" couplings that exist among such systems, they should "evolve together" rather than to assume the possibility of information sharing. This paper examines the context of co-evolution between pharmaceutical care systems at the level of "hospitals" and "Pharmacovigilance Centers" and proposes the use of a multi-agent Pharmacoinformatics reference model.

2 Methods

The methodology for this project was generally a descriptive analytical survey with both inductive and deductive methods applied including empirically driven qualitative and quantitative theories. To account for the diversity of information across the different managerial landscapes and to ensure the validity of the instruments of research, Anthony's taxonomy of managerial levels, information modeling, agent oriented software engineering, and other "process-centered" and "resource-oriented" approaches are used for the articulation of variables and the development of the entire model and its corresponding modules. The majority of the project's data was collected from electronic sources of many healthcare organizations and regulatory agencies such as the National Pharmacovigilance Center at the Food and Drugs Authority (FDA). The study also used a variety of tools and methods of analysis of data provided by international organizations such as WHO. The tools include data matrix, tables, diagrams, models and output of computer programs. Data has been also compiled from the use of questionnaires and interviews with key personnel in different healthcare organizations across the kingdom. In order to reduce the variance of estimators and gain sampling precision, the technique of optimum allocation has been used [2].

3 Related work

Software agents are computational entities that perform some tasks on behalf of their users, other agents or programs with some degree of autonomy using the appropriate information and communication platforms. Their roles include task delegation, users training, event monitoring, information search, matchmaking and filtering [3]. They possess important properties such as autonomy, social ability, reactivity and pro-activeness, learning, mobility, benevolence, rationality, independence, cooperation, reasoning, intelligence and adaptivity [4][5][6][7]. In complex systems such as healthcare, agents are used in the form of multi-agent organizations. A multi-agent system includes multiple heterogeneous agents who interact and exchange information in a decentralized and "social" manner to solve larger and complex problems. With regards to their use in pharmaceutical care, the use of multi-agent system is recognized as a sub module under the umbrella of the entire hospital information system. According to [8], multi-agent systems can be structured into different ways such as "organizational structuring", "contracting", "multi-agent planning" and "peer-to-peer negotiation". The complications experienced in the healthcare sector has been accompanied with the tendency to address pharmacy-related issues in separate applications, such as the use of multi-agent systems for monitoring and reporting of adverse drug events, electronic prescriptions, managing drug therapies and

mainstreaming pharmaceutical procurement activities, among others.

In their work about drug prescription, [9] used multi-agent systems to monitor the prescription of restricted use antibiotics within the context of an electronic Institution (i.e., hospital) incorporating agents, roles, dialogic framework, scenes, and performative structure. The architecture included six types of (functional) agents: patient (a.k.a guardian angel), physician secretary, laboratory manger, pharmacy expert, and nurse agents. Different scenes were used to address communication among the agents such as: Patients Room, Physicians Room, Laboratory, Pharmacy, Dialog scene, Waiting Room and halls. The main functionalities are antibiogram authorization, antibiogram results and modification of the entire electronic medical record. However, despite its advantages, the architecture has a limited scope to provide comprehensive Pharmacoinformatics assistance. Multi-agent systems have also been used for revising therapies (getting clinical information, deciding alternative therapies, etc.) and signaling of adverse drug events. Their use in these functions is motivated by the need for collaboration and exchange of complementary skills from different experts (e.g. epidemiologists, biostatisticians, pharmacists and physicians) for the analysis and interpretation of reports, collecting additional relevant information, and drawing reliable conclusions [10]. Also [11] and [10] used intelligent agents with a fuzzy recognition-primed decision model to develop a distributed adverse drug reaction detection system by utilizing distributed electronic patient data. The Recognition-Primed Decision (RPD) model is generalized to a fuzzy RPD model and utilizes fuzzy logic technology to represent, interpret, compute imprecise and subjective cues that were commonly encountered in ADRs and retrieve prior experiences reported by patients, physician and hospitals. Multi-agent systems have also been widely used for the management of pharmaceutical supply chains [12][13][14][15][16][17]. [18] proposed a generic process-centered methodological multi-agent supply chain management framework. [19] presented architecture for strategic information systems and [20] discussed the use of multi-agent systems and radio frequency identification (RFID) technologies to track pharmaceuticals supplies. [21] presented a multi-agents system collaborative production system to support the collaborative and autonomous mold manufacturing outsourcing processes. [22] focused on merging remote sensing data and population surveys in large, empirical multi-agent models. [23] combined a multi-agent framework with ontology-driven solutions to support and automate the procurement process. [24] developed a multi agent system to simulate the supply chain of the pharmaceutical industry.

Despite their benefits, the use of multi-agent systems has some limitations that need to be taken into account. Such

limitations include the problems of domain specification (agent-oriented problem formulation), communication problems (the most suitable protocols and languages necessary for enabling a possibly sophisticated and meaningful interaction among agents, co-ordination problems (the enforcement of the necessary teamwork), computational problems (designing and implementing multi agent systems in a way that avoids computational overload by means of load balancing strategies), implementation problems (the techniques and tools needed to support multi-agent system design and implementation in a safe, easy, and productive way) and the verification problem [25]. To relax such limitations a wide range of techniques and communication languages have been developed [26][27][28][29][30][31][32][33][34].

4 Co-evolutionary Pharmacoinformatics

In real terms, strong functional and structural couplings exist between the socio-technical configurations of the healthcare system and other systems. According to [35], functional coupling refers to the context where strong input-output relations exist between different regime elements of the same regime and/or across interacting regimes. Structural coupling refers to the situation in which the interacting regimes share the elements of socio-technical configurations (e.g., infrastructure, actor networks, technologies, institutions) or having a joint application regulations used by the two regimes. Such coupling necessitates the importance of understanding context-based interactions among healthcare systems (including pharmaceutical care) but most importantly to visualize and examine the way systems co-evolve together in a dynamic pattern. This calls for a migration from system-based interactions to ensure that the use of software agents in pharmaceutical care reflects an integrated pattern of information acquisition and visualization. Co-evolutionary positions focus on optimizing the functionality of pharmaceutical care information systems not only at their local levels (healthcare institutions) but also at their surroundings. Such objective can't be maintained by focusing on meeting management information requirements and decision support for the management of pharmaceutical care processes and providing inputs to other related (yet independent institutions such as Pharmacovigilance Centers operated by Food and Drug Authorities) by filling forms. Instead, the improvement of country-wide pharmaceutical processes and systems requires that such systems innovatively "co-evolve together" rather than act in an input-provision (and sometimes regulatory) format. Borrowing from the concepts of system innovation and transition thinking, such co-evolution allows for the appropriate examination of all inevitable functional and structural couplings and enables the operationalization of innovations and transitions in an agent-oriented format, at the levels of niches, regimes and landscapes of the entire pharmaceutical care processes. Niches represent new and relatively instable set of rules and

institutions for innovative practices. A regime represents a set of cognitive, regulative and normative rules or institutions that are coherent and guide the choices and behavior of the actors in that regime. The landscape offers a metaphor for the background setting and developments for regimes and niches and the possibilities for regime change, including structural socio-economic, demographic, political and international developments [36]. It represents the source of pressure on the regime to change and the behavior and choices of actors. Each regime may change itself automatically by restructuring their internal goals and resources but others may wait for the emergence of change in other interrelated domains of changing or stable regimes as it is the case of shared regulations. Figure (1) below depicts the pharmaceutical care as a socio-technical system.

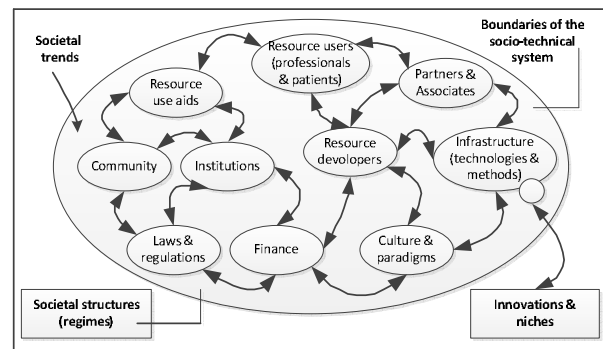


Fig. 1 pharmaceutical care socio-technical system

A typical out-patient pharmacy-based information system in a hospital follows a functional format by including patient-oriented, diagnosis-based and drug inventory management information. With different levels of abstraction, it includes information about patients, active diagnosis, pharmacy orders and delivery (new, existing and regular orders, batch, dose, route, frequency, description, restricted medications (i.e., medications with conditions), authorization, ordering doctor, service department etc). The basic objective of such system is to provide necessary management information to enable executives to improve pharmaceutical care processes and enhance patient-oriented outcomes. The niches of such regime are depicted in figure (2) below.

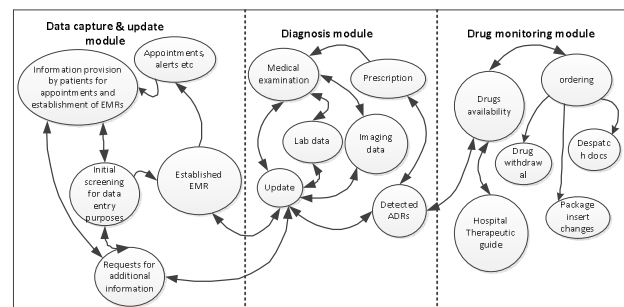


Fig. 2 hospital-based regime niche representation

A typical Pharmacovigilance Center (PV) aims at maintaining safety, quality, efficacy and accessibility of drugs

(human, veterinary, cosmetics and biological products) and providing accurate information to the public and healthcare professionals through administration of country-wide regulatory systems that incorporate international best practices. It undertakes regulatory processes (such as establishing and updating the regulations and licensing procedures that govern the establishment of pharmaceutical care institutions and the approval of the design and application of new drugs). To support the implementation of national plans, PV centers undertake research-and-development processes to administer post-market surveillance programs and develop port of entry inspection presence. They develop and update drugs and cosmetics listing databases and use them to advocate rational use of medicines and the detection of adverse drug reactions (ADRs) and their frequency rates. The processes used for the validation of signals examine three types of adverse drug events: A, B & C. Type (A) events are dosage-based and related to the pharmacokinetic properties of drugs; therefore, focus is usually made on improving dosage characteristics. Type (B) events are generally related to the patient's reactions and tend to be allergic, idiosyncratic, immunological, or non-immunological. Type (C) events tend to be serious and may result in significant implications on public health. They originate from the impacts of drugs used for improving the quality of life of patients with serious chronic diseases [37]. Figure (3) below depicts the niches of the entire Pharmacovigilance regime.

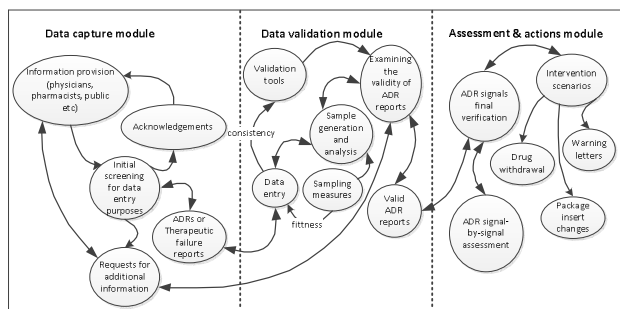


Fig. 3: PV regime niche representation.

The sustainability of the co-evolution processes and innovation experiments are benchmarked using niche attributes shown in figure (4) and figure (5) below. For the "data capture and update" niche of the hospital pharmacy regime the list of attributes include improved communication and data input, enhanced alerting abilities, improving the management of requests and managing established electronic medical records. The attributes for diagnosis niche include improved medical examination, diagnosis, prescription improved outcomes and reduced errors.

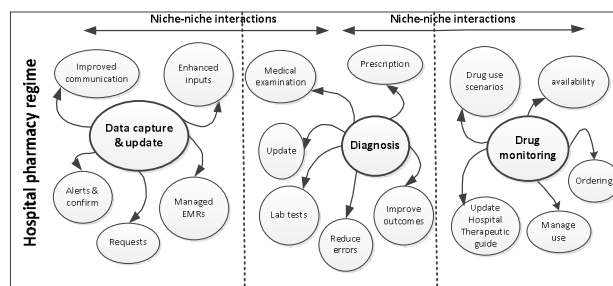


Fig. 4: hospital pharmacy regime attributes characterization

The attributes of the niches of the Pharmacovigilance regime are shown in figure (5) below.

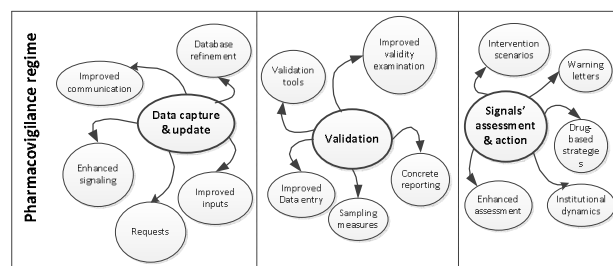


Fig. 5: Attribute characterization for the Pharmacovigilance regime

Based on such characterization, innovations and transitions that take place at the data capture niche of the hospital pharmacy regime for example (such as building and operating databases at outpatient pharmacies) significantly affect all niches of the PV regime. Niche-based transitions promote Pharmacoinformatics innovations and guide effective the design and implementation of organizational and structural transformations. A transition towards decentralized management of drug, for example, results into organizational change (such as organizational structures) and shifts of the context of decision making (decision partners and their roles and degree of workflow automation). Regime-based innovations shape the dynamics of Pharmacoinformatics applications and innovations, resource utilization and information sharing.

5 Model description

The co-evolutionary concepts discussed above have been used to develop the multi-agent Pharmacoinformatics model described in the following sub-sections.

5.1 Process modeling

The process model of the proposed multi-agent Pharmacoinformatics model is designed using TROPOS methodology. As shown in Fig. 1 below, the context of interaction among the stakeholders involved (directly or indirectly) in the process of pharmaceutical decision making reflects the environment of the entire problem. Patients, physicians, pharmacists and other supporting departments constitute the stakeholders whose interactions are reflected in

the functionality of the multi-agent organization and its model.

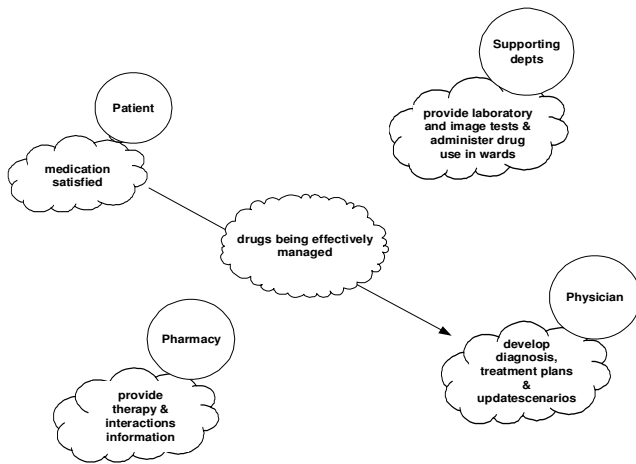


Fig. 1: Stakeholders representation and involvement [2]

The analysis of "flows" depicted in fig. 2 below shows the set of flows that take place within the domain of the problem solving process among patients, physicians, pharmacists and other departments. The flow also depicts the movement of "requests for drugs" and "drug orders" that govern the maintenance of drug availability for the treatment of patients. The central entity in the relational conceptualization is the "drug order" initiated and itemized by the physician (in the form of a prescription), hospital, outpatient or ward pharmacies. It represents an authorization to provide a specified type and quantity of a specified drug from a specific drug source to a specific user. Drug inventories exist at different locations such as hospital, outpatient and ward pharmacies and are assumed to be capable of satisfying "requests for drugs" and placing orders for different types of drug. The entity representing transactions that move drugs in, out and through the entire pharmacy network is a movement or a flow. When drugs (flow) are provided against a drug order (prescription), a flow (movement) is recorded by the drug provider from a particular location (e.g., outpatient pharmacy). Should there be an excess flow (drugs) provided by mistake to a patient, it should be monitored and addressed by the "providing" location or other entities in the hospital during the course of medication. This may entail another kind of movement-related transaction to record it. When "stock control" is carried out (with adjustments increasing or reducing the stock balances at different hospitals), an increase can be recorded as a movement to the "inventory" with no "from source of" being indicated and a decrease recorded as a movement from the "inventory" with no "to destination of" being identified. This means that physical flows are accompanied with information flows in different forms and formats.

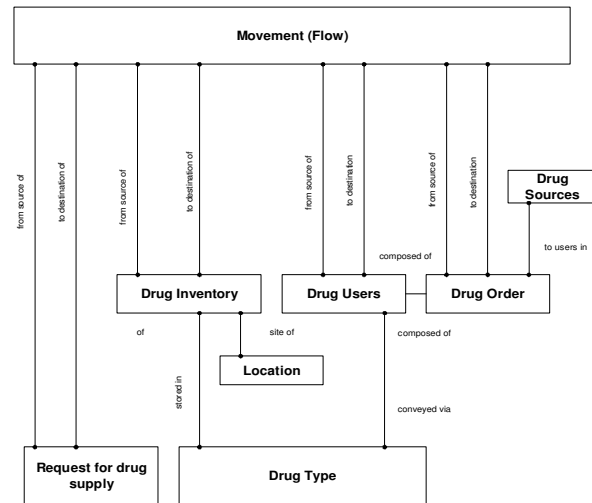


Fig. 2: Flow analysis in pharmaceutical care systems [2]

5.2 Multi-agent architecture

The architecture of the proposed reference model includes functional superior agents supported by subordinate ones (e.g. information, interface, task, etc) as shown in fig. 3 below. Superior agents representing health organizations (such as hospitals, Pharmacovigilance centers and Food and Drug Authorities (FDAs) act as "task mangers" and act on behalf of their users and subordinate agents for the implementation of hospital-wide functions organized in the form of business units such as surgery, pediatrics, etc. It also manages internal and external communication processes necessary for the management of the entire health organization and the coordination of its activities with other agencies such as Pharmacovigilance centers. They also manage the creation and access of electronic medical records (EMRs) and other data bases. Subordinate agents provide "functional", "information" or "interface" support to physicians, pharmacists and nurses, and other Superior agents. As shown in fig. 3 below, the hospital agent is responsible for the corporate management of the entire hospital by coordinating the efforts of other agents. It directs all other agents towards the realization of corporate objectives and represents a link between the agents interacting inside the hospital and other agents or entities in the external environment. Therefore, it accesses all functional databases and investigates the role of agents and their operating procedures in facilitating joint functionality. The patient agent exchanges information between the patient and other agents such as physicians and other personnel responsible for medical records, pharmacy and laboratory. Information exchange includes drug management, after-discharge complexities, drug reactions and alerts in relation to appointments, drug usages and confirmations. It also helps in reporting and signaling adverse drug reactions.

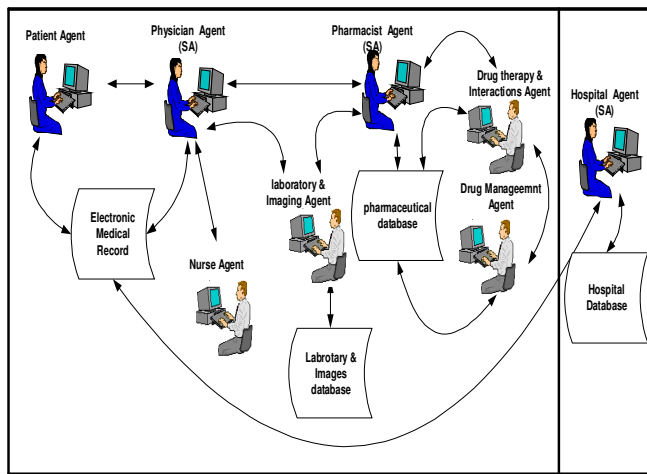


Fig. 3 Multi-agent Architecture [2]

The physician agent supports physicians to do processes, functions, schedules and communication (including alerts). It allows an interface between the physician, patients, pharmacists, nurses and other laboratory experts handling microbiological analyses, pharmaceutical screening and drugs ordering, modification and administration. It interacts with the patient's agent with regards to appointments, diagnosis, reporting of adverse drug reactions (especially post-discharge) and the confirmation and modification of drugs. The nurse agent collects medication instructions from the physician agent and monitors drug availability in wards in accordance with the medical order forms forwarded to it. The laboratory and images agent supports data acquisition, processing, update and communication among laboratory, imaging and other professionals in the hospital. It also interacts with other agents with regards to the requests and results of laboratory test requests. The pharmacy agent assists in the implementation of pharmaceutical functions and manages information acquisition and communication, analysis, reporting and recommendation of different alternative medication scenarios using its technical expertise and patient's data. It liaises between the hospital agent and its subordinate agents with regards to drug therapies, pharmaceutical and medical analyses, as well as signaling, analyzing and reporting of adverse drug reactions. It gets support from two task agents: (a) the drug therapy and interactions agent that uses information from the patient's records, medical and laboratory test results, and prescription records to assist in managing drug therapies and the recommendation of changes in medications and drugs based on a data-mining algorithm. It is responsible for screening drug therapies and signaling adverse drug reactions occurring for hospitalized and discharged patients throughout the medication process, in collaboration with the other agents and informing the pharmacy agent to take necessary actions. (b) the drug management agent responsible for drug management at the level of the hospital, outpatient and ward pharmacies, where it manages drug availability, procurement, ordering and removal of near-expiry drugs.

5.3 The data processing model

As shown in fig. 4 below, the data processing model includes three main modules: (a) data capture and update (implemented through the functionalities of the agents representing the patient, physician, laboratory and images, and nurses), (b) diagnosis (implemented through the agents representing physicians and pharmacists), and (c) drug monitoring (therapies, interactions analyses and requests for supply implemented through the agents representing physicians and pharmacists).

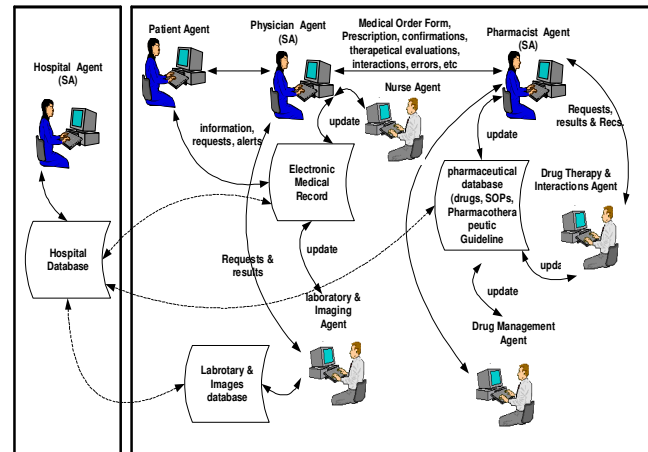


Fig. 4 data processing model [2]

In addition to its use of the database containing electronic medical records of patients, the model also incorporates a pharmaceutical database containing information about:

- Standard operating procedures that govern the examination of medical documents and requests coming to the pharmacy department, such as medical order forms and requests for drug supplies placed by different units in the hospital.
- Hospital Therapeutic Guide, which includes information about all routine drugs and medicines used in the hospital, as well as guidelines for acquiring and managing ad-hoc drugs. It also includes detailed recording of drug attributes such as correctness of drug types, dosage, route of administration, frequency, length of use etc.
- Allergy and microbiological information necessary for making associations between prescribed drugs and diseases, for example, isolated pathogen micro-organisms, and their corresponding sensitivities.
- Diseases-based drug groupings with each group including all drugs related to a particular disease for patients with certain allergies and symptoms.

6 Discussion

The use of multi-agent systems for the improvement of pharmaceutical care services brings to our attention a wider range of functional, organizational and technological concerns. The practical managerial context of the proposed

model is shown in its capacity to improve inter and cross-regime communication and the operating efficiency of pharmaceutical care processes. One of the main advantages is that drug prescription processes can be guided by a pharmaceutical cross-checking which investigates and alters the compliance with the standard operating procedures and requirements (such as microbiological analysis) in a way that minimizes medical errors and the occurrence of adverse drug reactions. This is expected to eventually improve signaling and analyzing ADRs information exchange at the level of hospital pharmacy and PV centers. The adoption of co-evolutionary concepts improves the performance of the corporate healthcare system and enhances cooperation. However, the use of such concepts has some critical impacts on healthcare organizations. The co-evolution of systems detects new axioms for developing multi-agent Pharmacoinformatics architectures and the level of change of focus, objectives and methods to be incorporated. Architecture specifies how the agent can be incorporated as a part of a multi-agent system and how these "parts" (hardware and/or software modules) should be made to interact using specific techniques and algorithms [38]. Existing methodologies are not satisfactory because they are based on the assumption that "any software life cycle, process or product model must be tailored towards the characteristic needs of the application domain of the target system [39]. The emerging co-evolutionary cross-regime interactions also incorporate information security issues that current risk-based models are not capable of addressing. Moreover, special attention is required with regards to the way the utility matrix of pharmaceutical and healthcare stakeholders is being viewed and optimized. Because of the rich domain of interaction, the co-evolution among systems and regimes significantly shapes the objective functions of policy makers in healthcare organization.

7 Conclusions

The emphasis of healthcare organizations on examining and analyzing ADRs will continue to grow and gather momentum attention of policy makers and the community at large. While the diversity of intervention mechanisms will continue to shape the responsiveness of healthcare organizations, deploying intelligent information systems in healthcare organizations is also expected to grow as a result of the foreseen technological developments. The limited use of multi-agent Pharmacoinformatics points towards high levels of expected support to be provided for pharmaceutical policy makers if additional effort is invested in addressing situation-specific and system development related issues. Especially in resource limited situations, the use of co-evolutionary measures proved to be useful for relaxing organizational, institutional and procedural issues. This is because such measures call for not only the change of "programs" but also the change of "mind sets". It is becoming of paramount importance that the evolution and

deployment of multi-agent Pharmacoinformatics being done in a co-evolutionary fashion in which the entire system is continuously cross-referenced with its operating environment. It is only through such thinking, radical innovations can be incorporated into its niches and the entire pharmaceutical care being made more and more patient-oriented.

8 References

- [1] Gasmelseid, Tagelsir. Pharmacoinformatics: Advanced information systems for improved pharmaceutical care. *Pharmacoinformatics and drug discovery: Technologies: Theories and Applications*, pp. 1-11, 2012.
- [2] Gasmelseid, Tagelsir. A reference multi-agent pharmacoinformatics model to improve hospital management. *Pharmacoinformatics and drug discovery: Technologies: Theories and Applications*, pp. 187-201, 2012.
- [3] Gasmelseid, Tagelsir. On the design of multi-agent, context aware and mobile systems. *Handbook on Modern System Analysis and Design Applications and Technologies*, Idea Group Publishing, USA, pp: 357-370, 2009.
- [4] Lisa, M; L. Hogg and N. Jennings. Socially intelligent reasoning for autonomous agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* **31** (5), pp. 381–393, 2001.
- [5] Persson; P; J. Laakso and P. Lönnqvist. Understanding socially intelligent agents: a multilayered phenomenon. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* **31** (5), pp. 349–360, 2001.
- [6] Bonarini, A. and V. Trianni. Learning fuzzy classifier systems for multi-agent coordination. *Information Sciences* **136** (1–4), pp. 215–239, 2001.
- [7] Hu, J and M. Weliman. Learning about other agents in a dynamic multi-agent system. *Cognitive Systems Research* **2** (1), pp. 67–79, 2001.
- [8] Anumba, C; A. Ren, O. Thorpe, O. Ugwu and L. Newnham. Negotiation within a multi-agent system for the collaborative design of light industrial buildings. *Advances in Engineering Software*. **34** (7), pp. 389-401, 2003.
- [9] Godo, L; J. Puyol-Gruart, J. Sabater, V. Torra, P. Barrufet and X. Fàbregas. A multi agent approach for monitoring the prescription of restricted use antibiotics. *Artificial Intelligence in Medicine*, **27** (3), pp. 259-282, 2003.
- [10] Yanqing, J; Y. Hao, J. Yen, Z. Shizhuo, M. Massanari and J. Barth. Team-based multi-agent system for early detection of adverse drug reactions in post marketing surveillance. In: *Proc Proceedings of the 24th North American Fuzzy Information Processing Society Ann Arbor, MI*, pp 644-649, 2005.
- [11] Yanqing, J; Y. Hao, S. Margo, S. Farber, Y. John, D. Peter, E. Richard and M. Michael. A Distributed, Collaborative Intelligent Agent System Approach for Proactive Post marketing Drug Safety Surveillance. *IEEE Transactions on information technology in Biomedicine*, **14** (3), pp. 826 – 837, May 2010..
- [12] Yanqing, J; Y. Hao, Y. John, Z. Shizhuo, C. Daniel, J. Barth, E. Richard and M. Michael. A distributed adverse drug reaction detection system using intelligent agents with a fuzzy recognition-primed decision model. *International Journal of Intelligent systems*, Volume **22**, pp.: 827–845, 2007.
- [13] Pathak, S; G. Nordstrom and S. Kurokawa. Modelling of supply chain: a multi-agent approach. *IEEE International Conference on Systems, Man, and Cybernetics*, **3**, pp. 2051-2056, 2000.
- [14] Fu, Y and R. Piplani. Multi-agent enabled modelling and simulation towards collaborative inventory management in supply chains. In *Proceedings of the 2000 Winter Simulation Conference J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.* pp.1763-1771, 2000.
- [15] Frey, D; T. Stockheim, P. Woelk and R. Zimmermann. Integrated Multi-agent-based Supply Chain Management. In *Proceedings of 1st International Workshop on Agent-based Computing for Enterprise Collaboration*. 2003.
- [16] Davidsson, P and F. Wernstedt. "A framework for evaluation of multi-agent system approaches to logistics network management". *Multi-Agent Systems: An Application Science*, Kluwer, 2004.

- [17] Walsh, W and M. Wellman. Modelling supply chain formation in multi-agent systems. In *Lecture Notes in Artificial Intelligence*, Vol. 1788, Agent Mediated Electronic Commerce II, Springer-Verlag, 2004.
- [18] Lu, L and G. Wang. A study on multi-agent supply chain framework based on network economy. *Computers & Industrial Engineering*, 54, pp. 288–300, 2008.
- [19] Govindu, R and R. Chinnam. MASCF: A generic process-centered methodological framework for analysis and design of multi-agent supply chain systems. *Computers & Industrial Engineering*, 53, pp. 584–609, 2007.
- [20] Shirazi, M and J. Soroor. An intelligent agent-based architecture for strategic information system applications. *Knowledge-Based Systems*, 20, pp. 726-735, 2007.
- [21] Turcu, C; C. Turcu, V. Popa and V. Gaitan. Identification and Monitoring of Patients Using RFID and Agent Technologies: Synergy and Issues. *Electronics and electrical engineering*, 6(86), pp. 17-22, 2008.
- [22] Trappey, A; Lu and L. Fu. Development of an intelligent agent system for collaborative mold production with RFID technology. *Robotics and Computer-Integrated Manufacturing*, 25, pp. 42–56, 2009.
- [23] Seyed, M; M. Rizzi, M. Maciej and G. Armando. Merging Remote Sensing Data and Population Surveys in Large, Empirical Multi-agent Models: The Case of the Afghan Drug Industry. Presented during the *Third World Social Simulation Congress* in Kassel, Germany. 2010. Retrieved April, 30, 2011 from: <http://www.css.gmu.edu/projects/irregularwarfare/remotesensing.pdf>.
- [24] Gottfried, K; M. Merdan, W. Lepuschitz, T. Moser and C. Reinprecht. Multi Agent Systems combined with Semantic Technologies for Automated Negotiation in Virtual Enterprises. *Multi-Agent Systems - Modelling, Control, Programming, Simulations and Applications*. InTech, pp. 221- 242, 2011.
- [25] Gaurav, J; R. Christian and H. Robert. A Multi-Agent Simulation (MAS) of the Pharmaceutical Supply Chain (PSC). *POMS 20th Annual Conference*, Orlando, Florida U.S.A., May 2009.
- [26] Sycara, K. Multiagent Systems. *AI Magazine* 19 (2), pp.79-92, 1998.
- [27] Cammarata, S; D. McArthur and R. Steeb. Strategies of Cooperation in Distributed Problem Solving. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, pp. 767–770. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, 1983.
- [28] Durfee, E. A Unified Approach to Dynamic Coordination: Planning Actions and Interactions in a Distributed Problem Solving Network, Ph.D. dissertation, Department of Computer and Information Science, University of Massachusetts, 1987.
- [29] Huhns, M and D. Bridgeland. Multiagent Truth Maintenance. *IEEE Transactions on Systems, Man, and Cybernetics* 21(6), pp. 1437–1445, 1991.
- [30] Mason, C and R. Johnson. DATMS: A Framework for Distributed Assumption- Based Reasoning. In *Distributed Artificial Intelligence, Volume 2*, eds. M. Huhns and L. Gasser, pp. 293–318. San Francisco, Calif.: Morgan Kaufmann, 1989.
- [31] Loui, R. Defeat among Arguments: A System of Defeasible Inference. *Computational Intelligence* 3 (1), pp. 100–106, 1987.
- [32] Hewitt, C. Offices are Open Systems. *ACM Transactions of Office Automation Systems* 4 (3): pp. 271-287, 1986.
- [33] W. Kornfeld and C. Hewitt. The Scientific Community Metaphor. *IEEE Transactions on Systems, Man, and Cybernetics* 11(1), pp. 24–33, 1981.
- [34] Tim, F; F. Richard Fritzon, M. Don and M. Robin. KQML as an agent communication language. *Proceeding of the CIKM '94 Proceedings of the third international conference on Information and knowledge management*, 1994.
- [35] Maes, P. Artificial life meets entertainment: life like autonomous agents. *Commun ACM* 38 (11), p. 0, 1995.
- [36] Bai, X. and Imura, H. A comparative study of urban environment in East Asia: stage model of urban environmental evolution, *Int. Rev. Environ. Strateg.* 1 (1) 135–158, 2000.
- [37] Meyboom, Ronald H; Lindquist, Marie; Egberts, Antoine C. An ABC of Drug-Related Problems. *Drug Safety*: 22 (1), pp 415-423, 2000.
- [38] Gasmelseid, Tagelsir. A system innovation oriented integration of Management Information Systems in Urban Water Management", *Handbook of Research on Hydroinformatics: Technologies, Theories and Applications*, pp. 389-405, 2010.
- [39] Basili, V; R., Caldiera, G. and H. Rombach. Experience Factory. In *Marciniak, J. J., Encyclopedia of Software Engineering*, volume 1, pp. 469-476, 1994.

Bernstein operational matrix method for solving physiology problems

K. Maleknejad, E. Hashemizadeh, and M. Mohsenyazadeh

Department of Mathematics, Karaj Branch, Islamic Azad University, Karaj, Iran

Abstract—A new approach implementing Bernstein operational matrix method for the numerical solution of differential equations, that arises in various physiology problems like oxygen diffusion, distribution of heat source in human head, tumor growth and etc. Operational matrix of derivative for Bernstein's polynomials function are presented to reduce these nonlinear differential equations to a system of nonlinear algebraic equation. Computational results are provided to demonstrate the viability of the new method.

Keywords: Bernstein polynomial; Operational matrix of derivative; nonlinear differential equations.

AMS Subject Classification: 34A34; 92C30.

1. Introduction

Nonlinear differential equations are indispensable tools for modeling many physiology problems such as study of steady state oxygen-diffusion in a cell with Michaelis-Menten uptake kinetics [1], [2], spring mass system [3] and bending of beams [4]. These equations are also useful in study of the distribution of heat sources in the human head [5], [6] and tumor growth [7], [8], [9], [10], [11].

We consider a class of singular boundary value problem

$$y''(x) + \left(a + \frac{m}{x}\right)y' = f(x, y), \quad 0 \leq x \leq 1, \quad (1)$$

$$\alpha_1 y(0) + \beta_1 y'(0) = \gamma_1, \quad (2)$$

$$\alpha_2 y(1) + \beta_2 y'(1) = \gamma_2, \quad (3)$$

which arising in physiology. we assume that $f(x, y)$ is continuous, $\frac{\partial f}{\partial x}$ exists and is continuous and also $\frac{\partial f}{\partial x} \geq 0$, $0 \leq x \leq 1$. Existence-uniqueness results for such problems have been established by several researchers [12]–[14].

It is a well-known fact that the solution of singularly boundary-value problem exhibits a multiscale character. That, there is a thin layer where the solution varies rapidly, while away from the layer the solution behaves regularly and varies slowly. This class of problems has recently gained importance in the literature for two main reasons. Firstly, they occur frequently in many areas of science and engineering, for example, combustion, chemical reactor theory, nuclear engineering, control theory, elasticity, fluid mechanics etc. Secondly, the occurrence of sharp boundary-layers as ε , the coefficient of highest derivative, approaches

zero creates difficulty for most standard numerical schemes, see for example [15], [16], [17], [18].

Bernstein polynomials play a prominent role in various areas of mathematics. These polynomials have been frequently used in the solution of integral equations, differential equations and approximation theory; see e.g., [19]–[23]. In recent years the various operational matrices of the polynomials have been developed to cover the numerical solution of differential, integral and integro-differential equations. In [24] the operational matrices of Bernstein polynomials are introduced.

In this paper we used Bernstein operational matrix of derivative for numerical solution of physiology problems. The advantage of Bernstein operational matrices method to other existing methods is its simplicity of implementation besides some other advantages.

This paper is organized as follows: In Section 2, we introduce Bernstein polynomials and their properties also we showed the operational matrix of derivative for Bernstein polynomials. In Section 3, the Bernstein polynomial approximation and its operational matrix of derivative together with collocation method are used to reduce the nonlinear singular ordinary differential equation to a nonlinear algebraic equation that can be solved by Newton's method. Section 4 illustrates some applied models to show the convergence, accuracy and advantage of the proposed method and compares it with some other existed method. Finally Section 5 concludes the paper.

2. Basic Definition

2.1 Definition of Bernstein polynomials basis

The Bernstein basis polynomial of degree n are defined by [24]

$$B_{i,n}(x) = \binom{n}{i} x^i (1-x)^{n-i}, \quad (4)$$

By using binomial expansion of $(1-x)^{n-i}$, we have

$$\binom{n}{i} x^i (1-x)^{n-i} = \sum_{k=0}^{n-i} (-1)^k \binom{n}{i} \binom{n-i}{k} x^{i+k}. \quad (5)$$

Now, we define

$$\Phi(x) = [B_{0,n}(x), B_{1,n}(x), \dots, B_{n,n}(x)]^T, \quad (6)$$

where we can have

$$\Phi(x) = A\Delta_n(x), \tag{7}$$

that A is an $(n + 1) \times (n + 1)$ upper triangular matrix with rows

$$A_{i+1} = \left[\overbrace{0, 0, \dots, 0}^{i \text{ times}}, (-1)^0 \binom{n}{i} \binom{n-i}{0}, (-1)^1 \binom{n}{i} \binom{n-i}{1}, \dots, (-1)^{m-i} \binom{n}{i} \binom{n-i}{n-i} \right], \tag{8}$$

and $\Delta_n(x)$ is an $(n + 1) \times 1$ matrix as follows

$$\Delta_n(x) = \begin{bmatrix} 1 \\ x \\ \vdots \\ x^n \end{bmatrix}.$$

2.2 Function approximation

A function $f(x)$, square integrable in $(0, 1)$, may be expressed in terms of Bernstein basis [24]. In practice, only the first $(n + 1)$ -terms Bernstein polynomials are considered. Hence if we write

$$f(x) \simeq \sum_{i=0}^n c_i B_{i,n}(x) = c^T \Phi(x), \tag{9}$$

where

$$c^T = [c_0, c_1, \dots, c_n], \tag{10}$$

then

$$c = \mathbf{Q}^{-1}(f, \Phi(x)), \tag{11}$$

where \mathbf{Q} is an $(n + 1) \times (n + 1)$ matrix and is said dual matrix of $\Phi(x)$ [24]

$$\begin{aligned} \mathbf{Q} &= (\Phi(x), \Phi(x)) = \int_0^1 \Phi(x)\Phi(x)^T dx \\ &= \int_0^1 (A\Delta_n(x))(A\Delta_n(x))^T dx \end{aligned} \tag{12}$$

$$= A \left[\int_0^1 \Delta_n(x)\Delta_n^T(x) dx \right] A^T = AHA^T,$$

A is defined in (8) and H is a Hilbert matrix

$$H = \begin{bmatrix} 1 & \frac{1}{2} & \dots & \frac{1}{n+1} \\ \frac{1}{2} & \frac{1}{3} & \dots & \frac{1}{n+2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n+1} & \frac{1}{n+2} & \dots & \frac{1}{2n+1} \end{bmatrix}. \tag{13}$$

The elements of the dual matrix \mathbf{Q} , are given explicitly by

$$\begin{aligned} \mathbf{Q}_{i+1,j+1} &= \int_0^1 B_{i,n}(x)B_{j,n}(x)dx \\ &= \binom{n}{i} \binom{n}{j} \int_0^1 (1-x)^{2n-(i+j)} x^{i+j} dx \\ &= \frac{\binom{n}{i} \binom{n}{j}}{(2n+1) \binom{2n}{i+j}}, \end{aligned} \tag{14}$$

where $i, j = 0, 1, \dots, n$.

2.3 Operational matrix of derivative

The differentiation of vector $\Phi(x)$ in Eq.(6) can be expressed as [24]

$$\Phi'(x) = \mathbf{D}\Phi(x) \tag{15}$$

where \mathbf{D} is the $(n + 1) \times (n + 1)$ operational matrix of derivatives for Bernstein polynomials. From (7) we have $\Phi(x) = A\Delta_n(x)$ and then

$$\Phi'(x) = A \begin{bmatrix} 0 \\ 1 \\ 2x \\ \vdots \\ nx^{n-1} \end{bmatrix}. \tag{16}$$

Defining $(n + 1) \times (n)$ matrix V and vector Δ_n^* as

$$V = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n \end{bmatrix}, \quad \Delta_n^* = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^{n-1} \end{bmatrix}, \tag{17}$$

equation (16) may then be restated as

$$\Phi'(x) = AV\Delta_n^*. \tag{18}$$

We now expand vector Δ_n^* in terms of $\Phi(x)$. By using [25], we get $\Delta_n^* = B^*\Phi(x)$ where

$$B^* = \begin{bmatrix} A_{[1]}^{-1} \\ A_{[2]}^{-1} \\ A_{[3]}^{-1} \\ \vdots \\ A_{[n]}^{-1} \end{bmatrix}, \tag{19}$$

so

$$\Phi'(x) = AVB^*\Phi(x), \tag{20}$$

therefor we have the operational matrix of derivative as

$$\mathbf{D} = AVB^*. \tag{21}$$

If we approximate $u(x) \simeq U^T \Phi(x)$, then for $n \geq 2$ (n is the order of derivatives), we get

$$u^{(n)}(x) \simeq U^T \Phi^{(n)}(x) = U^T \mathbf{D}^n \Phi(x). \tag{22}$$

3. Implementation of Bernstein method on physiology problems

In this section we solve nonlinear singular boundary value problem of the form Eq.(1) with the mixed conditions (2) and (3) by using Bernstein's operational matrix method.

From Eq. (9) we can approximate our unknown as

$$y(x) \simeq c^T \Phi(x), \tag{23}$$

Table 1: Approximate solutions for Example 1.

x	Present method with $n = 14$	Method in [26] with $m = 15$	Method in [27] with $n = 20$	Method in [28] with $n = 60$
0.0	0.82848329035981	0.82848329035968	0.82848329481355	0.82848327295802
0.1	0.82970609243393	0.82970609243380	0.82970609688790	0.82970607521884
0.2	0.83337473359113	0.83337473359100	0.83337473804308	0.83337471691089
0.3	0.83948991395383	0.83948991395370	0.83948991833986	0.83948989814383
0.4	0.84805278499619	0.84805278499606	0.84805278876051	0.84805277036165
0.5	0.85906492716936	0.85906492716923	0.85906492753032	0.85906491397434
0.6	0.87252831995841	0.87252831995828	0.87252831569855	0.87252830841853
0.7	0.88844530562332	0.88844530562319	0.88844529949702	0.88844529589927
0.8	0.90681854806693	0.90681854806680	0.90681854179965	0.90681854026297
0.9	0.92765098836571	0.92765098836558	0.92765098305256	0.92765098252660
1.0	0.95094579849659	0.95094579849648	0.95094579480523	0.95094579461056

where $\Phi(x)$ and c are defined in Eqs.(6) and (10). By using Eq. (22) we have

$$y'(x) = c^T \Phi'(x) = c^T \mathbf{D}^1 \Phi(x), \quad (24)$$

and

$$y''(x) = c^T \Phi''(x) = c^T \mathbf{D}^2 \Phi(x). \quad (25)$$

By substituting Eqs.(23), (24) and (25) in Eq. (1) we have

$$c^T \mathbf{D}^2 \Phi(x) + (a + \frac{m}{x}) c^T \mathbf{D} \Phi(x) = f(x, c^T \Phi(x)). \quad (26)$$

Also by using Eqs.(2), (3), (23) and (24) we have

$$\alpha_1 c^T \Phi(0) + \beta_1 c^T \mathbf{D} \Phi(0) = \gamma_1, \quad (27)$$

$$\alpha_2 c^T \Phi(1) + \beta_2 c^T \mathbf{D} \Phi(1) = \gamma_2. \quad (28)$$

Eqs.(27) and (28) give 2 linear equations. Since the total unknowns for vector c in Eq.(23) is $(n + 1)$, we collocate Eq.(26) in $(n - 1)$ Newton-Cotes points in the interval $[0, 1]$ as

$$x_p = \frac{2p - 1}{2(n + 1)}, \quad p = 1, 2, \dots, n - 1, \quad (29)$$

then we will have

$$c^T \mathbf{D}^2 \Phi(x_i) + (a + \frac{m}{x_i}) c^T \mathbf{D} \Phi(x) = f(x_i, c^T \Phi(x_i)), \quad (30)$$

for $i = 1, \dots, n - 1$. Now the resulting Eqs. (27), (28) and (30) generate a system of $(n + 1)$ nonlinear equations which can be solved using Newton's iterative method. We used the Mathematica 8 software to solve this nonlinear system.

4. Some applied models in physiology

To illustrate the effectiveness of the proposed method in the present paper, we implement it on two nonlinear singular boundary problems that arise in real physiology applications. Our results are compared with result in Refs. [26]–[30].

4.1 Example 1

Consider the following oxygen diffusion problem

$$y''(x) + \frac{2}{x} y'(x) = \frac{0.76129y}{y + 0.03119},$$

with boundary conditions:

$$y'(0) = 0, \quad 5y(1) + y'(1) = 5.$$

Table 1 shows the numerical results for various number of meshes, and present method solutions are compared with results in Refs. [26], [27] and [28].

4.2 Example 2

Consider this problem that is coincide by heat conduction model of the human head,

$$y''(x) + \frac{2}{x} y'(x) = -e^{-y},$$

we consider the solution of this problem with conditions as follows:

$$y'(0) = 0, \quad y(1) + y'(1) = 0.$$

Table 2 illustrates results for this example by proposed method alongside numerical solutions for this example that have been given in Refs. [26], [29] and [30].

5. Conclusions

This work present a numerical approach for solving a class of singular boundary value problems arising in physiology by the operational matrix of Bernstein polynomials. The operational matrix of derivative \mathbf{D} beside collocation method were used to transform the singular boundary value problems to a nonlinear system of algebraic equations that can be solved by Newton's method. This method is very simple and attractive. The implementation of current approach in analogy to existed methods is more convenient. The numerical examples that have been presented in the paper and the compared results support our claim.

Table 2: Approximate solutions for Example 2.

x	Present method with $n = 14$	Method in [26] with $m = 15$	Method in [29] with forth-order	Method in [30]
0.0	0.3675168151	0.3675168151	0.3675181074	0.3675169710
0.1	0.3663623292	0.3663623292	0.3663637561	0.3663623697
0.2	0.3628940661	0.3628940661	0.3628959378	0.3628941066
0.3	0.3570975457	0.3570975457	0.3570991429	0.3570975842
0.4	0.3489484206	0.3489484206	0.3489499903	0.3489484612
0.5	0.3384121487	0.3384121487	0.3384136581	0.3384121893
0.6	0.3254435224	0.3254435224	0.3254450019	0.3254435631
0.7	0.3099860402	0.3099860402	0.3099878567	0.3099860810
0.8	0.2919711030	0.2919711030	0.2919789654	0.2919711440
0.9	0.2713170101	0.2713170101	0.2713185637	0.2713170512
1.0	0.2479277233	0.2479277233	0.2479292837	0.2479277646

References

- [1] Lin, H. S. (1976). Oxygen diffusion in a spherical cell with nonlinear oxygen uptake Kinetics, *J. Theor. Biol.*, 60, 449–457.
- [2] McElwain, D. L. S. (1978). A re-examination of oxygen diffusion in a spherical cell with Michaelis-Menten oxygen uptake Kinetics, *J. Theor. Biol.*, 71, 255–263.
- [3] S.S. Ganji, A. Barari, D.D. Ganji, Approximate analysis of two-mass-spring systems and buckling of a column, *Computers and Mathematics with Applications*, Volume 61, Issue 4, February 2011.
- [4] J.B. Paiva, A.V. Mendonça, A coupled boundary element differential equation method formulation for plate-beam interaction analysis, *Engineering Analysis with Boundary Elements*, Volume 34, Issue 5, May 2010.
- [5] Flesch, U. (1975). The Distribution of heat sources in the human head: A theoretical consideration, *J. Theor. Biol.*, 54, 285–287.
- [6] Gray, B. F. (1980). The Distribution of heat sources in the human head: A theoretical consideration, *J. Theor. Biol.*, 82, 473–476.
- [7] A. D. Conger, M. C. Ziskin, Growth of mammalian multicellular tumor spheroids, *Cancer Res.* 43(1983), 556–580.
- [8] J.A. Adam, A simplified mathematical model of tumor growth, *Math. Biosci.* 81 (1986) 224–229.
- [9] J.A. Adam, A mathematical model of tumor growth II: effect of geometry and spatial non-uniformity on stability, *Math. Biosci.* 86 (1987) 183–211.
- [10] J.A. Adam, S.A. Maggelakis, Mathematical model of tumor growth IV: effect of necrotic core, *Math. Biosci.* 97 (1989) 121–136.
- [11] A.C. Burton, Rate of growth of solid tumor as a problem of diffusion, *Growth* 30 (1966) 157–176.
- [12] R.K. Pandey, *On a class of weakly regular singular two point boundary value problems II*, *J. Differential Equations* 127 (1996) 110–123.
- [13] M.M. Chawla, P.N. Shivkumar, *On the existence of solution of a class of singular two-point nonlinear boundary value problems*, *J. Comput. Appl. Math.* 19 (1987) 379–388.
- [14] R.D. Russell, L.F. Shampine, *Numerical methods for singular boundary value problems*, *SIAM J. Numer. Anal.* 12 (1975) 13–36.
- [15] U.M. Ascher, R.M.M. Mattheij, R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [16] A.K. Aziz, *Numerical Solution of Two Point Boundary-Value Problem*, Blaisdal, New York, 1975.
- [17] J.E. Flaherty, R.E. O' Malley, The numerical solution of boundary-value-problems for stiff differential equations, *Math. Comput.* 31 (1977) 66–93.
- [18] M.K. Kadalbajoo, R.K. Bawa, Variable mesh difference scheme for singularly-perturbed boundary-value problems using splines, *J. Optim. Theory Appl.* 90 (2) (1996) 405–416.
- [19] K. Maleknejad, E. Hashemizadeh, R. Ezzati, A new approach to the numerical solution of Volterra integral equations by using Bernstein's approximation, *Commun. Nonlinear. Sci. Numer. Simulat.* 16 (2011) 647–655.
- [20] E.H. Doha, A.H. Bhrawy, M.A. Saker, Integrals of Bernstein polynomials: An application for the solution of high even-order differential equations, *Appl. Math. Lett.* 24 (2011) 559–565.
- [21] E.H. Doha, A.H. Bhrawy, M.A. Saker, On the derivatives of Bernstein polynomials: An application for the solution of high even-order differential equations, *Boundary Value Problems* Volume 2011, (2011) Article ID 829543, 16 pages doi:10.1155/2011/829543.
- [22] B.N. Mandal, S. Bhattacharya, Numerical solution of some classes of integral equations using Bernstein polynomials, *Appl. Math. Comput.* 190 (2007) 707–1716.
- [23] T.J. Rivlin, *An introduction to the approximation of functions*, New York: Dover Publications; 1969.
- [24] S.A. Yousefi, M. Behroozifar, Operational matrices of Bernstein polynomials and their applications, *Int. J. Syst. Sci.* 41 (6) (2010) 709–716.
- [25] K. Maleknejad, B. Basirat, E. Hashemizadeh. A Bernstein operational matrix approach for solving a system of high order linear Volterra-Fredholm integro-differential equations *Mathematical and Computer Modelling*, 55 (2012) 1363–1372.
- [26] K. Maleknejad, E. Hashemizadeh. Numerical solution of nonlinear singular ordinary differential equations arising in biology via operational matrix of shifted Legendre polynomials *American Journal of Computational and Applied Mathematics*, 1(1) (2011) 15–19.
- [27] S.A. Khuri, A. Sayfy, *A novel approach for the solution of a class of singular boundary value problems arising in physiology*, *J. Math. Comput. Model.* 52 (2010) 626–636.
- [28] H. Caglar, N. Caglar, M. Ozer, *B-spline solution of non-linear singular boundary value problems arising in physiology*, *Chaos Solitons Fractals* 39 (2009) 1232–1237.
- [29] J. Rashidinia, R. Mohammadi, R. Jalilian, *The numerical solution of non-linear singular boundary value problems arising in physiology*, *J. Appl. Math. Comput.* 185 (2007) 360–367.
- [30] R.K. Pandey, Arvind K. Singh, *On the convergence of a finite difference method for a class of singular boundary value problems arising in physiology*, *J. Comput. Appl. Math.* 166 (2004) 553–564.

A Method of Breast Cancer Screening Based on Fractal Analysis

R.I. Andrushkiw¹, D.A. Klyushin², D.G. Shervarly², E.N. Golubeva², and N.V. Boroday³

¹New Jersey Institute of Technology, Newark, NJ 07102, USA

²Kyiv National Taras Shevchenko University, Kyiv, Ukraine

³R.E.Kavetsky Institute of Experimental Pathology, Oncology and Radiobiology, Kyiv, Ukraine

Abstract - In this paper we propose a new method of screening for breast cancer, based on fractal analysis of time series. The series is constructed using scanned digital images of the nuclei of interphase cells of buccal epithelium along the Peano curve. To do this, we compute the Hurst coefficients of the time series and use standard descriptive statistics. Digital images of interphase nuclei of buccal epithelium are studied in patients with benign tumors, malignant tumors and in individuals that were practically healthy (without tumors).

Keywords: Breast Cancer, Screening, Hurst coefficient, Peano curve.

1 Introduction

The analysis of malignancy-associated changes (MAC) in cells distant from a tumor is one of perspective methods for the effective screening of cancer. Such methods can be divided in two groups: methods involving the analysis of MAC in non-tumor cells located near a tumor [1, 2] and methods involving the analysis of MAC in non-tumor cells located far from a tumor, in particular, in buccal epithelium (oral mucosa) [3, 4].

In a recent study, Redon et al[5] demonstrated by using special cohorts of mice with B16 melanoma, MO5076 sarcoma, and COLON26 carcinoma, that a tumor may induce malignancy-associated changes in tissues distant from the sites of the implanted tumors. Also, in 2009 Lieberman-Aiden et al. [6] have shown that DNA in the cell nucleus is packaged as a fractal globula, i.e. as a polymer analogue of a 3D Peano curve.

The above two papers inspired us to study MAC in buccal epithelium, taking into account the fractal nature of DNA packaging in the chromatin. The results of this study led us to propose a new method of screening for breast cancer.

2 Materials

We consider two groups of patients: G_1 – joined group of patients suffering from breast cancer (68 cases)

and patients suffering from fibroadenomatosis (33 cases) and G_2 —group of practically healthy women (29 cases). Smears from various depths of the spinous layer were obtained (conventionally they were denoted as median and deep), after gargling and removing the superficial cell layer of the buccal mucous. The DNA content stained by Feulgen was estimated using the Olympus computer analyzer, consisting of the Olympus BX microscope, Camedia C-5050 digital zoom camera and a computer. We investigated from 40 to 60 nuclei in each preparation. The DNA-fuchsine content in the nuclei of the epitheliocytes was defined as a blue component of a RGB-value.

3 Methods

When scanning a digital image, one of the main condition that must be satisfied is invariance relative to the rotation of a scanogram, since the orientation of a nucleus on a slide of the microscope may be random. To provide for this invariance, we used space-filling fractal curves Peano [7-8]. This allowed us to consider an image as a time series, but not as a matrix.

The digital images contain 160×160 pixels. Since the Peano curve covers a square with 3^n pixels on a side, we had to use random squares in the of scanogram (fig. 1-3).

Based on the hypothesis of the fractal nature of chromatin distribution, we used the Hurst coefficient $H=2-D$, where D is the fractal dimension. The Hurst coefficient is computed using the following algorithm [9].

1. Compute the deviation of time series values from the mean value during current period:

$$\delta_{m,N} = \sum_{i=1}^m (x_i - \bar{x}_N),$$

where N is the length of a period varying from 2 to the length of the whole time series, m is the upper limit of summation varying from 1 to $N-1$, x_i is a

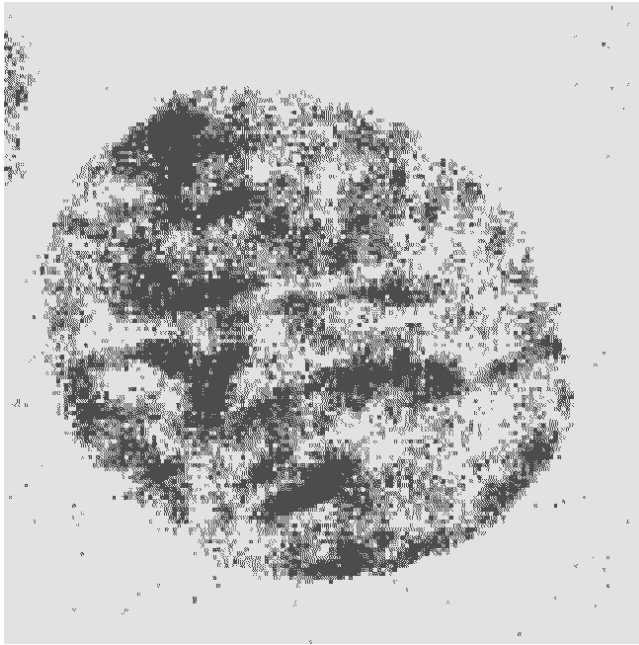


Fig. 1 Feulgen stained nuclei of cell in buccal epithelium

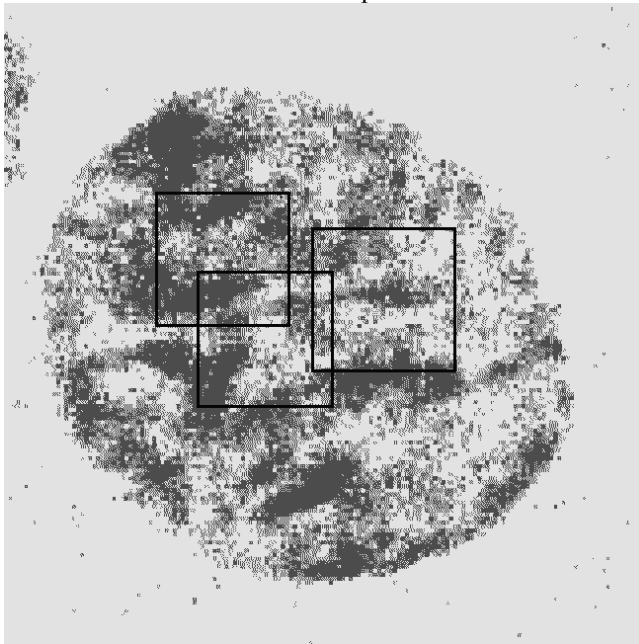


Fig. 2. Three random scanning fields

value of the time series and \bar{x}_N is the mean of the time series during the current period. Thus, we obtain $N - 1$ values $\delta_{2,N}, \dots, \delta_{N-1,N}$.

2. Compute the range of the deviation of the time series:

$$R = \max_{m=2, \dots, N} \delta_{m,N} - \min_{m=2, \dots, N} \delta_{m,N}$$

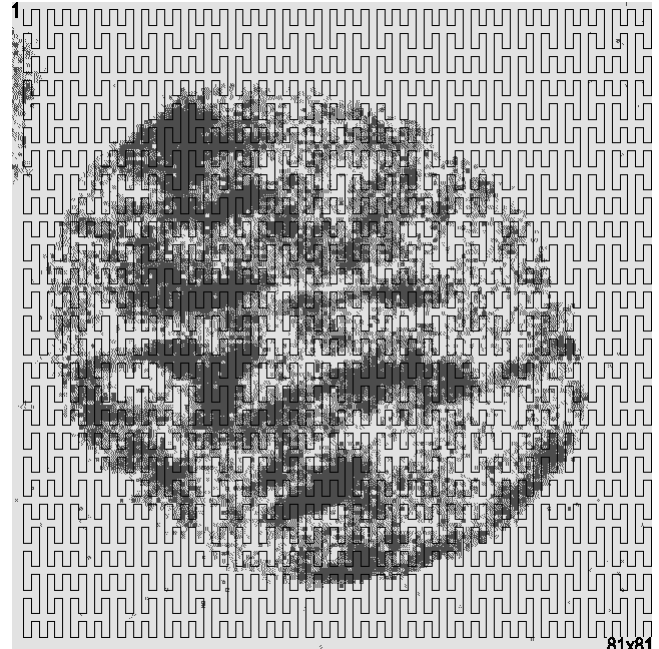


Fig. 3. Peano curve of 4th order ($3^4 \times 3^4$) covering part of scanogram

3. Normalize the range:

$$Q = \frac{R}{s},$$

where s is a standard deviation of the time series.

4. Take the logarithm of Q and N .

5. Compute $\lg Q$ and $\lg N$, and construct a linear approximation of the dependence of $\lg Q$ on $\lg N$.

6. Compute the Hurst coefficient, which is the tangent of the slope angle of the linear approximation of the dependence of $\lg Q$ on $\lg N$.

The Hurst coefficient characterizes the chaotic nature of the time series:

1. If $0 < H < 0.5$, then the time series is ergodic, i.e. if the time series increased during previous period then it is most probably that at the next moment the time series will decrease, and vice versa.
2. If $H = 0.5$, then the time series is chaotic, i.e. values of the time series do not affect to next values.
3. If $0.5 < H < 1.0$, then the time series has a stable trend, i.e. if the time series increased or decreased during the previous period, then it will be increasing or decreasing respectively during the next period.
4. If $H > 1$, then the time series is a random fractal time series, i.e. there are independent jumps of the amplitude during the time period and the time series is increasing.

For every patient we compute the following indexes:

1. The mean Hurst coefficient;
2. The maximal Hurst coefficient;
3. The minimal Hurst coefficients.

Then, by constructing a decision tree [10] we recognized 29 healthy cases, and 101 cases of breast cancer and fibroadenomatosis.

The above procedure describes completely our proposed method of breast cancer screening, which is based on fractal analysis of time series.

4 Conclusions

We discovered new malignancy-associated changes in chromatin of buccal epithelium in patients suffering from breast cancer and fibroadenomatosis: the time series constructed using Peano curves in buccal cells of healthy individuals are more chaotic, than in the buccal cells of individuals with breast cancer or fibroadenomatosis (the Hurst coefficients of healthy women are nearer to 0.5). The classification model based on 3 random Peano curves has the following characteristics:

Specificity = $27/29 \cdot 100\% = 93.1\%$,
 Sensitivity = $101/101 \cdot 100\% = 100\%$,
 Accuracy = $(27+101)/(29+101) \cdot 100\% = 98.46\%$.

5 References

- [1] Susnik, B., Worth, A., LeRiche, J. & Palcic, B. Malignancy-associated changes in the breast: changes in chromatin distribution in epithelial cells in normal-appearing tissue adjacent to carcinoma. *Analytical and Quantitative Cytology and Histology* 17(1): 62 – 68, 1995.
- [2] Mairinger, T., Mikuz, G. & Gschwendtner, A. Nuclear chromatin texture analysis of nonmalignant tissue can detect adjacent prostatic adenocarcinoma. *The Prostate* 41(1): 12 – 19, 1999.
- [3] Us-Krasovec M, Erzen J, Zganec M et al. Malignancy associated changes in epithelial cells of buccal mucosa: a potential cancer detection test. *Anal Quant Cytol Histol.* 27(5): 254-62, Oct. 2005.
- [4] Andrushkiw R.I., Boroday N.V., Klyushin D.A., Petunin Yu.A. Computer-aided cytogenetic method of cancer diagnosis. — New York: Nova Science Publishers, 2007.
- [5] Redon C.E. et al. Tumors induce complex DNA damage in distant proliferative tissues in vivo // *Proceedings of the National Academy of Sciences*, vol. 107 — no. 42. P. 17992-17997, Oct. 19, 2010
- [6] Lieberman-Aiden E., van Berkum N. L., et al. Comprehensive mapping of long-range interactions reveals folding principles of the human Genome. *Science* 326, 2009.
- [7] Nikolaou N. and Papamarkos N. Color image retrieval using a fractal signature extraction technique" *Engineering Applications of Artificial Intelligence*. Vol. 15., No. 1. P. 81-96, 2002.
- [8] Sagan H. Space-filling curves. — Springer-Verlag: New York–Berlin, 1994.
- [9] Butakov V., Grakovsky A. Evaluation of arbitrary time series stochastic level by Hurst parameter // *Computer Modelling and New Technologies*, Vol.9, No.2, P.27-32, 2005.
- [10] Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

SESSION

COMPUTER-BASED MEDICAL SYSTEMS, STATISTICAL METHODS, MODELLING, SIMULATION AND OPTIMIZATION OF BIOLOGICAL SYSTEMS

Chair(s)

TBA

An Agent-Based Modeling and Simulation Environment for Dynamic Biological Systems

Steven Phung¹, Rajdeep Singh², Jamie Lawson², Michael Hultner², Briam Peck², Ross Henderson³, Dennis Hsu², Desmond Lun^{4,5}, Vijayaraj Nagarajan³, Mariam Quiñones³, Darrell Hurt³, Yentram Huyen³, Mike Tartakovsky³

¹Autodesk, Inc., The Landmark @One Market, San Francisco, California 94105

²Lockheed Martin, Information Systems and Global Solutions, 4770 Eastgate Mall, San Diego, California 92121, USA

³Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, NIH Bethesda, Maryland 20817, USA

⁴Center for Computational and Integrative Biology and Department of Computer Science, Rutgers University, Camden, NJ 08102, USA

⁵Phenomics and Bioinformatics Research Centre and School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA 5095, Australia

ABSTRACT

Motivation: Understanding emergent behaviors of complex biological systems requires modeling and simulation of large and detailed models. Models must be both expressive and scalable to capture the size and complexity of molecular and cellular networks.

Results:

In this report we present GRANITE (Genetic Regulatory Analysis of Networks Investigational Tools Environment), an agent-based modeling (ABM) and multi-agent simulation (MAS) approach to modeling large, complex, and dynamic systems. We have demonstrated the GRANITE capability on metabolic networks: specifically the mycolic acid biosynthesis pathway of the *Mycobacterium tuberculosis*. The agent-based model has been compared to Flux Balance Analysis (FBA) and shown to be able to emulate the internal and external properties of the system as modeled by FBA. We show that the approach is scalable and computationally efficient to allow researcher interaction with a dynamically evolving simulation. The GRANITE tool enables the researcher to propose and test systems-level hypotheses and make predictions for laboratory experiments to validate or refute these hypotheses.

Availability and Implementation:

The GRANITE software is open-source and available from the corresponding author, Ross Henderson. Please indicate GRANITE in the subject line of correspondence.

Contact: rh@nih.gov*

1 INTRODUCTION

Living systems are complex systems. As such, they have emergent behaviors: input-response properties that can be observed but not

predicted by first order knowledge of the functions of the system's components. Systems biology is an approach to understand the general principles of living systems by elucidating the relationships between the components of a system and its emergent behaviors.

Only through understanding living things as systems can one hope to understand the mechanisms of cellular and molecular biology. These systems are formed from the many interactions between molecules within the cell and between cells. Examples include metabolic networks, signal transduction networks, gene regulatory networks, and other epigenetic networks. The interplay between these systems creates another level of complexity that makes the modeling and simulation of living systems a serious computational challenge.

Much of the focus of effort in systems biology involves the development of models for biological function at the systems level. To be useful these models must be expressive, computationally tractable, and should yield predictions that can be tested with laboratory experiments. Our initial gap analysis indicated the need for an interactive M&S (Modeling & Simulation) tool that allows for real-time interaction with the simulation. Since Cytoscape (<http://www.cytoscape.org/>) has limited ability to allow real-time dynamic interaction, we identified two other tools that study dynamics of biological networks and evaluate perturbation hypotheses. FERN (Erhard, *et al*, 2008) allows visualization of the dynamics but it does not allow for real-time interaction with the simulation. Even then, our attempts to integrate GRANITE with FERN proved cumbersome due to limitations of the available interfaces. Perturbation Analyzer tool (Fei Li, *et al*, 2009) was developed to investigate specifically the effects of single or combinatorial concentration perturbations by comparing two different steady states using law of mass action (LMA) in real-time and uses Cytoscape

*To whom correspondence should be addressed.

for visualization. In this report we present a modeling and simulation approach, GRANITE, that is expressive enough to capture any kind of interaction network, can modularly use any kinetic model, is computationally tractable and scalable, and allows researchers to interact and dynamically perturb the system at different hierarchical levels to learn its rules for emergent behavior.

2 METHODS

The Genetic Regulatory Analysis of Networks Investigational Tools Environment (GRANITE) software consists of:

- A simulation environment where software agents can be organized into dynamic models,
- A domain specific language (DSL) for expressing biological function, and
- A graphical user interface (GUI) for dynamic interaction with the simulation.

The agent based modeling and simulation components, and the DSL are implemented in Scala and the GUI is implemented in Java. The GRANITE software is available upon request from the corresponding author.

2.1 Agent-based Modeling (ABM) and Multi-agent Simulation (MAS)

The evolution of assemblies of biological components is often modeled as a system of ordinary differential equations (ODEs) that can be solved using numerical methods. Alternatively, the ABM approach (Eric, 2002) creates an assembly of computational components (agents) that would be governed by the same system of ODEs, but instead of explicitly solving the ODEs using classical numerical analysis, we simply allow the computational components to evolve directly in a MAS environment. This in effect solves the ODEs approximately in a distributed manner. The process of creating and running a system model for an experiment in the GRANITE context is as follows: A metabolic network model for the mycolic acid biosynthesis pathway (MAP) is instantiated from an SBML (Systems Biology Markup Language) model (Raman, *et al.*, 2005). A set of reaction agents and their associated metabolites are created by parsing the model, instantiating the agents, the environments, and populating the environment with metabolites. Simple Michaelis-Menten kinetics are used to model the agent reaction kinetics; GRANITE facilitates the use of other kinetic models by providing a generic agent-environment interaction interface. Similarly, the non-agent entities (e.g. metabolites and enzymes) are added to the environment and given initial conditions. The agents are then placed in the simulation framework with a set of parameters. A simulation scheduler strategy (deterministic or stochastic) is chosen and the progress of the simulation is meas-

ured in interaction time. For non-interactive simulations the simulation is allowed to evolve until it reaches a steady state. For interactive simulations the simulation evolves under control of the researcher via the Glimpse-GRANITE GUI.

2.2 Scalability

Agents interact with one another only indirectly using the environment as a mediator. This decoupled approach leads to modularity and scalability. The approach is modular because it uses agents to encapsulate biological function, and scalable because it avoids combinatorial interactions and therefore results in an efficient simulation. The performance of the GRANITE system has been benchmarked using the MAP network. We have shown that the computation scales linearly with the number of reaction agents. In the MAS framework, we employed a scheduling capability that provides control of the computational demands by modulating the simulation fidelity. A GRANITE simulation is configured to employ a deterministic or a stochastic scheduler. The deterministic scheduler evolves the simulation using all agent-environment interactions at all times based on the kinetic models of the agents; i.e., their strategies for turning reactants into products. However, an agent's interaction with the environment may not always lead to significant changes in the environment. For example, at very low substrate concentrations, the continuity assumption for the rate law does not hold and the reaction may not be moving forward at all times. Stochastic scheduling exploits this constraint and allows agents (reactions) to interact only if their interaction is significant; see Fig. 1. The uncoupled agents and the scheduling of their interactions with the environment produce simulations that scale linearly with the agent population size.

Let c be the continuity threshold and let the maximum saturation rate of a reaction agent, 'i', be given M_i . Let the relative rate of the reaction, r , at any time t , $r_{i,t} = v_{i,t}/M_i$. The mixed strategy used by the agent 'i' for deciding on whether to interact or not at any given time is as follows:

- If $r_{i,t} > c$, the agent can interact with the environment at time t . Let the set of all agents in this category be denoted as A_I .
- Else, let $r_{i,t} = r_{i,t} / \sum r_{i,t}$ define a normalized distribution, \underline{r} . We then choose a user defined percentage of the agents from the set A_I using roulette wheel selection based on \underline{r} .

An important measure of scalability of MAS is the time it takes to evolve to some steady state: the settling time. Factors contributing to settling time include the number of agents participating in the simulation and the fidelity of the underlying kinetic model of each agent (fidelity impacts how closely the computational components can evolve to the solution prescribed by the ODEs).

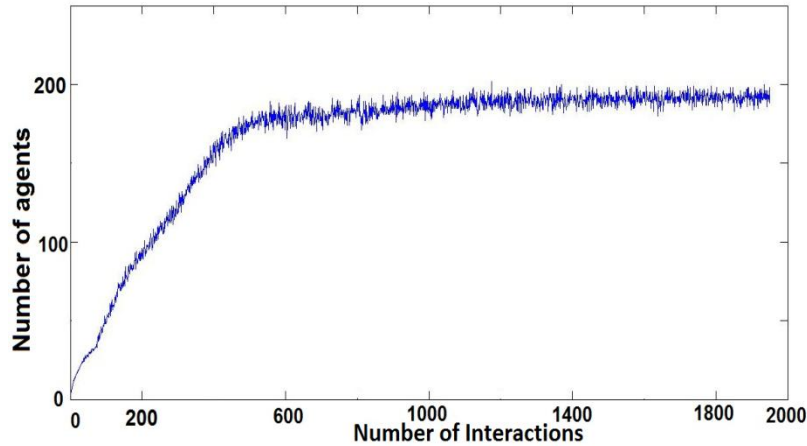


Fig. 1. Agent Population Size vs. Simulation Time. Stochastic Scheduler needs on average 75% agent-environment interactions (166 agents interact on average) and at most 190 agents, 87%, interact at any given time compared to Deterministic Scheduler (all 219 agents interact all the time).

The simulation framework manages this complexity by determining, at each time step, a trust region for each agent in which the agent can make a reliable contribution to environment evolution. The validity of this trust region is determined by the interactions of all the agents with the environment, driven by the scheduler. Figure 2 shows results from empirical experiments, demonstrating that a relatively coarse model of the trust region is sufficient to avoid very large settling times. We use the steady state flux to compare GRANITE to FBA where the GRANITE flux plots are scaled to compare with FBA on a gene by gene basis. The scaling approach we used is very straightforward and intuitive. We chose to group all reactions associated with each gene ' i ' (Raman et al, 2005). Let

f_j be the FBA flux and g_j be the GRANITE flux for reaction j , and S_i be the set of reactions influenced by gene ' i '. The affine scaling for all reactions in S_i is then computed as follows:

$$k = \frac{\max_{j \in S_i} [f_j] - \min_{j \in S_i} [f_j]}{\max_{j \in S_i} [g_j] - \min_{j \in S_i} [g_j]}$$

$$o = \max_{j \in S_i} [f_j] - k * \max_{j \in S_i} [g_j]$$

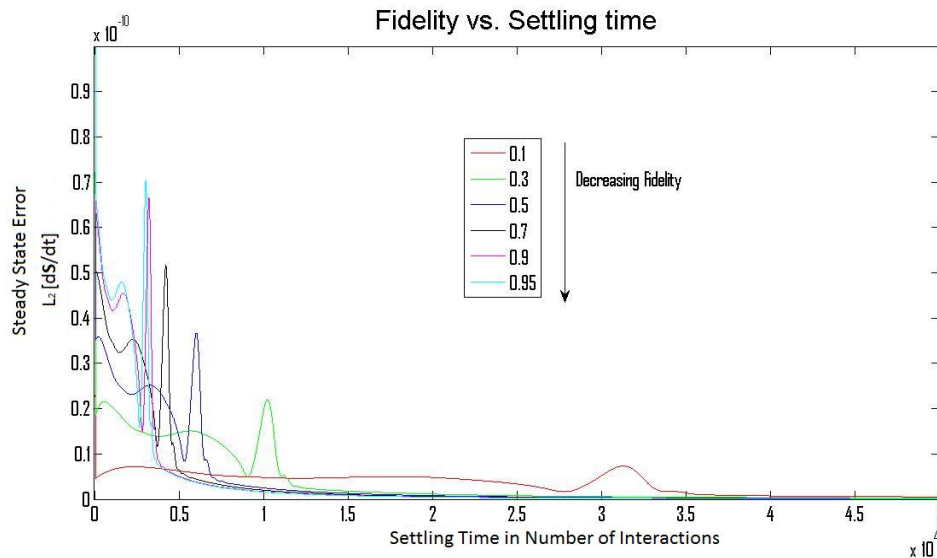


Fig. 2. Fidelity vs. Settling Time. Higher fidelity, trust parameter =0.1, moves the simulation slower to the steady state but the interactions are more accurate (the first order linear approximator defines the trust region using more support points in the same interval). Coarsest fidelity, trust parameter =0.99, corresponds to only 2 support points (the end points of the closed interval) and advances the simulation faster although with less accuracy.

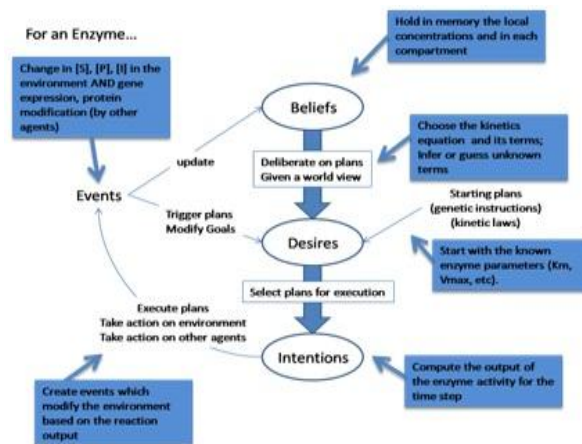


Fig. 3. The Belief-Desires-Intentions strategy is used to control the behavior of agents in a multi-agent simulation. For an agent that represents an enzymatic reaction, the Beliefs are the inputs to the reaction available from the environment and the entity properties; Desires are specified in one or more kinetic models for the reaction, and Intentions are the actions made by the agent onto the environment at each update step.

All of the GRANITE reaction fluxes associated with a gene are then scaled with the scale found for that gene. The correlation between the flux profiles improves significantly with this scaling. Note that one can apply a feedback loop by incorporating these scales into the catalyst concentration values to drive the GRANITE simulation.

3 RESULTS AND DISCUSSION

In this report we present a software framework for ABM of biological entities and a MAS environment for simulation of biological systems. This framework provides a means to create complex models of molecular networks that can evolve in an interactive simulation environment.

3.1 Agency

We employed classic ABM (Axelrod, 1997) to express units of biological function. Using the Belief-Desires-Intentions (BDI) model (Rao, 1995; and Weiss, 2000), as shown in Fig. 3, we created a framework for expressing biological agents that can be composed into complex systems. We discovered that this pattern works very well when agents represent biological function and specific entities. For example, an enzyme is represented by an agent that models its enzymatic reaction. Beliefs in the BDI model represent the world-view of the agent: the inputs to the agent from the environment, such as the state of mutable properties of substrates, enzymes, inhibitors, and other effectors. Desires represent the agent's goals such as the conversion of substrates to products, governed by stoichiometry and kinetic models for a reaction. Intentions are the actual steps an agent takes to affect its desires on the environment; the rate model for a reaction, for instance. The environment is then an arena where different agents compete through their intentions to achieve their desires.

This approach to modeling units of biological function is both expressive and modular. There are no limits placed on the techniques for expressing a functional response to environmental con-

ditions. Thus, alternative assumptions and models can be incorporated into the simulation and tested.

3.2 Simulation

We employed a multi-agent simulation with scheduling strategies to create a computationally tractable and scalable modeling and simulation capability. Agents compete with one another to achieve their goals in one or more environments. The simulation framework's job is therefore to manage the changes to the environment(s) resulting from agent activities scheduled in the system.

3.3 Domain Specific Language (DSL) for Dynamic Biological Systems

The feature that ties modeling and simulation together is a novel Domain Specific Language that enables the systems biologist to express agents and simulation context in a simple and concise form that they can relate to. Where SBML can express state, GRANITE DSL can express state, coordination, and activity. As such, the DSL can describe all of the dynamics of the system, i.e. the system's overall behavior with respect to time. The DSL is an extension of the popular Scala programming language, which is designed for domain specific extensions, and benefits from all of the tools and documentation developed in the Scala community. In addition to GRANITE's ability to use SBML models as inputs, the DSL facilitates creation of biological models in a more natural yet formal way which is biologist friendly. Consider units of measure as an example. Above, we stated that the GRANITE user can supply their own kinetic models. In fact, different reaction agents may employ different kinetic models as appropriate for the reaction. In order to maintain consistency among the different kinetic models, their units must be compatible. This is a hard bookkeeping problem, made harder when different models are developed by different people in different organizations. The GRANITE DSL provides "guardrails" for the user by supplying syntax for defining the units

of measure for the rate constants, or for any other values. The GRANITE system also supplies implicit conversions so that if one model assumes concentrations in moles/liter and another model assumes millimoles/liter, the GRANITE system will automatically make the appropriate conversions. When incompatible or unknown units are combined, GRANITE alerts the user rather than producing meaningless results. Adding two values in units of molarity produces an error because concentrations are not addable. Corresponding volumes are needed for that operation to make sense, and so the GRANITE DSL prevents it.

We discuss some simple steps to illustrate the use of DSL in the context of a metabolic network. The first step defines a meme called "a". Memes are first class modeling objects that have mutable and immutable properties. An immutable property, like molecular weight, always has the same value. A mutable property, like concentration, may vary at different times and in different environments.

```
val a = Species called "a" build
```

The second step defines a simulation; interaction models that will be bound to an environment using a simulation context are created. In the example below, the interaction model is a metabolic network containing two reactions. Reaction *r1* produces *b* and consumes *a*, whereas reaction *r2* produces *c* and consumes *b* using given stoichiometric coefficients, kinetic laws, rate parameters, and a deterministic scheduler (in this example) to decouple and synchronize the agent interactions.

```
def metabolicNetwork = CreateMetabolicNetwork of (
  Reaction called "r1" of (1*a) -> (1*b)
    using (MichaelisMenten withSpecificityConstants(a->0.1)
      catalyzedBy(p) withCatalyticConstant(0.1)),
  Reaction called "r2" of (1*b) -> (1*c)
    using (MichaelisMenten withSpecificityConstants(b->0.1)
      catalyzedBy(p) withCatalyticConstant(0.1))
) scheduledBy DeterministicMetabolicModelScheduler(0.01)
```

The specificity constant and the catalyst constant are typically denoted in the Michaelis-Menten kinetics as K_m and K_o respectively, and may be referenced as such in the DSL.

The third step creates the simulation contexts; this involves the creation of environments and the assignment of interaction models affecting those environments. The environment is defined using a *containing* clause which specifies memes and associated properties. Specifying which interaction models to use is accomplished by a *using* clause. Below is an example of defining a simulation context where memes a, b, c, and p are associated with concentration properties which use a metabolic network interaction model.

```
val sc1 = SimulationContext called "sc1" containing (
  a where ConcentrationIs(1000.0),
```

```
  b where ConcentrationIs(0.0),
  c where ConcentrationIs(0.0),
  p where ConcentrationIs(1.0)
) using metabolicNetwork
```

Finally, a simulation is constructed by defining which simulation contexts are part of the simulation. Below is an example of defining a simulation. The conciseness reflects the power of the DSL.

```
Simulation of sc1
```

3.4 Validation

As the use of agents is a departure from traditional methods of modeling biological systems, we performed a set of experiments designed to validate the approach. Our basis for validation criteria was the ability to emulate results from established systems, as well as from accepted modeling or simulation methods. For this study we chose to model a metabolic network, the mycolic acid biosynthesis pathway of the *Mycobacterium tuberculosis* which involves 197 metabolites, 219 reactions, and 28 enzymes driving these reactions. The pathway has been defined (Barry, 1998) and models exist in the SBML format (Raman, *et al.*, 2005). Furthermore, systems-level analyses exist in the literature that provide metrics about the internal states of the system against which we can compare the states of the agents and the environment. We chose to compare the ABM-MAS results to a Flux Balance Analysis of the mycolic acid pathway using the same SBML model of MAP as used by Raman *et al.* Instantiating the MAP model into a set of reaction agents with the same stoichiometric parameters, we attempted to emulate the internal and external states of the metabolic pathway at steady-state using Michaelis-Menten kinetics. We examined the ability of the ABM-MAS system to emulate the output of the pathway in terms of the observed proportions of mycolic acids and the flux profiles of the reactions in the network. Initial results showed that we could either emulate the mycolate ratios or the flux profile (Table 1). Using group scaling based on gene-reaction associations, and an optimized set of parameters, the ABM-MAS system was able to reproduce both the observed mycolic acid proportions and the reported flux profiles (Fig. 4). The method for optimizing and discovering the system parameters involves a novel use of genetic algorithms (Lawson, Singh, *et al.*) that will be published separately.

3.5 Dynamic Systems

Agent systems are particularly useful in modeling dynamic systems in a manner that allows the modeler to directly interact with the evolving system. The modeler can make changes to the system and immediately observe the response in real-time. We assert this is a novel technique for proposing and testing hypotheses at the systems level of molecular biology.

Table 1. A comparison of the observed and simulated mycolate ratios.

	methoxy-mycolate to alpha-mycolate	keto-mycolate to alpha-mycolate	trans to cis forms of methoxy-mycolate and keto-mycolate
Observed	0.54	0.49	0.14
Randomly Chosen Parameters	0.49	0.47	1.0
Manually Chosen Parameters	0.36	0.28	0.15
Learned Parameters	0.54	0.49	0.14

The first row presents the published output of the mycolic acid pathway (Watanabe, 2001). Initial experiments with randomly chosen parameters were able to approximate the ratios except for the cis:trans bias (row 2). Altering the specific activity levels of the methylases MmaA1 and MmaA4 did improve the cis:trans bias but reduced the fidelity of the other mycolates (row3). Learned parameters using a genetic algorithm approach (Lawson, Singh, *et al.*) were ultimately used to create a model that produced the desired mycolate ratios (row 4).

The Glimpse-GRANITE tool was developed to provide that capability to systems biologists. This GUI, see Fig. 5, provides a command-and-control interface to the simulation that includes the ability to create an agent system, start a simulation, observe the internals and externals of the simulation environment, and configure all aspects of the system and the simulation. The temporal aspect of the simulation is measured in number of interactions within the system. Since the agents are uncoupled from the simulation interactions, a change to an agent is immediately reflected in the simulation – no re-compilation or re-start is necessary. Furthermore, the state of the simulation can be check-pointed, or saved, such that if perturbations of the system destroy the integrity of the steady-state

model, the simulation can be brought back to a stable state and new perturbations can be tested.

3.6 Predictive Power of ABM-MAS

In addition to the expressivity, scalability, and evolutionary properties of the ABM-MAS method, it also has the capability of making and testing predictions. Outcomes of the multi-agent simulations are not determined by a global objective or control function. Thus, the system, as a function of initial conditions, will evolve dynamically into a steady-state, an oscillating state, or possibly degenerate into a chaotic state that is not sustainable. The observable features of the system state(s) are important components in measuring the predictive power of the model. If a change to the system model

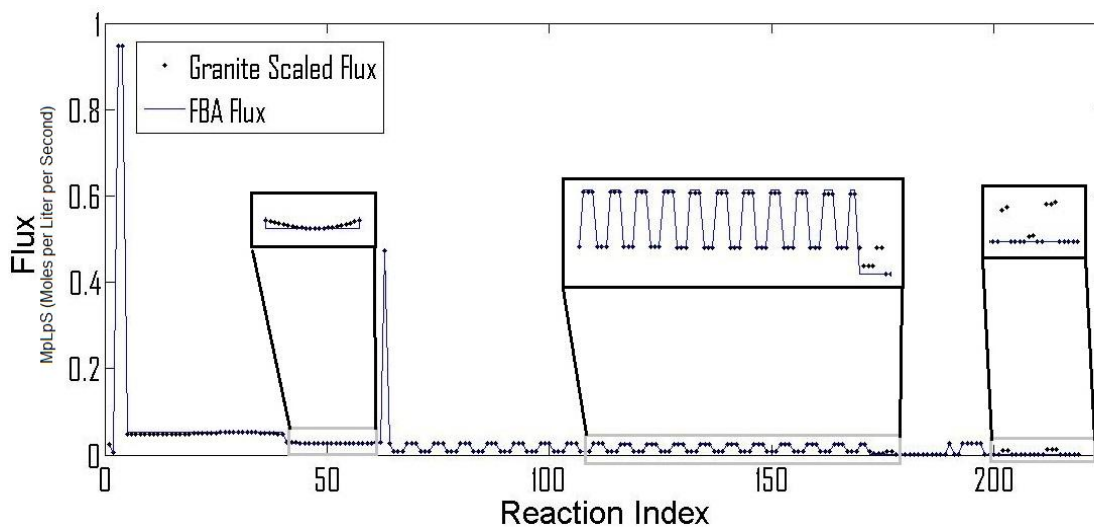


Fig. 4. Comparison of the FBA and ABM-MAS flux profiles. The reaction flux across each reaction point in the mycolic acid pathway (MAP) was compared in this plot. The x axis values represents the numbered reactions in the MAP SBML model while the y axis values represents the flux value calculated using FBA (blue line) and the ABM simulation (black dot). The inset shows the comparison of the high-complexity region of the Flux plot. The comparison shows that the ABM approach is able to emulate the internal flux properties of the FBA analyses with a correlation of 0.99

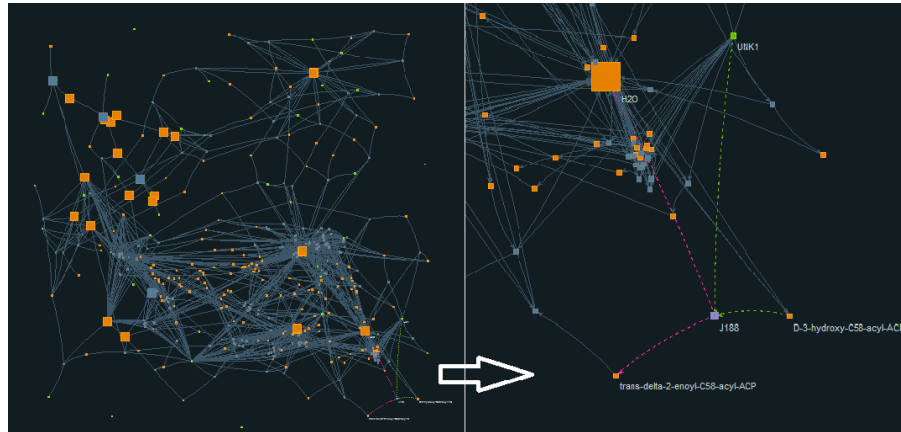


Fig. 5. The GRANITE tool includes a visualization application that allows direct interaction with the simulation. The GUI view displays a directed graph in which nodes are GRANITE memes and directed edges are the relationships between them (the right hand plot is zoomed in on a specific reaction). An edge from a meme to a reaction node implies that the meme is a reactant; an edge to a meme from a reaction node implies that the meme is a product of that reaction. Node color and size are configured based on the interaction model. In a metabolic network, color represents meme type such as reaction or metabolite. The size of a node represents reaction flux or concentration. Researchers can easily identify highly active reactions (agents) and select a subset of memes to compare their property values in real time in a chart view. Users can also perturb the system by changing some meme properties at any time and observe the effects of those perturbations on the system evolution.

(initial conditions inclusive) results in a new system-state with new observable features that can be recreated in the lab, the perturbation is informative and the predictive capability of the model increases for the next round of *in silico* experimentation. Iterations on this hypothetico-deductive cycle promise to build more accurate predictive models and reveal the general principles of the biological system under study. Thus GRANITE is an expressive, scalable, and predictive environment for modeling and simulating biological systems that enables bench researchers to integrate existing system descriptions with current hypotheses, and construct *in silico* exper-

iments that lead to predictions which can be tested in the laboratory. The results of those experiments can inform refinements to the system model that improve the prediction capability and focus lab experimentation. We intend to apply this technique to other interaction networks and integrated systems of metabolic, gene regulatory, and signal transduction networks, that are of interest to systems biology researchers and developers.. Ongoing work is being directed toward establishing a GRANITE user community, so that comprehensive systems-level *in silico* simulations of biological and biochemical networks can be collaboratively designed, created, and developed.

ACKNOWLEDGEMENTS

The authors would like to thank the sponsoring team at NIH, led by Dr. J.J. McGowan, for providing insightful comments on this manuscript. Brian Peck at Lockheed Martin deserves a special acknowledgement for his recent development efforts on GRANITE.

REFERENCES

- Erhard., *et al.* (2008) FERN – a Java framework for stochastic simulation and evaluation of reaction networks, *BMC Bioinformatics*
- Fei Li., *et al.* (2009) PerturbationAnalyzer: A tool for investigating the effects of concentration perturbation on protein interaction networks, BioInformaitcs, Oxford University Press.
- Bonabeau, Eric (2002) Agent-based modeling: methods and techniques for simulating human systems. *Proc. National Academy of Sciences* 99(3): 7280-7287
- Raman, K. *et al.* (2005) Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-Tubercular Drugs. *PLoS Comput Biol.*, 1(5) e46. doi:10.1371/journal.pcbi.0010046
- Axelrod, Robert (1997) *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton: Princeton University Press. ISBN 978-0-691-01567-5
- A. S. Rao and M. P. Georgeff. (1995) BDI-agents: From Theory to Practice, In Proceedings of the First International Conference on Multiagent Systems (ICMAS'95), San Francisco.
- Weiss, G. ed. (2000) *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, MIT Press.
- Barry CE III, Lee RE, Mdluli K, Simpson AE, Schroeder BG, et al. (1998) Mycolic acids: Structure, biosynthesis and physiological functions. *Prog Lipid Res* 37: 143–179.
- Watanabe M, Aoyagi Y, Ridell M, Minnikin D (2001) Separation and characterization of individual mycolic acids in representative mycobacteria. *Microbiology* 147: 1825–1837.
- J. Lawson, R. Singh, *et al* (submitted 2012), A Retrodiction Approach Using a Genetic Algorithm to Analyze the Mycolic Acid Pathway, *PLoS*

Population Structure and Related Attribute-Weighting Schemes Under the Assumption of Infectious Disease Scenarios

I. Gomez-Lopez¹, O. Loza¹, and A. R. Mikler¹

¹Computer Science, University of North Texas, Denton, Tx, United States

Abstract—*Epidemic modeling has been utilized to gain a better understanding of the infectious disease spread by means of studying the associated factors to the epidemic. In this regard, demographics are factors that affect the preference of individuals to interact with others of similar characteristics. Similarly, geographic characteristics influence the contacts dissemination within a physical boundary. The identification of specific groups within the population that favors the progression of the disease and their interconnection with other population subgroups are fundamental to assess the correlation between the population and disease dynamics. Furthermore, the correlation of the dominant demographic features and their influence in the disease behavior would not only give us insights of the disease dynamics, but also permit to track down possible dissemination of secondary infections to similar clusters. In this work, the resulting distribution of the population into clusters is mapped into graph representation so the clusters network properties can be observed and studied.*

Keywords: Population, Dynamics, Demographics, Geographics, Epidemics, Structures

1. Introduction

Diverse population properties, such as demographics and geographic characteristics, exert influence on the disease dynamics and the distribution of people into groups. For instance, changes on the final number of infectious individuals or the velocity and the duration of a disease are dictated by the relationship of the disease itself and the population characteristics[1]. The Center for Disease Control and Prevention has observed that characteristics of the population, such as sex, age, race and ethnicity, and socioeconomic factors are related to prevalence or emergence of a particular disease[2]; nevertheless, the correlation between those factors and a disease has to be studied in order to gain a better understanding of the disease dynamics. Methodologies and theories have been proposed and implemented to learn more about the disease attributes and its spatial-temporal structure; also, observations of diverse infectious processes within different population groups have been researched[3]. In this regard, studies of varying population characteristics and the impact in the disease progression have been made, and the characteristics of population subgroups have been taken into consideration for analysis. For instance, the disease

parameters such as the number of secondary infections and the velocity of the disease transmission are influenced by the population groups properties such as density[4]. Current approaches take into consideration the population characteristics at the group level and analyze their effects on population aggregations with diverse densities. The distribution of people into groups is influenced by multiple factors such as geographics and demographics. Under these assumptions, a social environment within a group of individuals with similar characteristics influences the number of interactions a person can have inside and outside its group, and then as a consequence the disease spread as well[5]. The social elements that influence a disease spread are measured in terms of the group characteristics such as individuals age and density[6]. School children is the sector of the population that bears the highest risk for disease transmission due to their high contact rate and limited immune response[7]. Current work focuses on the population structure that entails a distribution into groups by taking into account demographics, such as age, and geographics such as schools zone belongingness.

2. Methodology

There exists a natural segregation of the population into groups with people of similar characteristics[8]. In this methodology, a synthetic population constructed from aggregated data from the US Census 2000, is utilized to find arrangements of people into groups in a sample county. Demographics such as age, ethnicity, gender, and school grade; and geographics such as school zone, are to be utilized to build a structure of the population. A hierarchy of the population attributes is generated to assign larger weights on attributes of the interest of this study. The identification of groups within the population follows a hierarchization of attributes as seen in the Figure 1.

In this ranked chart it can be observed that, the geographic location and the age attributes are assigned more weight than the rest of the attributes. Concurrently, the weight among the remaining demographic and geographic attributes is not differentiated, and as a result, the distribution of the population into groups is highly influenced by the age and the geographic location. In order to obtain a distribution that follows these rules, a document retrieval technique is utilized for processing the synthetic population database. The Vector Space model allows to represent entities, such as

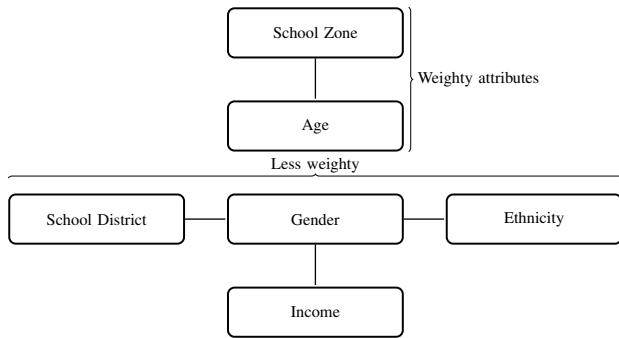


Fig. 1: Population Attributes Hierarchy

documents, as a collection of vectors of words. In this model, each dimension of the vector is represented by a document-term and its weight-value[9]. In our model, each individual is an entity to be represented as a vector of attributes, and each attribute is assigned a weight according to our hierarchy of attributes depicted in Figure 1. The synthetic population P with elements $P = \{p_1, p_2, p_3, \dots, p_n\}$, where $p_i = \{a_1, a_2, a_3, \dots, a_m\}$ is an individual represented as an array of weighted geographic and demographic attributes, and the dimension m of the multidimensional space, is equal to the number of the population attributes. Having mapped the population into a multidimensional space, the similarity between any two individuals is computed by means of a distance metric. The *cosine similarity* permits to calculate the similarity between any two individuals by computing the cosine angle between them as seen in Equation 1.

$$sim(p_i, p_j) = \frac{\sum_{k=1}^m a_{p_i,k} \times a_{p_j,k}}{\sqrt{\sum_{k=1}^m a_{p_i,k}^2} \times \sqrt{\sum_{k=1}^m a_{p_j,k}^2}} \quad (1)$$

In summary, this similarity metric is utilized to construct a matrix of distances among the population individuals, which in turn is used for the population clustering computation.

2.1 Clustering

In this section, a distribution of the population into groups of people with similar characteristics is generated. The criterion for people clustering is the hierarchy of attributes previously introduced. Two experiments with different assumptions are performed in order to show the effects of attribute-weighting on the population structure and the disease dynamics as well. The presuppositions for the clustering experiments on the synthetic heterogeneous population are the following:

- Clustering with homogeneous attribute-weighting
- Clustering with non-homogeneous attribute-weighting following a hierarchy scheme.

The algorithm to cluster the synthetic populating is shown in the Algorithm 2.1.

Algorithm 2.1: CLUSTERING(P)

```

procedure HIERARCHYCLUSTERING( $P$ )
  while  $K == 1$ 
  do {
    DISTANCE( $C_K, C_{K-1}$ )
    if  $Distance \leftarrow \min \text{DISTANCE}(C_K, C_{K-1})$ 
    then {MERGE( $C_K, C_{K-1}$ )
  
```

```

procedure GOODNESS( $clustering$ )
   $hubberts \leftarrow \text{HUBBERTS}(clustering)$ 
   $dunn \leftarrow \text{DUNN}(clustering)$ 
   $silhouette \leftarrow \text{SILHOUETTE}(clustering)$ 
   $goodness \leftarrow \text{sum}(hubberts, dunn, silhouette)$ 
  return ( $goodness$ )
  
```

```

global  $P, K = N$ 
 $optimal \leftarrow 2$ 
 $P \leftarrow \{p_1, p_2, p_3, \dots, p_n\}$ 
while  $C_{optimal} \leq optimal$ 
  {
    comment: Find nearest pair of clusters in  $P$ 
     $Tree \leftarrow \text{HIERARCHYCLUSTERING}(P)$ 
    comment: Cut the Tree
  }
  do {
     $numberC \leftarrow \text{CUTTREE}(Tree)$ 
    comment: Goodness Metrics
     $clustering \leftarrow \text{K-MEANS}(numberC)$ 
     $C_{optimal} \leftarrow \text{GOODNESS}(clustering)$ 
  }
  
```

According to current clustering procedure, the final distribution of the population is the C clustering with k groups of the population P within a partition $C = \{c_1, c_2, c_3, \dots, c_k\}$, with $k \geq 1$ clusters that satisfy $\forall (c_i, c_j)$ such that $i \neq j$, $c_i \cap c_j = \emptyset$ and $\cup_{i=1}^k c_i = P$. This partitioning of the population is subject to change if any new assumptions for the clustering are made. In current work, k-means clustering is utilized to find groups within the population, however, this method needs to know the k number of clusters in advance. For this reason it is necessary to assess the appropriate choice of k with alternate methods. In this regard, a hierarchical clustering algorithm of the population is used in order to estimate a *tentative* value of k . Hence, a dendrogram is generated and studied so that the number of clusters can be inferred. According to the properties of the dendrogram, such as height and shape of the branches, a cut is performed at an appropriate height of the tree, and a *tentative* value of k is determined. Then, k -means makes use of the choice of k to generate the partition of the population. Finally, the quality of the resulting clustering has to be evaluated by means of the within-cluster compactness and between-

cluster distances. Goodness metrics, such as Hubert's gamma coefficient, the Dunn index, and Silhouette index, are utilized to assess whether a new estimation of k is required or the *tentative* k is chosen as the *final* k to perform the partition of the population[10], [11], [12].

3. Results

Making use of the methodology described in the previous section 2 for clustering the synthetic population of the sample county, two different structures of the population were generated. First, a clustering $C_\alpha = \{c_1 = 178, c_2 = 323, c_3 = 270, c_4 = 117, c_5 = 402, c_6 = 210\}$ with six clusters, and a second clustering $C_\beta = \{c_1 = 182, c_2 = 1120, c_3 = 31, c_4 = 18, c_5 = 22, c_6 = 42, c_7 = 19, c_8 = 48, c_9 = 18\}$ with nine clusters. In this regard, C_α was constructed under the assumption of zero *attribute-weighting* for the population features, whereas C_β was calculated making use of the *attribute-weighting* hierarchy scheme for the population features as shown in the Figure 1. On one hand, in clustering C_α , the distribution of the population into clusters has been roughly evenly distributed into medium size clusters ranging from 117 to 402 individuals. On the other hand, in the clustering C_β , the cluster size is non-uniformly distributed with one very large cluster of 1120 individuals and seven small clusters varying size from 18 to 48 individuals. This partitioning of the population into clusters with different properties, shows the correlation between the assignment of weights to the population attributes and the non-uniform distribution of individuals into the clusters when the weighted-population is utilized. In this context, both C_α and C_β are utilized as two distinct scenarios to study an infectious disease spread and ascertain the potential correlation between these two differentiated structures of the population and the disease dynamics itself. The simulation of a disease progression over two different scenarios C_α and C_β starts when a single individual within a given cluster is randomly infected. Once the disease takes over, more individuals are infected locally and infectious contacts are exported globally to other clusters. Interactions among infectious and non-infectious individuals take place locally within clusters and globally between clusters. This makes contagion to describe a path that is originated from the cluster of the onset of the disease, following other clusters in sequence. The resulting path of the disease spread over the clustering is mapped into a graph representation so that a contagion network is produced and its properties can be observed and studied.

3.1 Scenario I

The contagion network for the C_α scenario is shown in Figure 2. A graph $G_\alpha = (V, E)$ where $G_\alpha(V) = C_\alpha$ and $G_\alpha(E) = (c_i, c_j)$.

In this contagion network it can be observed that there exist a correlation between the density of the population

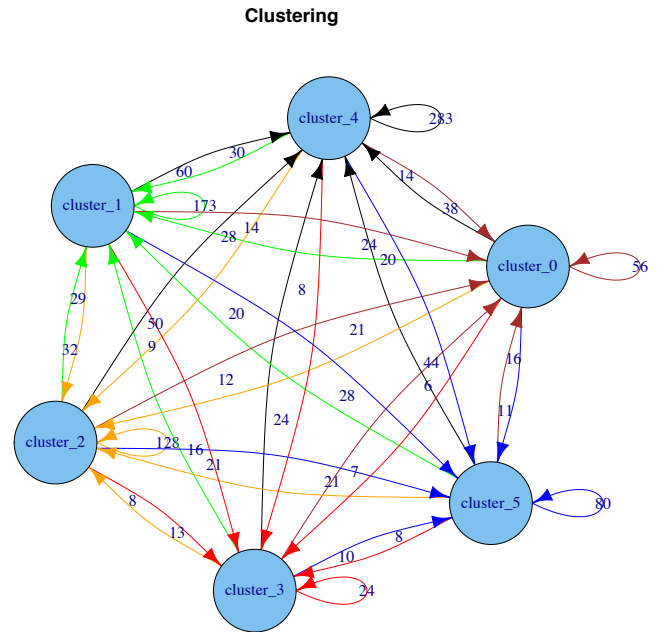


Fig. 2: Contagion Network non-weighted population attributes

and the number of individuals infected in each cluster. It has been previously ascertained the quasi-proportional distribution of the number of individuals into the clusters, and such structure has provoked that the disease is also evenly distributed in the clusters infecting approximately from 50% to 60% of the total number of individuals in each cluster. The disease spread follows a path that cover all the clusters favoring the interactions between infectious and non-infectious individuals back and forth among all the clusters facilitating the global contagion. The correlation between the disease spread and the population dynamics also has effects on other epidemic characteristics such as the disease duration and the time of onset. In addition, this behavior can be observed In the Figure 3 where the six clusters epidemics are presented with the same scale so the differentiated dynamics can be visually discerned. Also, it can be noted that the time between the onset and the conclusion is about 30 days for each of the single cluster outbreak. In addition, different onset times are observed in each cluster, slightly shifted on time due to the fact that the disease is randomly initiated in a single cluster and exported to others in apparently different times. However, the lack of synchronization between these times is barely perceptible due to the intrinsic properties of the clustering, such as cluster size and the between-cluster distance, that promotes such behavior. Finally, an illustration of another perspective of the six outbreaks is shown in the Figure 4. In this Figure, every cluster is depicted with a different color so that the six outbreaks that were started sequentially can be observed. The order in which the clusters are infected depends on the

population density of each cluster. As a result, in the Figure 4 it can be observed that the first outbreak onset occurs in the most populated cluster, that is the highest bell curve, and the rest of the outbreaks start sequentially following an order of clusters that decrease in size.

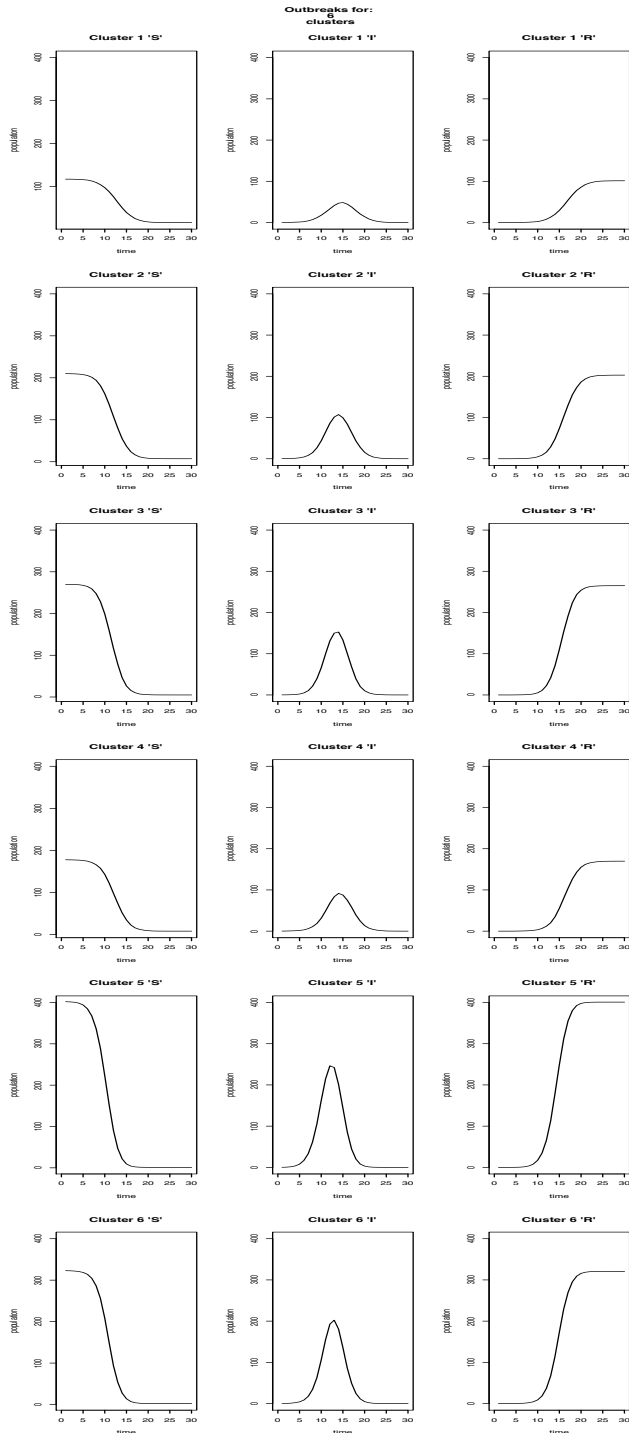


Fig. 3: Outbreaks SIR in each cluster: Scenario I

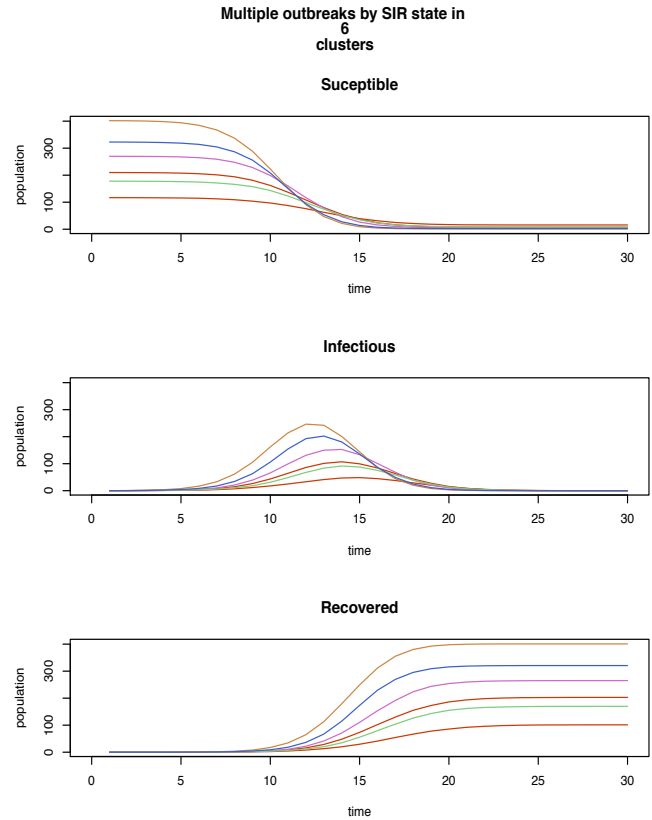


Fig. 4: Multiple outbreaks by SIR state in 6 clusters: Scenario I

3.2 Scenario II

The contagion network for the C_β scenario is shown in Figure 5. A graph $G_\beta = (V, E)$ where $G_\beta(V) = C_\beta$ and $G_\beta(E) = (c_i, c_j)$.

The contagion in this network shares some properties with the contagion in the Subsection 3.1 but also shows contrasting behavior. The correlation between the density of the population and the number of individuals for the scenarios persists but in different proportions. The non-uniform distribution of individuals into nine clusters has affected the number of infected individuals per population subgroup. In this experiment it can be seen that the smaller the size of the cluster is, the smaller the number of infected individual a cluster have. For instance, in the clusters where the size oscillates around 38 individuals the proportion of the local population being infected is about 35 percent. In this contagion network the traffic of individuals from one cluster to another is rather restricted. The size of the clusters influences intrinsically and externally the disease dynamics within-cluster and between-clusters. In the Figure 5 it can be noted multiple zeroes in the between-clusters inward and outward arrows, and within-clusters self pointing arrows as well. This is because the number of interactions between

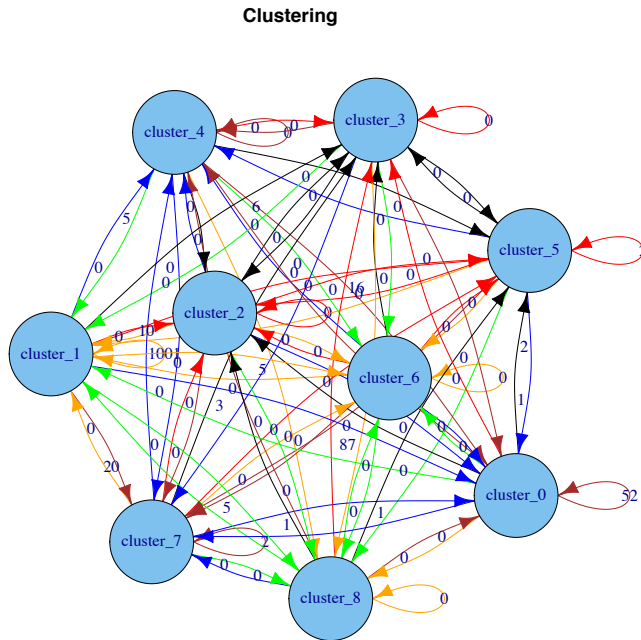


Fig. 5: Contagion Network weighted population attributes

two individuals that lead to a successful infection within the cluster c_k , and between the clusters c_k and c_{k+1} , are accounted in the final number of interactions of the cluster where they originally were spawned. Also, in this experiment it can be noted that the starting time of the disease spread in each cluster is out of phase with each other. Furthermore, the onset times for both experiments, C_α and C_β , are out of phase with each other as well. The shifting on the onset time of the disease from cluster to cluster goes from three to eight days and is larger than the scenario in the Subsection 3.1, where the shifting-time of the onsets is imperceptible. In addition, details can be noted in the Figure 6 where the same scale was preserved for the first eight clusters to facilitate the comparison and contrast of the particular cluster properties. For instance, in the cluster four the onset is around day three, whereas in cluster number six it starts around day eight. This shows a delay up to five days between onsets, whereas in previous experiment it was barely of one day. Finally, an illustration of another perspective of the nine outbreaks is shown in the Figure 7. In this Figure, every cluster is depicted with a different color so that the nine outbreaks that were started sequentially can be observed. Nonetheless, the scale utilized in the plot is the same as the smallest seven clusters so that the onset of every epidemic can be observed along with the two largest epidemics. The order in which the clusters are infected depends on the population density of each cluster as seen in the Scenario I. Consequently, in the Figure 7 it can be also observed that the first outbreak onset occurs in the most populated cluster, that is the highest bell curve, and the rest of the outbreaks start sequentially

following an order of clusters that decrease in size.

3.3 Two Scenarios

A final outbreak graph for the C_α scenario is shown in the Figure 8, and in the Figure 9, the scenario of C_β is depicted.

Making use of these figures facilitates the comparison and contrast of the properties of the disease dynamics. For instance, in the Scenario II it can be noted that the velocity of the spread of the disease in general is faster than the one of the Scenario I. In the Scenario I the outbreak takes up to 20 days to fade out, while on the contrary, it takes 15 days to the outbreak to die in the Scenario II. The attribute of the population structures that plays an important role in this contrasting characteristic is the population density. The number of individuals clearly influences the number of contacts that are been made locally and globally. This is reflected on the total number of infected individuals for both scenarios at the end of the outbreaks. Scenario I equals its final number of infectious individuals to the total population number, whereas, the final number of infectious individuals in the Scenario II is 10% smaller than those in Scenario I.

4. Conclusion

Taking into consideration different assumptions to generate the partitions of the population into clusters produces two different structures of the population. Assigning different weights to the attributes of the population permits to study different distributions of the same population: clustering C_α and C_β . Every arrangement of the population has distinct properties in terms of population density. Under these circumstances, the spread of the disease is affected by the structure of the population itself. Changes in onset times for the disease, the duration of an outbreak and the final number of infected individuals are parameters that are influenced by the structure of the population. Particularly, the quasi-even distribution of people in the C_α 's clustering causes the disease spread to have the same behavior in every cluster, whereas, in C_β 's irregular distribution of people in its clusters generates diverse behaviors of the disease dynamics. Finally, it can be concluded that there exists a correlation between the clusters density, affected by the selected weighting scheme that derives a particular structure of the population, and the velocity and the duration of the outbreak in the population. Current work studies an abstraction of the disease dynamics. Under the assumption of an infectious disease, demographics and geographics of the population are mapped into a contagion network to facilitate the analysis of the disease and the population dynamics; nevertheless, the visualization is an important aspect of the analysis that needs to be addressed. Making use of the *latitude-longitude* coordinates of every individual of the clustering, the future work involves the spatial mapping of the contagion network information into a geographic map

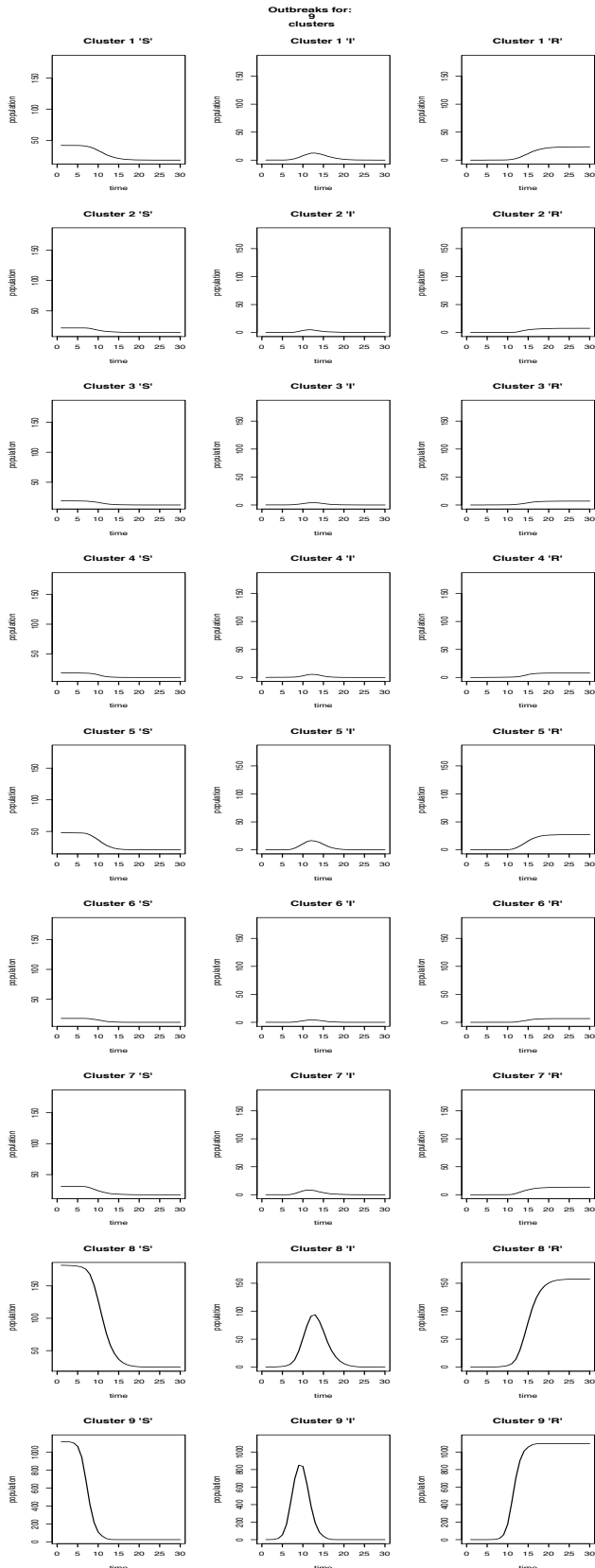


Fig. 6: Outbreaks SIR in each cluster: Scenario II

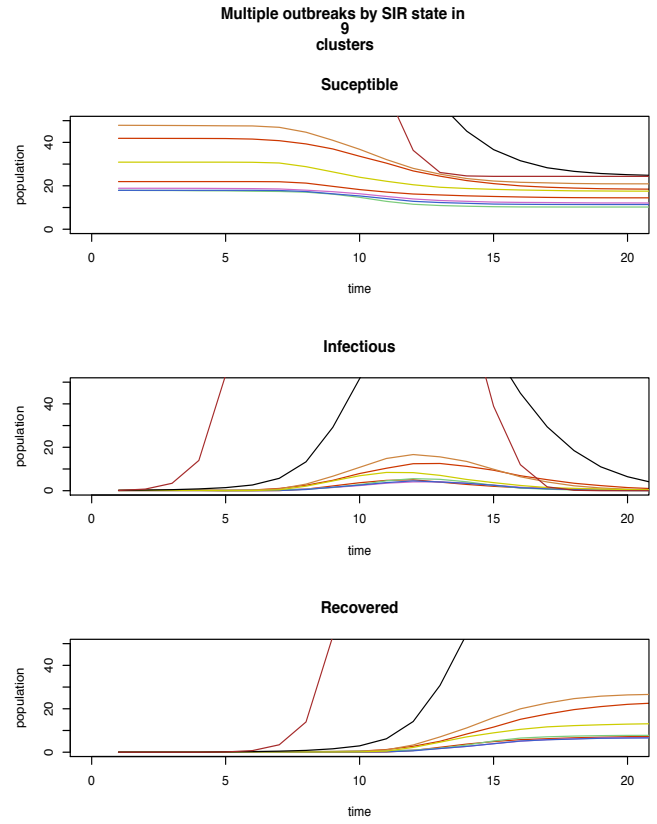


Fig. 7: Multiple outbreaks by SIR state in 9 clusters: Scenario II

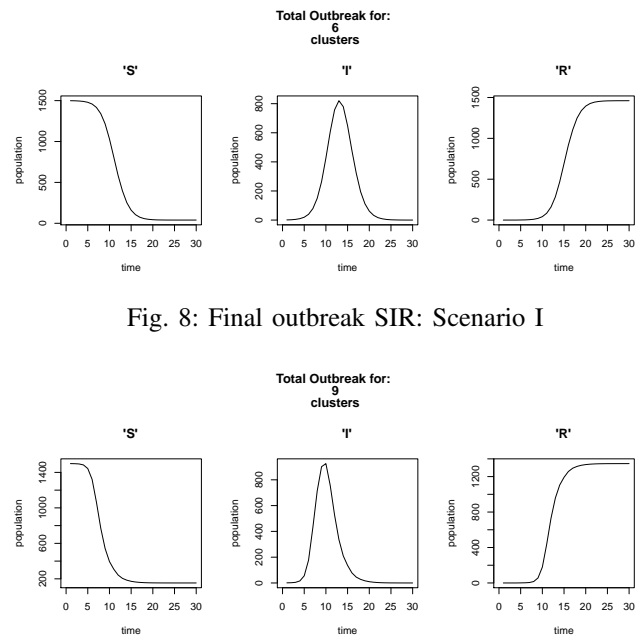


Fig. 8: Final outbreak SIR: Scenario I

Fig. 9: Final outbreak SIR: Scenario II

taking into consideration the *latitude-longitude* coordinates of the cluster elements. Allowing to geographically observe the physical distribution of the population in a given spatial region.

References

- [1] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds, "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Med*, vol. 5, no. 3, p. e74, 03 2008. [Online]. Available: <http://dx.doi.org/10.1371/journal.pmed.0050074>
- [2] C. for Disease Control and Prevention, "Population characteristics and environmental health," 01 2012. [Online]. Available: <http://ephtracking.cdc.gov/showPopCharEnv.action>
- [3] D. Wang and S.-J. Xiong, "Effects of disease characteristics and population distribution on dynamics of epidemic spreading among residential sites," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 13, pp. 3155 – 3161, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378437108000150>
- [4] D. Mollison, Ed., *Epidemic Models: Their Structure and Relation to Data*, 1st ed., ser. Publications of the Newton Institute. Cambridge University Press, June 1995. [Online]. Available: <http://www.worldcat.org/isbn/0521067286>
- [5] S. Syme, "Social determinants of disease," *Annals of Clinical Research*, pp. 44–52, 1987.
- [6] P. Du, Bruce, P. O'Campo, and L.-A. McNutt, "Changes in population characteristics and their implication on public health research," *Epidemiologic Perspectives & Innovations*, vol. 4, pp. 6+, July 2007. [Online]. Available: <http://dx.doi.org/10.1186/1742-5573-4-6>
- [7] I. M. Longini and M. E. Halloran, "Strategy for Distribution of Influenza Vaccine to High-Risk Groups and Children," *American Journal of Epidemiology*, vol. 161, no. 4, pp. 303–306, Feb. 2005. [Online]. Available: <http://dx.doi.org/10.1093/aje/kwi053>
- [8] M. Mcpherson, Smith-Lovin, Lynn, and C. J. M., "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001. [Online]. Available: <http://dx.doi.org/10.2307/2678628>
- [9] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975. [Online]. Available: <http://doi.acm.org/10.1145/361219.361220>
- [10] L. Hubert and J. Schultz, "Quadratic assignment as a general data analysis strategy," *British Journal of Mathematical and Statistical Psychology*, vol. 29, no. 2, pp. 190–241, 1976. [Online]. Available: <http://dx.doi.org/10.1111/j.2044-8317.1976.tb00714.x>
- [11] J. C. Dunn, "Well separated clusters and optimal fuzzy-partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.
- [12] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, November 1987. [Online]. Available: <http://dl.acm.org/citation.cfm?id=38768.38772>

Blended HMMs: Reducing Redundancy in the SCOP HMM Database

Mingming Liu¹, Lenwood S. Heath¹, Layne T. Watson², and Liqing Zhang¹

¹Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

²Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Abstract—Hidden Markov models (HMMs) have been widely used to represent families or superfamilies of proteins that are regarded as evolutionarily-related groups. In a previous study, we have systematically analyzed the relationship between HMMs using a network method for the Structural Classification of Protein (SCOP) database and found high similarity among HMMs in the database. Based on the HMM network built in our previous study, we propose the concept of a blended HMM, aiming to reduce the redundancy of HMM models in the SCOP database. We construct a single HMM to integrate multiple HMM models into one model based on the similarity between HMMs reflected on the network. HMMER3 is used to build blended HMMs that represent the connected components (CC) in the original HMM network. The performance of each blended HMM is evaluated by measuring its ability to identify the correct superfamilies or families. Results show that these blended HMMs identify the correct protein sequence sets with accuracy over 95%. Blended HMMs provide a more compact representation of the protein families and superfamilies of the SCOP database, thus their use can reduce the size of an HMM database and decrease the computational cost of a large number of database queries.

Keywords: HMM, SCOP, Redundancy, Superfamily

1. Background

Protein sequence homology detection is an important task for understanding the evolutionary origin of different protein families. Homology among protein or DNA sequences is typically inferred on the basis of sequence similarity. Sequence-sequence comparison methods, such as FASTA or BLAST [1], [2], are used to detect conserved regions between sequences. However, pairwise comparisons have less sensitivity in detecting remote homology than profile-based methods[3], such as hidden Markov models (HMMs), which have been described very effective in detecting conserved patterns in multiple sequences [4], [5].

The Structural Classification of Proteins (SCOP) database is a comprehensive protein database with a hierarchical structure to classify proteins on the basis of their evolutionary and structural relationships. It is organized in a hierarchical structure consisting of four levels: family, superfamily, fold, and class. Protein domain sequences are

classified according to these four levels. SCOP uses HMMs to represent superfamilies or families. The basic procedure of building an HMM for a particular superfamily starts with a seed protein. Then, it performs a sequence search in a database to obtain other proteins that have sequence similarities above a set threshold. Finally a profile HMM is built based on a multiple sequence alignment (MSA) of these sequences. A previous study[6] demonstrated that multiple HMMs, each of which was constructed from a different seed sequence, produce better identifying results than a single HMM that was constructed from one seed. However, this is at the expense of redundancy in the HMM database[6]. To understand how the HMMs in the same or different superfamilies are related, we performed an analysis of all the HMMs in SCOP using a network approach[7]. We used 13,730 HMMs from seven protein classes to build an HMM network using the HHsearch program[8]. The final network consists of many connected components. Nodes in the network represent HMMs and edges similarity between HMMs. Consequently, HMMs within each connected component show high similarity to one another. If we can achieve similar performance by one HMM rather than multiple models, then a more compact representation of the original database is achieved. In this paper, we propose a method to build blended HMMs that can represent two or more HMMs in the same connected component obtained from our previous paper[7].

Our method to construct a blended HMM consists of two steps. First, a set of sequences is sampled from those used to build the HMMs in a connected component in the network and a multiple sequence alignment is constructed by MUSCLE[9]. Second, a profile HMM of the set is built from the multiple sequence alignment by HMMER3[10]. A model-scoring program is used to assign a score to any sequence of interest with respect to the blended profile HMM; the better the score, the greater the chance that the query sequence is a member of the protein family represented by the profile HMM. In this way, each sequence in a database can be scored to find the members of the family present in the database. The performance of the blended HMM was measured by its ability to identify members of a protein family in sequence databases. Our results verify that the blended HMM maintains high sensitivity without losing resolution. It is also compared to the coverage of the original

HMMs in a connected component to verify similarity and resolution.

2. Methods

2.1 Notation

We first introduce notation. Let D_s denote the SCOP HMM database and D_b denote the blended HMM database. The original HMMs in a connected component C_i with size m_i are referred to as $M_{i1}, M_{i2}, \dots, M_{im_i}$. The blended HMM corresponding to C_i is B_i . Each HMM M_{ij} was built from multiple sequence alignment (MSA) A_{ij} , and A_{ij} is created by the set of sequences S_{ij} . All the sequences for C_i are in the set $F_i = \bigcup_{j=1}^{m_i} S_{ij}$. The MSA for the blended HMM B_i is sampled from the sequence set F_i .

2.2 Data Description

Our data have two sources. The first source is the HMM network built in [7], where the nodes represent SCOP HMMs and the edges represent the similarity between HMMs. The network contains 151,461 edges and 11,929 vertices. There are 1524 connected components (CCs), 1236 of which are fully connected. Overall, 566 CCs have size 2, 261 size 3, 140 size 4, 85 size 5, 60 size 6, and 124 sizes greater than 6. The second source is the multiple sequence alignments (A3M format, derived from aligned FASTA format) used to build the corresponding profile HMMs, which were downloaded from <ftp://ftp.tuebingen.mpg.de/pub/protevo/HHsearch/databases>.

2.3 Building blended HMMs

There is one blended HMM B_i for each connected component C_i . For a given C_i , B_i was built based on a multiple sequence alignment sampling from $F_i = \bigcup_{j=1}^{m_i} S_{ij}$. Sampling was performed to select sequences from each S_{ij} for C_i . Let $N_i = |F_i|$ and t_i be the number of sequences sampled (at most 2000 sequences were sampled), then $n_i = t_i/N_i$ is the fraction of sequences sampled from each S_{ij} . Because a blended HMM is used to represent a protein family connected component without losing much coverage rather than using multiple HMMs, we assume that all members in the same CC belong to the same family or superfamily more often than expected under a random network connection model. Our previous work [7] shows that more than 95% of connected components have only members from the same superfamily.

MUSCLE[9] was used to do multiple sequence alignment, `muscle -in seqs.fa -out seqs.afa -maxiters 2`, where `seqs.fa` is the input FASTA file and `seqs.afa` the output MSA file. Since the number of sequences is large (typically > 1000), the MUSCLE option `-maxiters 2` was used to compromise between speed

and accuracy. Based on the resulting MSA, HMMER[10] was used to build the blended HMM B_i with default parameters using the command: `hmmbuild blendhmm seqs.afa`, where `blendhmm` is the blended HMM B_i . B_i is called a blended HMM B_i for the connected component C_i . All blended HMMs were combined into a database $D_b = \{B_i \mid 1 \leq i \leq 1524\}$.

2.4 Scoring

The `hmmsearch` program in the HMMER package takes a query sequence and searches it against a profile HMM database. A bit score is assigned to a target model if it significantly matches the sequence. A bit score is a log-odds ratio (base two) comparing the likelihood of the profile HMM to the likelihood of a null hypothesis (an independent, identically distributed random sequence model, as in BLAST). More precisely, for a hidden Markov model M , $s(M) = \log \frac{P_m}{P_r}$, where P_m is the probability of the alignment to the HMM M and P_r the probability of the sequence given the random overall sequence model.

3. Results and Discussion

3.1 Comparison between blended HMM and original HMMs

Given a connected component C_i , B_i was compared with M_{ij} ($1 \leq j \leq m_i$) by matching each sequence in S_{ij} . First, for a given C_i , M_{ij} ($1 \leq j \leq m_i$) and B_i were concatenated as an HMM database D_{mi} , then each S_{ij} was matched against the database with `hmmsearch` `hmmdb` `seq.fas`. The parameter `hmmdb` is the HMM database (D_{mi}) to be searched, and the file `seq.fas` contains the query sequences (S_{ij}). For each sequence, the program will assign a bit score to a profile HMM in the database if the sequence significantly matches the HMM.

All sequences F_i associated with C_i were classified into two groups for a given HMM model M_{ij} , the training set H_{ij} and the testing set T_{ij} . The training set for M_{ij} is the sequence set used to build M_{ij} , that is $H_{ij} = S_{ij}$, and the testing set contain all other sequences in F_i , that is $T_{ij} = F_i - H_{ij}$. The comparisons were categorized into two cases. In the first case, when the numbers of sequences in all S_{ij} ($1 \leq j \leq m_i$) are small (typically $|S_{ij}| < 50$), B_i has better performance than M_{ij} if matched against the sequences in T_{ij} but not for the sequences in H_{ij} . Example 1. For instance, suppose C_i contains five HMMs, referred to as $M_{i1}, M_{i2}, M_{i3}, M_{i4}, M_{i5}$ and $|S_{ij}|$ ($1 \leq j \leq 5$) equals 20, 1, 1, 6, and 6 respectively. Figure 1 shows that the blended HMM achieves a greater score for testing sequences but a lower score for training sequences. In other words, the loss of B_i is primarily on the training sequences. Another instance is a CC with size $m_i = 2$. Let s_{B_i} and $s_{M_{ij}}$ denote the bit score for blended HMM B_i and HMM M_{ij} respectively. If we use sequence set S_{i1} to match against the database,

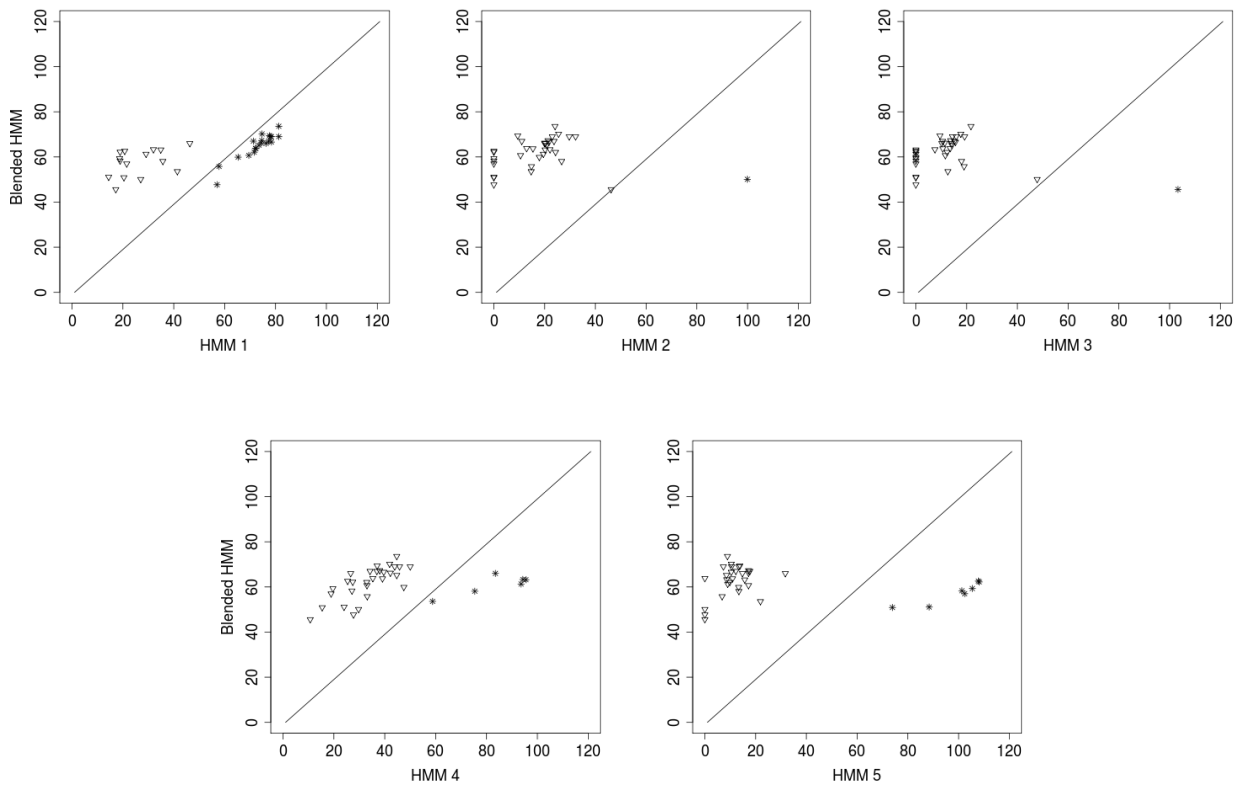


Fig. 1: Pairwise comparison between blended HMM and original HMMs for Example 1 (small $|S_{ij}|, |F_i| = 34$). HMM models are d1bnba_, de14ra_, de14ta_, d1fd3a_, and d1kj6a_ in order respectively (all with SCOP ID g.9.1.1 (A:)). Each point represents a sequence. The triangles and stars are the corresponding training set H_{ij} and testing set T_{ij} respectively for a given M_{ij} . Axes are bit scores.

over 95% of the sequences fit the model M_{i1} the best, i.e., $s_{M_{i2}} < s_{B_i} < s_{M_{i1}}$. For set S_2 , we have $s_{M_{i1}} < s_{B_i} < s_{M_{i2}}$ for most of the sequences. The average scores are shown in Figure 2.

In the second case, when the numbers of sequences in all S_{ij} ($1 \leq j \leq m_i$) are large (typically $|S_{ij}| > 1000$), the performance of B_i and the M_{ij} tend to be similar. Example 2. For instance, for a CC with size $m_i = 5$ and associated with a large number of sequences, the scores for both HMMs tend to be close to each other for a majority of sequences. For a given sequence, we define the loss of a blended HMM B_i with respect to an original HMM M_{ij} as

$$L_{ij} = s_{M_{ij}} - s_{B_i}. \quad (1)$$

The distributions for the loss of the blended HMM for this example is shown in Figure 3. It shows that the original HMMs and the blended HMM tend to have similar scores. One may hypothesize that, in example 2, the loss function has a normal distribution. To test normality, we combined all the loss values for the blended HMM with respect to each

original HMM model in example 2 using all the sequences in F_i and did a Shapiro-Wilk test[11]. The test showed that $W = 0.9597$ with p -value $< 2.2 \times 10^{-16}$, indicating that the loss function likely does not follow a normal distribution. For each connected component C_i , based on F_i , we define a loss matrix, where each entry $\Lambda^{(i)}$ is the loss score (Equation 1) of B_i with respect to M_{ij} based on sequence k ($1 \leq k \leq |F_i|$). We randomly tested 100 connected components each with approximately 6,000 sequences, and pick the maximum loss score from each sequence (row of the matrix) and fit the data to a type I extreme value distribution[12]. The scaled density distribution of maximum extreme value is shown in Figure 3. The probability of the loss score being less than or equal to zero is 0.746.

Based on the loss distribution, we hypothesize that, with increasing numbers of sequences, the bit scores of the blended HMM will be greater than or equal to those for the original models. To test this hypothesis, we applied Wilcoxon's signed rank test. Randomly select 100 connected components C_i each with around 6,000 sequences and combine all the bit scores of the blended HMMs and their

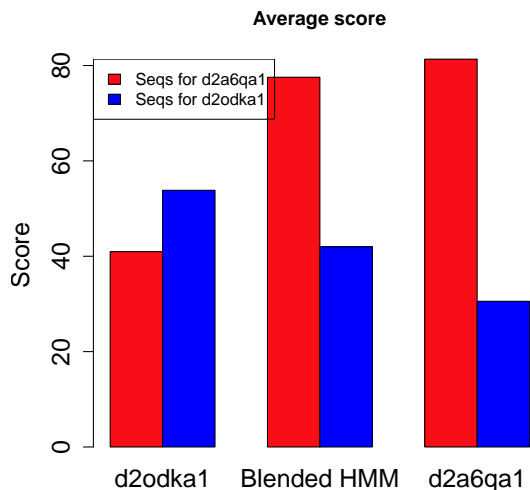


Fig. 2: Average score comparison between blended HMM and original HMMs.

original HMM models based on all sequences in all the F_i . The p -value $< 2.2 \times 10^{-16}$ illustrates that the bit scores of the blended HMMs are likely significantly greater than those of the original HMMs.

3.2 Homology Detection

A blended HMM database D_b was built by gathering all the blended HMMs for the CCs. This HMM database is used to identify the protein family that a protein sequence belongs to. For testing the homology detection ability of D_b , a test sequence set X was made by randomly selecting 10,000 sequences from the pool of sequence sets $\bigcup_i F_i$ and D_b was searched to see if the sequences were assigned to the correct superfamilies (the correct blended HMMs). Each query sequence to SCOP is attached to an HMM that represents one of the protein superfamilies. For any sequence in X , if B_i obtained the maximum bit score Δ_{B_i} , we classify the sequence to the superfamily represented by B_i . A query sequence is misidentified when it selects the incorrect model B_i or is ignored when it fails to select any model B_i in the database. Table 1 shows that the precision of D_b (94.93%) is less than that of D_s (98.76%), however, the precision of D_b per HMM model is much greater than that for D_s due to the size of the database. We also sampled 30 sets of queries each with 5,000 sequences and compute the accuracies by searching D_b ($98.99\% \pm 0.0035$) and D_s ($95.34\% \pm 0.0049$). To illustrate the efficiency of the blended HMM database D_b , we compare the computing time for queries to the blended HMM database D_b versus the original database D_s . Figure 4 shows a linear relationship between the computing time and the number of query searches, and

Table 1: Accuracy for Blended HMM database

	Size	Precision	Time (s)	misidentified	ignored
Blended HMM	1524	94.93%	842	20	487
SCOP HMM	11,929	98.76%	2008	24	100

searching the blended HMM database is about 2.35 times faster than the original SCOP database. Thus the blended database D_b improves computational efficiency by providing a more compact representation of the SCOP protein families and superfamilies.

3.3 Redundancy Measurement

We measure the redundancy of an HMM database using the network density defined by

$$R(G) = \frac{2|E|}{|V|(|V| - 1)}, \quad (2)$$

where $|E|$ is the number of edges, $|V|$ the number of models (HMMs), and G the HMM network or connected component in the network. Two HMMs are connected in the network if they have high similarity according to the all-against-all comparison using HHsearch. The density of the entire network G_{orig} for the original HMMs is only 0.0017 ($R(G_{orig}) = 151461 / \binom{13547}{2}$), but individual CCs tend to have high densities [7], with more than 82% of the CCs having densities more than 0.95, which illustrates a high redundancy in the HMM database.

We create a network for the blended HMMs in the same way as our previous work by using HHSearch[8]. HHsearch, similar to BLAST, uses a query that can be either a protein sequence or an HMM to search a sequence or HMM database. The HMM network was built in two steps. First, HHsearch [8] was used to perform an all-against-all HMM comparison with default parameters. Two HMMs are matched if the E-value is below 0.001 [7]. As a result, all 1,524 blended HMMs met the criterion, with 1,484 only having matches with themselves. Second, an undirected network (graph) $G_{new} = (V, E)$ was constructed, where the vertices V are HMMs, and there is an edge in E between two HMMs, if their E-value is below the threshold. Since we are trying to match a query HMM to our own HMM database, it is necessary to calibrate all the query HMMs. To calibrate, we use the command `hhsearch -cal -i query.hmm -d cal.hmm`, where `query.hmm` is the query HMM (B_i) and `cal.hmm` the calibrating HMM database that contains only one HMM per SCOP folder. Then the command, `hhsearch -i query.hmm -d hmddb` was used for searching the database, where `hmddb` is D_b . The density of the entire network G_{new} is 2.4127×10^{-5} ($R(G_{new}) = 28 / \binom{1524}{2}$) with 15 connected components. Therefore, the redundancy in the new network is greatly decreased from the original network ($R(G_{orig}) \gg R(G_{new})$). There are 40

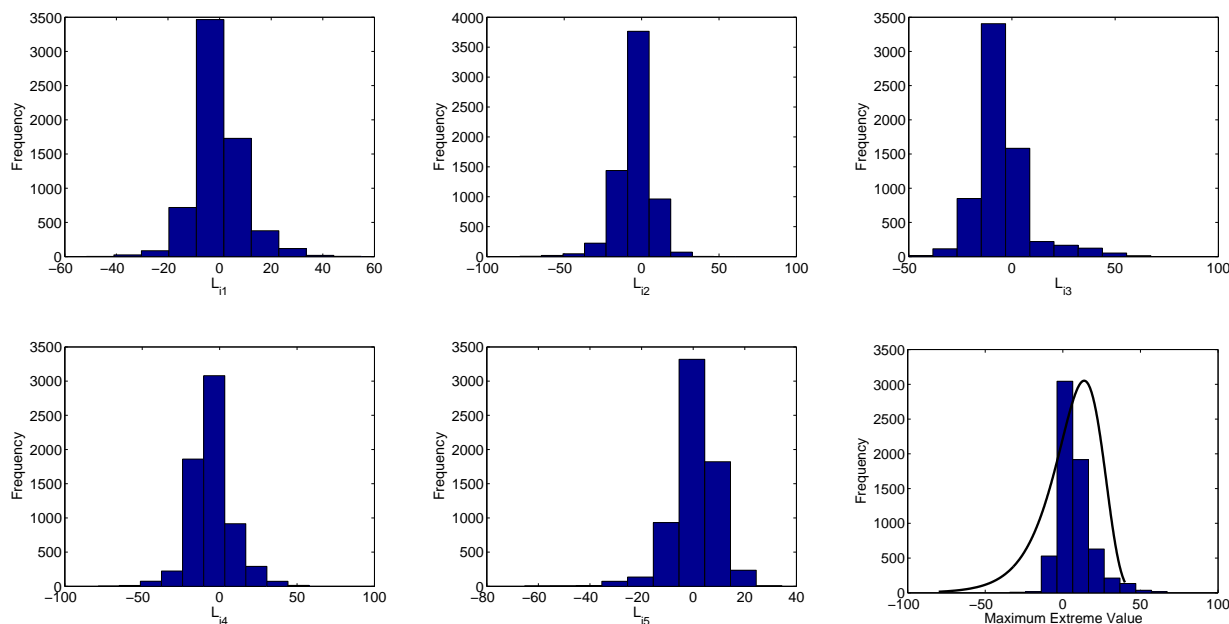


Fig. 3: Histograms of L_{ij} , $1 \leq j \leq 5$, for Example 2 (large S_{ij} , $|F_i| = 6544$). HMM models are d1h6ha_, d1kmda_, d1kq6a_, d1lcsa_, and d1xtea_ in order respectively (all with SCOP ID d.189.1.1 (A:)). x axis is the loss score L_{ij} . The curve in the last figure is the fitted extreme value distribution.

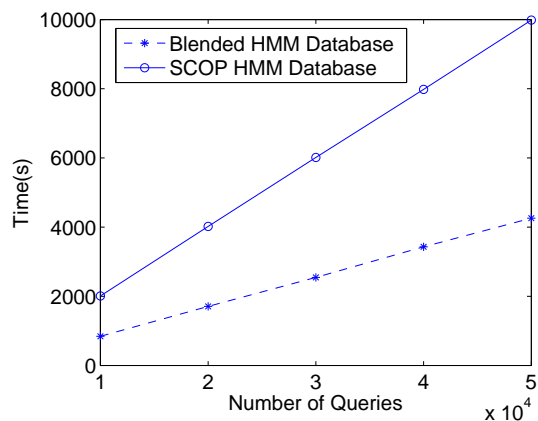


Fig. 4: Efficiency comparison between two databases

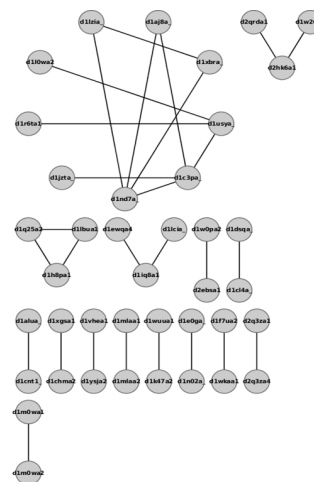


Fig. 5: The blended HMM connected components

blended HMMs that match other blended HMMs apart from themselves. If we select these blended HMMs to build a subnetwork, the density is 0.036, which includes 11 CCs with size 2, 3 CCs with size 3, and one CC with size 9 and density 0.28 as shown in Figure 5.

4. Conclusions

An HMM database provides a way for detecting members of protein families or superfamilies. It is an important resource to understand protein evolution and function. The HMM library in the SCOP database is widely used to

identify and conduct SCOP domain assignment for new sequences. However, the redundancy among HMMs influences its efficiency as our previous study revealed, which will influence its efficiency. In this paper, we proposed to use blended HMM to reduce the size of the original HMM database but preserve its performance. Using blended HMMs has obvious advantages when many queries are required, which happens frequently when a new genome is sequenced and many predicted protein sequences need to be

annotated. Moreover the blended HMM database can also serve as a preprocessing step. Sequences with high e-value can be filtered out using the blended HMM database, the remaining sequences with uncertainty may then be matched against the SCOP HMM database. It is noted that our approach of constructing blended HMMs can be one of many methods to integrate multiple HMM models into one model. Our method depends on connected components in the HMM network and multiple sequence alignments. In future work, we will extend our method. For example, instead of using a network to imply the similarity among HMMs, we can construct a hierarchical structure for HMM and integrate two HMMs into one according to the characteristics of the models. By adjusting the distance between two HMMs, we can decide the size of clusters and thus the size of the reduced database.

References

- [1] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Nat Genet*, vol. 85, pp. 2444–24448, 1988.
- [2] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *J Mol Biol*, vol. 215, pp. 403–410, 1990.
- [3] M. Madera and J. Gough, "A comparison of profile hidden markov model procedures for remote homology detection," *Nucleic Acids Res*, vol. 30, pp. 4321–4328, 2002.
- [4] R. Hughey and A. Krogh, "Hidden markov models for sequence analysis: extension and analysis of the basic method," *Bioinformatics*, vol. 12, pp. 95–107, 1996.
- [5] C. Y. H. T. Baldi, P and M. McClure, "Hidden markov models of biological primary sequence information," *Proc. Natl Acad. Set*, vol. 91, pp. 1059–1063, 1994.
- [6] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure," *J Mol Biol*, vol. 313, pp. 903–919, 2001.
- [7] L. Zhang, L. Watson, and L. Heath, "A network of scop hidden markov models and its analysis," *BMC Bioinformatics*, vol. 12, p. 191, 2011.
- [8] J. Söding, "Protein homology detection by hmm-hmm comparison," *Bioinformatics*, vol. 21, pp. 951–960, 2005.
- [9] R. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res*, vol. 32, pp. 1792–1797, 2004.
- [10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.
- [11] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52(3-4), pp. 591–611, 1965.
- [12] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

Multiscale Discretization for Reaction Diffusion Systems

Fei Li

Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061
Email: felix@vt.edu

Yang Cao

Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061
Email: ycao@cs.vt.edu

Abstract—Stochastic simulation of reaction-diffusion systems presents a great challenge because of the high computational cost in these systems. Straightforward extension of Gillespie's stochastic simulation algorithm (SSA) to reaction-diffusion systems leads to so-called Inhomogeneous Stochastic Simulation Algorithm (ISSA). However, the ISSA can be prohibitively expensive in computation if the discretization size is too small and results in a large system. Thus a proper size of the discretization for a reaction-diffusion system is critical. In this paper we present a multiscale discretization method for stochastic reaction-diffusion system simulation. With proper discretization scale, we can greatly reduce the size of the system and achieve high efficiency.

I. INTRODUCTION

Reaction-diffusion processes are used extensively in modeling of complex systems in areas including biology, social sciences, ecosystems, and materials processing. In recent years, stochastic modeling and simulation of reaction-diffusion processes have drawn more and more attention because of their applications in spatially inhomogeneous biological systems. Theoretically, the dynamics of spatially inhomogeneous stochastic system is governed by the reaction-diffusion master equation (RDME) [1], which was developed in 1970s. But the RDME is computationally impossible to solve for almost all practical problems. Stochastic methods were then proposed to simulate reaction diffusion systems. Spatial stochastic simulation is an extremely computationally intensive task, due to the large size resulted from the discretization of the system.

Gillespie's stochastic simulation algorithm (SSA) [2] is a widely used method to simulate stochastic biochemical systems under the assumption that the reaction system is in thermal equilibrium (also called well-stirred system or spatially homogeneous system under different circumstances). There exist several implementation of the SSA, such as the direct method [2], the first reaction method [2], and the next reaction method [3]. Great effort has also been taken in order to develop efficient approximation algorithms, such as the τ -leaping method [4] and the slow scale SSA [5], since the SSA is computationally intensive for most practical models.

When the "well-stirred" assumption is not valid, the SSA cannot be directly applied. Instead, it needs to be extended to spatially inhomogeneous system and that results in the inhomogeneous SSA (ISSA). The spatial domain is discretized into small voxels. Each voxel is well-stirred where the reactions

remain the same as in the homogeneous case. The diffusion is modeled as the Brownian motion between neighboring voxels. Each state variable of the system will have a local copy as the number of molecules of each species in each voxel at a given time. The key assumption is that within each voxel, the system is well-stirred. Hence the discretization size should be bounded by this homogeneity assumption. The discretization size has been studied theoretically and numerically [12], [14] for uniform 1-D discretization. Moreover, 2-D and 3-D uniform discretization methods have been applied to simulate nonlinear reaction-diffusion systems [15]. Non-uniform 1-D discretization strategies, such as adaptive and unstructured meshes, have been applied in stochastic process simulation. Additionally, Drawert et al. [11] have developed the finite state projection (FSP) algorithm [8] to simulate reaction-diffusion systems. Recent efforts, including the next subvolume method (NSM) [6], MesoRD [7], MSA [9], and the DFSP [11] methods, focused on speeding up the ISSA. The next subvolume method (NSM) [6] utilizes the priority queue structure originally proposed in the next reaction method. MesoRD [7] implements this method and has been widely used. The binomial tau-leap spatial stochastic simulation algorithm [10] uses a similar technique by combining the idea of aggregating diffusive transitions with the priority queue structure used in the NSM. However, these methods can still be prohibitively slow, due to the presence of fast diffusion. The MSA [9] was developed for the scenario where the diffusion rates are much greater than the reaction rates. The MSA uses an approximation method to calculate the net intervoxel diffusion transfers by realizing that the number of diffusion events conforms to a multinomial distribution which can be calculated and sampled. Drawert et. al. developed a novel formulation of the finite state projection (FSP) method [8], called diffusive FSP (DFSP) method [11] for efficient and accurate simulation of diffusive processes.

In this paper, we introduce a multiscale discretization method for multispecies systems with different diffusion rates. By assigning different discretization sizes to different species, we greatly reduce the diffusive transitions between neighboring voxels, resulting in improvement on simulation efficiency. The paper is organized as follows. Section II briefly reviews the mathematical background, including the chemical

master equation (CME), the stochastic simulation algorithm (SSA), and the reaction-diffusion master equation (RDME). In section III we present a simple one variable model and the theoretical analysis for a proper diffusion subvolume length. In section IV, we present numerical experiments that demonstrate the efficiency and accuracy of our multiscale discretization method. Finally, we conclude with an assessment of this approach, the applications, and discussion about future development.

II. BACKGROUND

A. Chemical Master Equation and Discrete Stochastic Simulation

Consider a biochemical system of N species $\{S_1, S_2, \dots, S_N\}$ interacting through M reaction channels $\{R_1, R_2, \dots, R_M\}$. The state vector is denoted by $X(t) \equiv (X_1(t), X_2(t), \dots, X_N(t))$, where $X_i(t)$ is the number of the molecules of species S_i at time t . The system is confined to a constant volume Ω , and is well-stirred. Each reaction channel R_j can be characterized by the propensity function a_j and the state change vector $\nu_j \equiv (\nu_{1j}, \nu_{2j}, \dots, \nu_{Nj})$. $a_j(X)dt$ gives the probability that one R_j reaction will occur in the next infinitesimal time interval $[t, t+dt)$, and ν_{ij} gives the change in the S_i molecule population induced by one R_j reaction. The matrix ν makes the stoichiometric matrix.

Once the propensity functions and stoichiometric matrix are determined, the chemical master equation (CME) completely depicts the dynamics of the system:

$$\begin{aligned} & \frac{\partial P(x, t|x_0, t_0)}{\partial t} \\ = & \mathcal{R}P(x, t|x_0, t_0), \\ = & \sum_{j=1}^M [a_j(x - \nu_j)P(x - \nu_j, t|x_0, t_0) - a_j(x)P(x, t|x_0, t_0)], \end{aligned} \quad (1)$$

where \mathcal{R} denotes the generating matrix for the Markov chain that describes the chemical reactions and $P(x, t|x_0, t_0)$ denotes the probability that $X(t)$ will be x given that $X(t_0) = x_0$. However, the CME is both theoretically and computationally intractable due to the huge number of possible combinations of states.

Gillespie's stochastic simulation algorithm (SSA) is an "exact" simulation algorithm as it follows the same probability assumption that rules the CMEs. Instead of solving for time evolution of the probabilities, the SSA generates sample trajectories step by step. In each step, the SSA answers two questions: when will the next reaction fire and which reaction will fire. Let $p(\tau, j|x, t)$ denote the probability that given $X(t) = x$, an R_j reaction will fire in the infinitesimal time interval $[t + \tau, t + \tau + d\tau)$. It can be derived that

$$p(\tau, j|x, t) = a_j(x)e^{-a_0(x)\tau}, \quad (2)$$

where $a_0(x) \equiv \sum_{j=1}^M a_j(x)$. Equation (2) is the mathematical basis of the SSA approach. It implies that the time τ to the next reaction is an exponential random variable with mean

and standard deviation $1/a_0(x)$, while j is a statistically independent integer random variable with point probability $a_j(x)/a_0(x)$. There are several Monte Carlo procedures for generating samples of τ and j according to their distributions. The simplest is the direct method, which generates two uniformly distributed random numbers r_1 and r_2 in the unit interval, and take

$$\begin{aligned} \tau &= \frac{1}{a_0(x)} \ln\left(\frac{1}{r_1}\right), \\ j &= \text{the smallest integer satisfying } \sum_{j'=1}^j a_{j'}(x) > r_2 a_0(x). \end{aligned} \quad (3)$$

The system is then updated according to $X(t + \tau) = x + \nu_j$. This process will repeat until the simulation end criterion is reached.

The SSA is exact in the sense that the sample paths it generates are distributed according to the solution of the CME. However the SSA is computationally intensive. There have been many improvements over the direct method to improve the efficiency, such as Tau-leaping method [4] etc.

B. Reaction Diffusion Master Equation

The dynamics of spatially inhomogeneous stochastic system is governed by the reaction-diffusion master equation (RDME), developed in the early works of Gardiner [1]. To apply a similar strategy as the SSA method, the spatial domain for inhomogeneous system is partitioned into voxels such that species within each voxel are considered well-stirred.

Assume the domain Ω is partitioned into K voxels V_k , $k = 1, 2, \dots, K$. For simplicity, we assume at this moment that the space Ω is one dimensional (1D). Each molecular species in the domain is represented by the state vector $X_i(t) = (X_{i,1}(t), X_{i,2}(t), \dots, X_{i,K}(t))$, where $X_{i,k}(t)$ is the number of molecules of species S_i in the voxel V_k at time t . Molecules in a voxel can react with molecules within the same voxel, and diffuse between neighboring voxels. The dynamics of diffusion of species S_i from voxel V_k to V_j is characterized by the *diffusion propensity function* $d_{i,k,j}$ and the *state change vector* $\mu_{k,j}$, where $\mu_{k,j}$ is a vector of length K with -1 in the k th position and 1 in the j th position and 0 everywhere else, and $d_{i,k,j}(x)dt$ gives the probability that, given $X_{i,k}(t) = x$, one copy of species S_i at voxel V_k diffuses into voxel V_j in the next infinitesimal time interval $[t, t + dt)$. If $j = k \pm 1$, then $d_{i,k,j}(x) = D/l^2$, where D is the diffusion rate and l is the characteristic length of voxel; Otherwise $d_{i,k,j} = 0$.

Similar to the CME, the diffusion dynamics can be expressed by diffusion master equation (DME).

$$\begin{aligned} & \frac{\partial P(x, t|x_0, t_0)}{\partial t} \\ = & \mathcal{D}P(x, t|x_0, t_0), \\ = & \sum_{i=1}^N \sum_{k=1}^K \sum_{j=1}^K [-d_{i,k,j}(x_i)P(x, t|x_0, t_0) + \\ & d_{i,k,j}(x_i - \mu_{k,j})P(x_1, \dots, x_i - \mu_{k,j}, \dots, x_N, t|x_0, t_0)], \end{aligned} \quad (4)$$

where \mathcal{D} denotes generating matrix for the Markov chain that describes the diffusion of molecules in the system. The usual method of solution of the DME is to simulate each diffusive jump event explicitly. This is the method used by the ISSA and NSM [6] algorithms. Combining the CME and DME yields the reaction-diffusion master equation (RDME)

$$\frac{\partial P(x, t|x_0, t_0)}{\partial t} = \mathcal{R}P(x, t|x_0, t_0) + \mathcal{D}P(x, t|x_0, t_0). \quad (5)$$

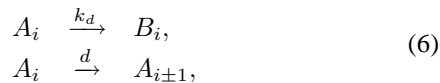
The RDME has many more possible states than the corresponding CME. Thus, it is more difficult to solve. Many techniques for accelerating the SSA can be applied to the ISSA. Much effort has been focused on the improvement of the ISSA, but ISSA remains computationally expensive. The problem is that fast diffusive movements between adjacent voxels dominate the computation time.

III. DISCRETIZATION SIZE IN ONE DIMENSION

A. Theoretical Analysis

In this section, we will present theoretical analysis of optimal discretization length for a reaction-diffusion system. For the convenience of discussion, we will use a simple one dimensional model. We note that the discussion here can be easily extended to a general case.

Suppose a biochemical model with a 1D spatial domain of size L . Species A is the only reactive species in this model, which gives a stoichiometry product B . A diffuses within this 1D space, while B stays where it is generated. Suppose the whole 1D space is discretized with voxels with length l . The reaction firing in the i -th cell is specified with a subscript i . The product B is generated at the place where A is. The diffusion process can be expressed as chemical reactions across the voxels. The reaction schema can be expressed as follows for this simple model.



where $d = D/l^2$ and D is the diffusion rate for species A . By the finite-difference schema, the propensity function for the reaction diffusion system (6) can be formulated as

$$\begin{aligned} a_1(A_i) &= k_d A_i, \\ a_2(A_i) &= D \frac{A_{i+1} - 2A_i + A_{i-1}}{l^2}, \end{aligned} \quad (7)$$

where A_i denotes the population of A at the i -th bin and l is the length of the bin.

The assumption of the discretization requires that the lengths of the bins be small enough, such that species within the bins can be considered well stirred. Previous work [12] has shown that this criterion is equivalent to

$$\frac{\tau_r}{\tau_d} \gg 1, \quad (8)$$

where τ_r is the mean free time with respect to the reactive collision and τ_d denotes the mean free time, during which a molecule will remain within a bin. For the first order

degradation with reaction rate k_d , the mean life time of a molecule can be expressed as $\tau_r = \frac{1}{k_d}$. The diffusion can also be considered as a first order reaction with the reaction rate constant $d = D/l^2$. Hence, the mean free time for a molecule staying within a bin can be expressed as $\tau_d = l^2/D$. Kuramoto's criterion (8) can be rewritten as

$$\frac{\tau_r}{\tau_d} = \frac{D}{k_d l^2} \gg 1, \quad \text{or} \quad l \ll \sqrt{\frac{D}{k_d}}. \quad (9)$$

Note that the discretization size l have a lower boundary $l \gg l_0$, where l_0 is the mean intermolecular distance. l_0 does not depend on specific type of the chemical reactions. Equation (9) gives a large upper bound for general cases. Here we will derive a more specified expression for the ideal discretization length.

For spatially inhomogeneous systems, the discretization of the space aims to result in a set of homogeneously populated voxels, where in each voxel the CME is applicable. Apparently, the smaller the voxel size is, the more accurate the resulted system will be. However, a small voxel size often results in large propensities for species jumping to neighboring voxels, but the heavy computational cost on the back-and-forth jumping makes little contribution to the actual population distribution evolution. Thus, it is important to find a optimal voxel size that is as large as possible while still maintaining a reasonable simulation error.

The ideal discretization should be based on the assumption that molecules of a species within a voxel are well-stirred, such that any two copies of that species within a voxel have similar probability distributions before one of them fires a reaction. Suppose we have two molecules located at positions $x = 0$ and $x = l$ initially. After diffusing for time t , the probability distribution of the two molecule's position can be solved from the following equation:

$$\frac{\partial u_i}{\partial t} = D \frac{\partial^2 u_i}{\partial x^2}, \quad \text{where } i = 1, 2, \quad (10)$$

$$\begin{aligned} \text{initial condition } u_1(x, 0) &= \delta(x), \\ u_2(x, 0) &= \delta(x-l), \end{aligned}$$

where $\delta(x)$ is the Dirac delta function, with $\int_{-\infty}^{\infty} \delta(x) dx = 1$. The solutions to the two diffusion functions are

$$\begin{aligned} u_1(x, t) &= \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}}, \\ u_2(x, t) &= \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{(x-l)^2}{4Dt}}. \end{aligned} \quad (11)$$

The difference of two probability distribution functions can be calculated by the Kullback-Leibler divergence (K-L divergence). In probability theory, the Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distribution functions. For probability distribution functions P and Q , the K-L divergence is defined as the integral:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx, \quad (12)$$

where $p(x)$ and $q(x)$ denote the probability density functions of P and Q . Thus the difference of the two distributions u_1, u_2 can be formulated as:

$$\begin{aligned}
 & D_{KL}(u_1||u_2) \\
 &= \int_{-\infty}^{\infty} u_1(x) \ln \frac{u_1(x)}{u_2(x)} dx, \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi Dt}} e^{-x^2/(4Dt)} \ln \frac{\frac{1}{\sqrt{4\pi Dt}} e^{-x^2/(4Dt)}}{\frac{1}{\sqrt{4\pi Dt}} e^{-(x-l)^2/(4Dt)}} dx, \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi Dt}} e^{-x^2/(4Dt)} \left(\frac{l^2}{4Dt} - \frac{2lx}{4Dt} \right) dx, \\
 &= \frac{l^2}{4Dt}. \tag{13}
 \end{aligned}$$

From equation (13), it is apparent that large l causes large diffusion difference. We are concerned with the diffusion time scale between chemical reactions. Thus we use the mean life time with respect to the reaction. In the simple reaction diffusion model, the mean life time of reaction is $\tau_r = 1/k_d$. We require that the two molecules' diffusion probability distribution difference be smaller than a tolerable threshold. If we set this threshold as 5%, we will have an analytic solution for the critical discretization size:

$$l_c^2 = 0.05 \times 4D\tau_r = 0.20 \frac{D}{k_d}, \quad \text{or} \quad l_c \approx 0.45 \sqrt{\frac{D}{k_d}}. \tag{14}$$

If we are a little more conservative to set the divergence threshold as 1%, where we will achieve a more accurate and safer simulation. The critical discretization size with respect to the conservative threshold is:

$$l_c^2 = 0.01 \times 4D\tau_r = 0.04 \frac{D}{k_d}, \quad \text{or} \quad l_c = 0.2 \sqrt{\frac{D}{k_d}}. \tag{15}$$

In the paper we will use 5% as the threshold (and thus formula (14)). Equation (14) and (15) satisfy Kuramoto's boundary $l \ll \sqrt{\frac{D}{k_d}}$. And it specifies the “ \ll ” relationship. The discretization size smaller than l_c will provide a more accurate simulation with heavier computational expense. For a larger l , it will lead to larger simulation error with respect to the homogeneous assumption. Note that when the diffusion rate $D \rightarrow \infty$, from (14) and (15) we get $l_c \gg 0$. That is the “well-stirred” situation, and the ISSA reduces to the SSA.

In a general system, a species will be involved in many reactions. However, the optimal discretization size for this species still depends on the time scales of diffusion and reaction. Equation (14) and (15) can still be applied, although k_d will be the overall change rate of this species rather than the reaction rate for a simple reaction.

B. Numerical Experiment

Below we present the numerical experiment results for the simple reaction-diffusion model (6). The parameters for this simple model are given in table I.

We solve the chemical kinetics equations and simulate the stochastic simulation with different discretization sizes. Figure 1 gives the plot for the mean population within each

TABLE I
THE PARAMETER SET FOR THE ONE-VARIABLE MODEL

parameter	value
L	1.0
k_d	1.0
D	0.0005

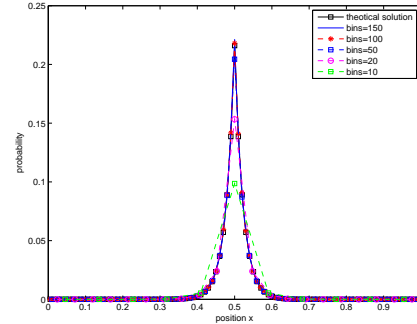


Fig. 1. Distribution density for B in the one-variable one-dimension simple model. The parameters are: total length $L=1.0$, $k_d = 1.0$, and $D = 0.0005$. The stochastic simulation is averaged over 10,000 runs.

voxel for this simple model. As we derived above, a good discretization size $l = \sqrt{0.2 \times \frac{D}{k_d}} = 0.01$ for the model parameter set, which requires the domain be discretized into about $L/l = 100$ voxels. We can see from the plot figure 1 that, When the discretization size reduces to 0.01, the stochastic result matches well to the theoretical result, and larger discretization sizes lead to greater errors. In this simple model, the 5% K-L divergence threshold is good enough to obtain an accurate simulation.

IV. MULTISCALE DISCRETIZATION

A. Theoretical Analysis

Typical biological systems contain multiple species and reactions. The reaction and diffusion rates may come across a wide scale. From equation (14) we can see that fast reaction (large k_d) and slow diffusion (small D) lead to small discretization size l . For a multispecies system, if a uniform discretization is used, l has to using the largest k_d and the smallest D , and a small discretization size l leads to heavy computational burden. Here with the ideal discretization size calculated in previous section, we propose a multiscale discretization method, which assigns a proper discretization size to each species, if necessary, depending on the diffusion rates and the chemical kinetics.

Consider a biochemical system with N species $\{S_1, \dots, S_N\}$ interacting through M reaction channels $\{R_1, \dots, R_M\}$. The domain Ω is partitioned differently for different species, according to the diffusion and reaction rate constants of each species. $\{h_1, \dots, h_N\}$ and $\{l_1, \dots, l_N\}$ are the specific discretization bin numbers and bin sizes for species $\{S_1, \dots, S_N\}$ respectively. Each species in the domain is represented by the state vector $X_i(t) = [X_{i,1}(t), \dots, X_{i,h_i}(t)]$, where $X_{i,k}(t)$ is the number

of molecules of species S_i in k -th voxel of that species at time t .

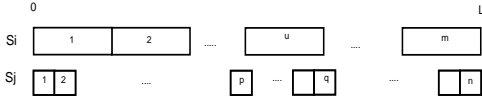


Fig. 2. The diagram for multiscale discretization. The total space have a length L . Species S_i is assigned m bins, while S_j n bins. Within the range of the u -th bin for S_i lies the p -th to q -th bins for S_j

With the multiscale discretization method, the RDME needs some modification to describe the newly formed system. The diffusion can still be modeled as the Brownian motion across neighboring bins, except that different species may diffuse across different-sized bins. The dynamics for diffusion of species S_i from k -th bin into the neighboring bins are characterized by the diffusion propensity function $d_{ik} = D/l_i^2$, where $d_{ik}dt$ gives the probability that one molecule of S_i at k -th voxel will jump into the neighboring bins in the infinitesimal time interval $[t, t + dt)$. More concerns must be put into the calculation of the reactive transformation propensity. With the multiscale discretization strategy, the chemical reaction between different species may not take place within the same bins. However, they may react when their resided bins overlap with each other. The reaction propensity function has to be modified accordingly, based on the assumption that each species in each of its own voxels must be homogeneous. The propensity functions for several typical reactions are given in Table II. Here we show two species S_i and S_j , where the u -th bin of S_i overlaps with the q -th bin of S_j , as in figure 2. Table II gives the propensity for some typical chemical reaction, along with the chemical kinetics reaction rates.

With multiscale discretization, we reduce the propensity for diffusion by increasing the discretization bin length to lower the computational cost. The total chemical reaction propensity and reaction firings number is not affected by the multiscale discretization.

B. Numerical Experiment

Our numerical experiment is based on a reaction-diffusion model of Turing pattern. In this model, the full domain length L is set to 1.3 unit length. The reactions schema and propensity functions in the domain are shown in equation (16), where P denotes polymer, M denotes monomer, and U denotes a catalyst that promotes the polymerization. The monomer is constantly synthesized and exponentially degraded. The monomers M polymerize to P with the help of the catalyst U . Catalyst U is constantly synthesized and degraded with a constant reaction rate k_{du} . The monomers, polymers, and catalysts diffuse within the whole domain length. At one end, there exist a binding site for catalyst U , which causes the catalyst level higher than the other end. Because of the inhomogeneity of the catalyst, the polymers and the monomers distribution are inhomogeneous too. The polymers, monomers,

TABLE II
PROPENSITIES FOR SOME TYPICAL CHEMICAL REACTION FOR THE SPECIES S_i IN u -TH VOXEL IN FIGURE 2 UNDER MULTISCALE DISCRETIZATION METHOD, ALONG WITH THE CHEMICAL REACTION RATE

reaction type	propensity $a_{iu}(X_{iu})dt$	reaction rate r
$null \rightarrow S_{iu}$	k_{syn}	k_{syn}
$S_{iu} \rightarrow null$	$k_{deg}X_{iu}$	$k_{deg}[X_{iu}]^a$
$S_{iu} \rightarrow S_{jl}^b$	k_1X_{iu}	$k_1[X_{iu}]$
$S_{iu} + S_{jl} \rightarrow S_k$	$\frac{k_a}{l_i} \sum_l X_{iu}X_{jl}$	$k_a[X_{iu}][X_{jl}]^d$

^a X_{iu} denotes the population number of S_i in u -th bin, while $[X_{iu}]$ denotes the concentration of S_i in u -th bin

^b l is the bins of S_j that overlap with i -th bin of S_i .

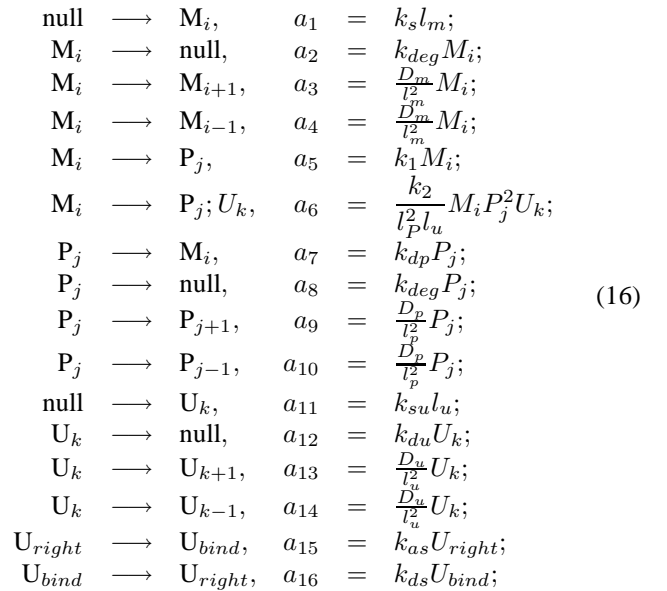
^c X_{jl} is the total population of S_j with the bins overlapping with u -th bin of S_i . For the bins of which only part of it are in that range, the population included in the formula is proportional to the length within the range

^dthe $[X_{jl}]$ is the average concentration of S_j within the length of the u -th bin of S_i

TABLE III
THE PARAMETER SET AND INITIAL CONDITIONS FOR THE 2-VARIABLE MODEL

parameter	value	parameter	value
k_s	9.1	D_p	0.1
k_{deg}	0.05	D_m	100
k_{dp}	5	D_u	1.0
k_1	10		0
k_2	0.01		
k_{su}	0.1	k_{du}	0.01
k_{as}	2	k_{ds}	0.001

and the catalysts diffuse at different rates, making it a simple model to test our multiscale discretization method.



The parameter values are in the table III. All the species population are set to zero as the starting point.

For this system, the mean life time τ_r of one species is proportional to the summation of propensities of all reactions involved with this species. The table IV shows the

TABLE IV
THE CRITICAL DISCRETIZATION BIN SIZE FOR THE 2-VARIABLE MODEL

Species	l_c	number of bins
M	0.18(0.009) ^a	7(150)
P	0.028	46
U	0.14	10

^a U only have a high population at right end, so the discretization size needs to be much smaller than the other bins

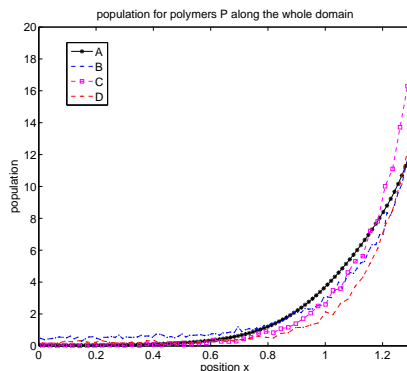


Fig. 3. The mean population over the whole domain for different discretization. A shows the deterministic solutions. B is the stochastic simulation result where the domain is partitioned into 100 bins for all species. C is for multiscale discretization, where the domain was equally partitioned for each single species. D is the result where P and U is uniformly partitioned into different bins, while M has a much smaller bins at the right end, due to the large reaction rate at the end. The other domain is equally partitioned into 10 bins for M .

critical discretization length for polymers P , monomers M and catalysts U , based on the parameter set. Polymers are spatially inhomogeneous. Besides, the monomer and catalyst with in the whole domain are always at very low level, 0 for most bins. The critical discretization size for M is a little complicated, because of the high order polymerization reaction with catalyst.

To run the Gillespie's algorithm over this model, it needs hundreds of reaction channels to depict the system. Here we applied the rule-based model technique, where the same reaction type of a species is grouped into one rule. By the rule based modeling, we not only simplify the the reaction propensity calculation, but also improve the time efficiency. A particle based method, network-free algorithm (NFA) [13] was proposed for the rule based model. The NFA calculates the propensity for a rule. The firing time and rule index are calculated as in Equation (3). For each rule, reactant candidate list are created to store all particles that satisfy the rule condition. After a rule is selected, reacting particles in that rule are selected from corresponding candidate list by generating uniform numbers to calculate the indices of the reacting particles.

To test the accuracy and time efficiency of this multiscale discretization method, we run the simulation for different discretization sizes. Figure 3 shows the average population distribution over the whole domain under some typical dis-

TABLE V
THE AVERAGE CPU TIME OVER 100 RUNS

	Multiscale discretization Model		
Mbins	100	50	10
Pbins	100	100	50
Ubins	100	50	10
CPU time	28min	11min	1min15s

TABLE VI
THE AVERAGE PERCENTAGE OF THE DIFFUSION RULE FIRINGS OVER 100 RUNS

	Multiscale discretization Model		
Mbins	100	50	10
Pbins	100	100	50
Ubins	100	50	10
$M_i \rightarrow i+1$	3.1e8(45.2%)	7.5e7(38.4%)	6.8e6(26.1.0%)
$M_i \rightarrow i-1$	3.1e8(45.0%)	7.4e7(38.0%)	6.5e6(26.8%)
$P_i \rightarrow i+1$	2.1e7(3.0%)	2.0e7(10.3%)	4.8e6(19.0%)
$P_i \rightarrow i-1$	2.1e7(3.2%)	2.2e7(11.2%)	5.2e6(22.34%)
$U_i \rightarrow i+1$	1.2e7(1.7%)	1.8e6(0.9%)	1.8e4(0.03%)
$U_i \rightarrow i-1$	1.2e7(1.7%)	1.8e6(0.8%)	1.9e4(0.04%)

cretization cases: plot A shows the deterministic solutions. Plot B is the stochastic simulation result where the domain is partitioned into 100 bins for all species. C is for multiscale discretization for different species, where the domain was uniformly partitioned for each single species. For plot D, P and U are uniformly partitioned into corresponding bins, but we use a nonuniform discretization strategy for M . Because the catalytic polymerization has a much larger propensity in the right end, we assign smaller bin size for M at the right end. From the result, even though the plot D has a larger discretization bin size, the result is still closer to the deterministic result. Besides, when we apply the nonuniformly discretization for M , the larger discretization size leads to less simulation CPU time. Table V shows the simulation time for this model under different discretization strategy.

The multiscale discretization technique does not affect the number of the reactive firings, only the diffusive transitions between the neighboring bins are reduced. Table VI shows the percentage of the diffusion rule firings. As it shows in the table, when the discretization bin length is twice large, the diffusive transitions are almost one quarter of the original firings.

For the other chemical reactions, the reactions firings keeps similar, only fluctuated over the stochastic effect. The average firings for different runs are listed in table VII.

From the statistics we collected, we can conclude that the

TABLE VII
THE AVERAGE NUMBER OF CHEMICAL REACTION FIRING OVER 25 RUNS

Rules	num of firings	Rules	num of firings
null \rightarrow M	2989.5	P \rightarrow null	2701.1
M \rightarrow null	46.5	M \rightarrow P; U	263977.5
M \rightarrow P	9190.5	null \rightarrow U	32.5
P \rightarrow M	270233.4	U \rightarrow null	5.75
U \rightarrow ur	25.7	ur \rightarrow U	1.8

multiscale discretization for different species, based on their different kinetics, can decrease the CPU time for stochastic simulation. A larger discretization bin size leads to smaller diffusion propensity between neighboring bins, which further shortens the simulation time.

V. CONCLUSION

In this work we have introduced a new method for efficient stochastic simulation of reaction-diffusion system. With the idea of critical discretization bin length, we assign each species with a proper bin length, which reduces the possible diffusion transition between neighboring bins. Larger discretization sizes may bring larger simulation errors. However, we can set a threshold for tolerable error and make a trade off between accuracy and efficiency.

The critical discretization bin length is easier to compute for the systems with only simple chemical reactions, such as first order reaction. The higher order reaction will make the case complicated. We believe the idea of multiscale discretization can be extended to 2-D or 3-D reaction-diffusion system simulation in a rather straightforward way.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under award CCF-0953590, and the National Institutes of Health under award GM078989.

REFERENCES

- [1] C. W. Gardiner, K. J. McNeil, D. F. Walls and I. S. Matheson, "Correlations in stochastic theories of Chemical Reactions" *J. Stat. Phys.* vol. 14, pp. 307–331. 1976
- [2] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions" *J. Chem. Phys.*, vol. 81, pp. 2340–2361. 1977.
- [3] M. Gibson and J. Bruck "Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels" *J. Chem. Phys. A*, vol. 104, pp. 1876–1889, 2000.
- [4] D. Gillespie, "Approximate accelerated stochastic simulation of chemically reacting systems" *J. Chem. Phys.*, vol. 115, pp. 1716–1733, 2001.
- [5] Y. Cao, D. Gillespie, L. Petzold "The slow-scale stochastic simulation algorithm" *J. Chem. Phys.* vol. 122, 014116, 2005.
- [6] Elf, J.; Ehrenberg, M.; , "Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases" *Systems Biology, IEE Proceedings*, vol. 1, ppp. 230–236. 2004
- [7] J. Hattne, D. Fange, and J. Elf "Stochastic reaction-diffusion simulation with MesoRD" *Bioinformatics* vol. 21, pp. 2923–2924, 2005.
- [8] B. Munsky, and M. Khammash; "The finite state projection algorithm for the solution of the chemical master equation" *J. Chem. Phys.* vol. 124, 044104, 2006.
- [9] S. Lampoudi, Dan T. Gillespie, and Linda R. Petzold, "The multinomial simulation algorithm for discrete stochastic simulation of reaction-diffusion systems" *J. Chem. Phys.* vol. 130, 094104, 2009.
- [10] T. T. Marquez-Lago and K. Burrage, "Binomial tau-leap spatial stochastic simulation algorithm for applications in chemical kinetics" *J. Chem. Phys.* vol. 127, 104101, 2007.
- [11] B. Drawert, M.J. Lawson, L. Petzold, M. Khammash "The diffusive finite state projection algorithm for efficient simulation of the stochastic reaction-diffusion master equation" *J. Chem. Phys.* vol. 132, 074101. 2010.
- [12] Y. Kuramoto, "Effects of diffusion on the fluctuations in open chemical system." *Progr. Theor. Phys.* vol. 52, pp. 711, 1974.
- [13] M. Sneddon, J. Faeder, and T. Emonet. "Efficient modeling, simulation and coarse-graining of biological complexity with NFsim" *Nature Methods*, vol. 8, pp. 177–183, 2011
- [14] C. Van Den Broeck, W. Horsthemke, M. Malek-Mansour, "On the diffusion operator of the multivariate master equation", *Physica A* vol. 89(2), pp. 339–352, 1997.
- [15] D. Rossinelli, B. Bayati, P. Koumoutsakos "Accelerated stochastic and hybrid methods for spatial simulations of reaction diffusion systems." *Chem. Phys. Lett.* vol. 415, pp. 136–140, 2008.

Quantum Cellular Combinatorics

- An Equilibrium or Non-Equilibrium Approach

Wen-Ran Zhang

Dept. of Computer Science, College of Engineering and Information Technology

Georgia Southern University, Statesboro, GA 30460

Email: wzhang@georgiasouthern.edu

Abstract - It is shown that YinYang bipolar dynamic logic and bipolar quantum linear algebra make quantum cellular combinatorics possible. Basic structures of the new type of combinatorics are introduced and discussed from an equilibrium and non-equilibrium perspective. These include YinYang-1-element, YinYang-2-element, ..., up to YinYang-n-element cellular networks. The utility of the new combinatorics in biological computing is highlighted. Philosophical distinctions of the new approach are drawn from existing approaches.

Keywords: YinYang Bipolar Quantum Cellular Combinatorics (QCC); Bipolar Dynamic Logic (BDL); Bipolar Quantum Linear Algebra (BQLA); YinYang-1-element; YinYang-2-element; YinYang-N-element

1 Introduction

Based on YinYang bipolar dynamic logic (BDL) and bipolar quantum linear algebra (BQLA) [Zhang 2011a], this paper introduces the concepts of quantum cellular combinatorics (QCC). Algebraic models and graphs of YinYang-1-element, YinYang-2-element, ..., up to YinYang-n-element cellular networks are presented from an equilibrium or non-equilibrium perspective. The utility of the new approach in biological computing is highlighted. Philosophical distinction of this work is drawn from existing approaches.

The remaining sections are organized as follows: Section 2 presents a review on BDL, logically definable causality, and bipolar quantum linear algebra. Section 3 presents the basic concepts and graphs of QCC. Section 4 examines the properties of different QCC structures. Section 5 draws a few conclusions.

2 Background

2.1 Bio-Quantum Computing

Despite the desire and efforts to bring quantum computing into the biological world, bio-quantum computing is still in its infant stage. A key barrier is the mystery of quantum entanglement. Although, numerous reported experimental successes in testing quantum entanglement have been reported, quantum non-local connection or entanglement remains logically unresolved. Many scientists believe that

something fundamental must still be missing from the big picture. The missing fundamental is often traced to the ultimate unknown cause-effect relationship in quantum entanglement. Without logically definable causality, quantum entanglement could be deemed something beyond the realm of science because, if Aristotle's causality principle is the doctrine of all sciences, there should be no science beyond the doctrine.

2.2 YinYang Bipolar Quantum Lattice and Bipolar Dynamic Logic (BDL)

Aristotle's causality principle became controversial in the 18th century after David Hume challenged it from an empirical perspective. Hume argued that causation is irreducible to pure regularity. Bipolar dynamic logic (BDL) has changed this situation in a fundamental way.

BDL is defined with Eq. (1)-(12) on a bipolar quantum lattice $B_1 = \{-1,0\} \times \{0,+1\}$ in background independent YinYang bipolar geometry (Fig. 1). It provides logically definable causality [Zhang 2011]. In B_1 , (0,0), (0,1), (-1,0), and (-1,1) stand, respectively, for eternal equilibrium, non-equilibrium, another non-equilibrium; and equilibrium. The laws in Fig. 2 hold on B_1 .

Bipolar Partial Ordering: $(x,y) \geq (u,v)$, iff $ x \geq u $ and $y \geq v$.	(1)
(Note: The use $ x $ through this paper is for explicit bipolarity only.)	
Complement: $\neg(x,y) \equiv (-1,1) - (x,y) \equiv (-x, -y) \equiv (-1-x, 1-y)$.	(2)
Implication: $(x,y) \Rightarrow (u,v) \equiv (x \rightarrow u, y \rightarrow v) \equiv (\neg x \vee u), \neg y \vee v)$.	(3)
Negation: $\neg(x,y) \equiv (-y, -x)$.	(4)
Bipolar least upper bound (blub):	
$blub((x,y),(u,v)) \equiv (x,y) \otimes (u,v) \equiv (x \vee u , y \vee v)$;	(5)
Bipolar greatest lower bound (bglb):	
$bglb((x,y),(u,v)) \equiv (x,y) \& (u,v) \equiv (-(x \wedge u), y \wedge v)$;	(6)
-blub: $blub^-(x,y),(u,v)) \equiv (x,y) \oplus (u,v) \equiv (-y \vee v), (x \vee u)$;	(7)
-bglb: $bglb^-(x,y),(u,v)) \equiv (x,y) \&^-(u,v) \equiv (-y \wedge v), (x \wedge u)$;	(8)
Cross-pole greatest lower bound (cglb):	
$cglb((x,y),(u,v)) \equiv (x,y) \otimes (u,v) \equiv (-(x \wedge u) \vee y \wedge v), (x \wedge u) \vee y \wedge v)$;	(9)
Cross-pole least upper bound (club):	
$club((x,y),(u,v)) \equiv (x,y) \oplus (u,v) \equiv (-1,1) - ((x,y) \otimes (u,v))$;	(10)
-cglb: $cglb^-(x,y),(u,v)) \equiv (x,y) \oplus (u,v) \equiv -(x,y) \otimes (u,v)$;	(11)
-club: $club^-(x,y),(u,v)) \equiv (x,y) \oplus (u,v) \equiv -(x,y) \otimes (u,v)$.	(12)

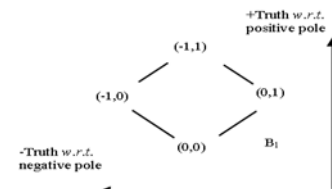


Fig. 1. Hasse diagrams of B_1 in YinYang bipolar geometry

Excluded Middle	$(x,y) \oplus \neg(x,y) \equiv (-1,1); (x,y) \oplus^- \neg(x,y) \equiv (-1,1);$
Non-contradiction	$\neg((x,y) \& \neg(x,y)) \equiv (-1,1);$ $\neg((x,y) \&^- \neg(x,y)) \equiv (-1,1);$
Linear Bipolar DeMorgan's Laws	$\neg((a,b) \& (c,d)) \equiv \neg(a,b) \oplus \neg(c,d);$ $\neg((a,b) \oplus (c,d)) \equiv \neg(a,b) \& \neg(c,d);$ $\neg((a,b) \&^- (c,d)) \equiv \neg(a,b) \oplus^- \neg(c,d);$ $\neg((a,b) \oplus^- (c,d)) \equiv \neg(a,b) \&^- \neg(c,d);$
Non-Linear Bipolar DeMorgan's Laws	$\neg((a,b) \otimes (c,d)) \equiv \neg(a,b) \oslash \neg(c,d);$ $\neg((a,b) \oslash (c,d)) \equiv \neg(a,b) \otimes \neg(c,d);$ $\neg((a,b) \otimes (c,d)) \equiv \neg(a,b) \oslash \neg(c,d);$ $\neg((a,b) \oslash (c,d)) \equiv \neg(a,b) \otimes \neg(c,d)$

Fig. 2. Bipolar laws

Unipolar Axioms (UAs): UA1: $\phi \rightarrow (\phi \rightarrow \psi);$ UA2: $(\phi \rightarrow (\phi \rightarrow \chi)) \rightarrow (\phi \rightarrow \psi) \rightarrow (\phi \rightarrow \chi);$ UA3: $\phi \rightarrow \psi \rightarrow ((\phi \rightarrow \psi) \rightarrow \phi);$ UA4: (a) $\phi \wedge \psi \rightarrow \phi;$ (b) $\phi \wedge \psi \rightarrow \psi;$ UA5: $\phi \rightarrow (\psi \rightarrow \phi \wedge \psi);$	Bipolar Linear Axioms: BA1: $(\phi, \psi) \rightarrow ((\phi, \psi) \rightarrow (\phi, \psi));$ BA2: $((\phi, \psi) \rightarrow ((\phi, \psi) \rightarrow (\chi, \chi))) \rightarrow ((\phi, \psi) \rightarrow (\phi, \psi)) \rightarrow ((\phi, \psi) \rightarrow (\chi, \chi));$ BA3: $(\neg(\phi, \psi) \rightarrow (\phi, \psi)) \rightarrow ((\neg(\phi, \psi) \rightarrow (\phi, \psi)) \rightarrow (\phi, \psi));$ BA4: (a) $(\phi, \psi) \& (\phi, \psi) \rightarrow (\phi, \psi);$ (b) $(\phi, \psi) \& (\phi, \psi) \rightarrow (\phi, \psi);$ BA5: $(\phi, \psi) \rightarrow ((\phi, \psi) \rightarrow ((\phi, \psi) \& (\phi, \psi)));$
Inference Rule – Modus Ponens (MP): UR1: $(\phi \wedge (\phi \rightarrow \psi)) \rightarrow \psi.$	Non-Linear Bipolar Universal Modus Ponens (BUMP) BR1: IF $((\phi, \psi) \rightarrow (\psi, \psi)),$ $[(\phi, \psi) \rightarrow (\phi, \psi)] \& ((\psi, \psi) \rightarrow (\chi, \chi)),$ THEN $[(\phi, \psi) \rightarrow (\chi, \chi)];$
Predicate axioms and rules UA6: $\forall x, \phi(x) \rightarrow \phi(t);$ UA7: $\forall x, (\phi \rightarrow \psi) \rightarrow (\phi \rightarrow \forall x, \psi);$ UR2-Generalization: $\phi \rightarrow \forall x, \phi(x)$	Bipolar Predicate axioms and Rules of inference BA6: $\forall x, (\phi(x), \psi(x)) \rightarrow (\phi(t), \psi(t));$ BA7: $\forall x, (\phi, \psi) \rightarrow (\phi, \psi) \rightarrow (\phi, \psi) \rightarrow \forall x, (\phi, \psi);$ BR2-Generalization: $(\phi, \psi) \rightarrow \forall x, (\phi(x), \psi(x))$

Fig. 3(a) Bipolar axiomatization

$\forall \phi = (\phi, \psi^+), \psi = (\psi^+, \psi^+), \psi = (\psi^-, \psi^+),$ and $\chi = (\chi, \chi^+) \in B_1,$ $[(\phi \Rightarrow \psi) \& (\psi \Rightarrow \chi)] \Rightarrow [(\phi * \psi) \Rightarrow (\psi * \chi)].$
Two-fold universal instantiation:
1) Operator instantiation: * as a universal operator can be bound to $\&, \oplus, \&^-, \oplus^-, \oslash, \otimes, \otimes^-, \oslash^-.$ ($\phi \Rightarrow \psi$) is designated (bipolar true); $((\phi^-, \psi^+) * (\psi^-, \psi^+))$ is undesignated.
2) Variable instantiation: $\forall x, (\phi^-, \psi^+)(x) \Rightarrow (\psi^-, \psi^+)(x); (\phi^-, \psi^+)(A); \therefore (\psi^-, \psi^+)(A).$

Fig. 3(b). Bipolar Universal Modus Ponens (BUMP)

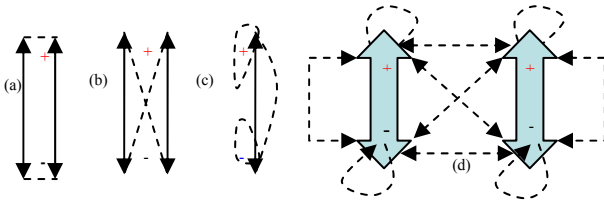


Fig. 4. (a) Linear; (2) Cross-pole; (c) Oscillatory; (d) Entangled

An axiomatization of BDL (Fig. 3) has been proven sound and complete [Zhang & Zhang 2004; Zhang 2005, 2011]. The key element in the axiomatization is bipolar universal modus ponens (BUMP) which is a bipolar tautology, an equilibrium-based non-linear bipolar dynamic generalization of classical modus ponens (MP) and a logical representation of bipolar quantum entanglement.

BDL generalizes Boolean logic to a quantum logic where \oplus and \oplus^- are “balancers”; \oslash and \otimes are intuitive “oscillators”; \oslash^- and \otimes^- are counter-intuitive “oscillators”; $\&$ and $\&^-$ are “minimizers.” The linear, cross-pole, bipolar fusion, oscillation, interaction/entanglement properties are depicted in Fig. 4. Based on BDL, bipolar equilibrium relations, bipolar linear algebra (BLA), bipolar cellular networks [Zhang et al. 2009], bipolar quantum computing have been presented [Zhang 2011]. Most interestingly,

equilibrium-based bipolar causality is now logically definable¹.

2.3 Bipolar Causality and Relativity

BUMP makes bipolar causality logically definable in physical terms. It simply states: For all bipolar equilibrium functions $\phi, \psi, \chi,$ IF $(\phi \Rightarrow \psi) \& (\psi \Rightarrow \chi),$ THEN the bipolar interaction $(\phi * \psi)$ implies that of $(\psi * \chi).$ With the emergence of space and time, BUMP leads to a complete background independent theory of YinYang bipolar relativity defined by Eq. (13) and a partial solution to Hilbert’s Problem 6 [Zhang 2011].

$$\forall a, b, c, d, \psi(a(t_x, p_1)) \Rightarrow \chi(c(t_y, p_3)) \& [\phi(b(t_x, p_2)) \Rightarrow \psi(d(t_y, p_4))] \Rightarrow [\psi(a(t_x, p_1)) * \phi(b(t_x, p_2)) \Rightarrow \chi(c(t_y, p_3)) * \psi(d(t_y, p_4))]. \quad (13)$$

In Eq. (13), $a(t_1, p_1), b(t_1, p_2), c(t_2, p_3), d(t_2, p_4)$ are any bipolar agents where $a(t, p)$ stands for “agent a at time t and space p ” (t_x and t_y can be the same or different points in time and p_x and p_y can be the same or different points in space). An agent without time and space is assumed at any time t and space $p.$ An agent at time t and space p is therefore more specific. Time and/or space can be omitted in some discussion for simplicity.

2.4 Bipolar Quantum Linear Algebra (BQLA)

The bipolar lattice $B_I = \{-1, 0\} \times \{0, 1\}$ and bipolar fuzzy lattice $B_F = [-1, 0] \times [0, 1]$ can be naturally extended to the infinite bipolar lattice $B_\infty = [-\infty, 0] \times [0, +\infty].$ While B_I and B_F are bounded complemented unit square crisp/fuzzy lattices, respectively, B_∞ is unbounded. $\forall (x, y), (u, v) \in B_\infty,$ two major operations can be defined as shown in Eq. (4a) and (4b).

Tensor Bipolar Multiplication:

$$(x, y) \times (u, v) \equiv (xv + yu, xu + yv); \quad (14a)$$

Bipolar Addition:

$$(x, y) + (u, v) \equiv (x + u, y + v). \quad (14b)$$

In Eq. (14a), \times is a bipolar cross-pole multiplication operator with the infused non-linear bipolar tensor semantics of $--=+, -+=-1,$ and $++=+;$ in Eq. (14b) $+$ is a linear bipolar addition or fusion operator. With the two basic operations, classical linear algebra is naturally extended to BQLA with bipolar fusion, diffusion, interaction, oscillation, and quantum entanglement properties. These properties enable biological agents to interact through bipolar bioelectromagnetic fields such as atom-atom, cell-cell, heart-heart, heart-brain, brain-brain, organ-organ, and genome-genome bio-electromagnetic quantum fields as well as biochemical pathways in energy equilibrium or non-equilibrium. Thus, the bipolar properties are suitable for equilibrium/non-equilibrium based bipolar dynamic modeling with quantum aspects where one kind of equilibrium/non-equilibrium can have causal effect to another such as I/O energy equilibrium/non-equilibrium.

¹ Causality discussed in this work meant to be those with bounded cause and effect. Of course, causality is undefinable without bounded cause or effect.

Given an input bipolar row vector matrix $E=[e_i]=[e_i^-, e_i^+]) \in B_\infty$, $i=1,2,\dots,k$, and a bipolar connectivity matrix $M=[m_{ij}]=[m_{ij}^-, m_{ij}^+]$, $i=1,2,\dots,k$ and $j=1,2,\dots,n$, we have $V=E \times M=[V_j]=[v_j^-, v_j^+]$. While E is the input vector to a dynamic system characterized with the connectivity matrix M , V is the result row vector with n bipolar elements.

$$V_j = \sum_{i=1}^k (e_i \times m_{ij}). \quad (15)$$

Eq. (15) has the same form as in classical linear algebra except for: (i) e_i and m_{ij} are bipolar elements; (ii) the multiplication operator is defined in Eq. (14a) on bipolar variables with bipolar (quantum) entanglement; and (iii) the \sum operator is based on the addition operation defined on bipolar variables in Eq. (14b).

BQLA provides a new mathematical tool for modeling YinYang-n-element or YinYang-n-element cellular networks with explicit YinYang representation and equilibrium, quasi- or non-equilibrium states for energy and stability analysis. In this case, energy in a row matrix can be considered as biological energy of biological elements or agents such as energy for repression and activation of regulator proteins [Shi et. al 1991]; energy embedded in a connectivity matrix can be considered organizational energy of the biological agents such as the bipolar capacities of biological pathways.

YinYang Bipolar Elementary Energy. Given a bipolar element $e=(e^-, e^+)$,

- (i) $\varepsilon^-(e) = e^-$ is the *Yin or negative energy of e*;
- (ii) $\varepsilon^+(e) = e^+$ is the *Yang or positive energy of e*;
- (iii) $\varepsilon(e) = (\varepsilon^-(e), \varepsilon^+(e)) = (e^-, e^+)$ is the *YinYang bipolar energy measure of e*;
- (iv) The absolute total $|\varepsilon|(e) = |\varepsilon^-(e)| + |\varepsilon^+(e)|$ is the *total energy of e*;
- (v) $\varepsilon_{imb}(e) = |\varepsilon^+(e)| - |\varepsilon^-(e)|$ is the *imbalance of e*;
- (vi) $EnergyBalance(e) = (|\varepsilon|(e) - |\varepsilon_{imb}(e)|)/2.0 = \min(|e^-|, |e^+|)$;
- (vii) $Harmony(e) = Balance(e) = (|\varepsilon|(e) - |\varepsilon_{imb}(e)|)/|\varepsilon|(e)$.

YinYang Bipolar System Energy. Given an $k \times n$ bipolar matrix $M=[m_{ij}]=[M^-, M^+] = ([m_{ij}^-], [m_{ij}^+])$, where M^- is the *Yin half* with all the negative elements and M^+ is the *Yang half* with all the positive elements,

- (i) $\varepsilon^-(M) = \sum_{i=1}^k \sum_{j=1}^n \varepsilon_{ij}^- = \sum_{i=1}^k \sum_{j=1}^n m_{ij}^-$ is the *negative or Yin energy of M*;
- (ii) $\varepsilon^+(M) = \sum_{i=1}^k \sum_{j=1}^n \varepsilon_{ij}^+ = \sum_{i=1}^k \sum_{j=1}^n m_{ij}^+$ is the *positive or Yang energy of M*;
- (iii) the polarized total, denoted $\varepsilon(M) = (\varepsilon^-(M), \varepsilon^+(M))$ is the *YinYang bipolar energy of M of M*;
- (iv) the absolute total, denoted $|\varepsilon|(M) = |\varepsilon^-(M)| + |\varepsilon^+(M)|$, is the *total energy of M*;

- (v) the energy subtotal for row i of M is denoted

$$|\varepsilon|(M_{i*}) = \left| \sum_{j=1}^n \varepsilon_{ij} \right|;$$

- (vi) the energy subtotal for column j of M is denoted $|\varepsilon|(M_{*j})$

$$= \left| \sum_{i=1}^k \varepsilon_{ij} \right|;$$

- (vii) $\varepsilon_{imb}(M) = \sum_{i=1}^k \sum_{j=1}^n \varepsilon_{imp}(m_{ij}) = \sum_{i=1}^k \sum_{j=1}^n (m_{ij}^+ - |m_{ij}^-|)$ is the

YinYang imbalance of M;

- (viii) balance or harmony or stability of M is defined as $Harmony(M) = Balance(M) = Stability(M) = (|\varepsilon|(M) - |\varepsilon_{imb}(M)|)/|\varepsilon|(M)$;

- (ix) the *average energy of M* is measured as $h = (\varepsilon^-(M)/(kn), \varepsilon^+(M)/(kn))$ where $kn=k \times n$ is the total number of elements in M .

Law 1. Elementary Energy Equilibrium Law.

$\forall (x,y) \in B_\infty = [-\infty, 0] \times [0, +\infty]$ and $\forall (u,v) \in B_F = [-1, 0] \times [0, 1]$, we have

- (a) $[|\varepsilon|(u,v) \equiv 1.0] \Rightarrow [|\varepsilon|((x,y) \times (u,v)) \equiv |\varepsilon|(x,y)]$;
- (b) $[|\varepsilon|(u,v) < 1.0] \Rightarrow [|\varepsilon|((x,y) \times (u,v)) < |\varepsilon|(x,y)]$;
- (c) $[|\varepsilon|(u,v) > 1.0] \Rightarrow [|\varepsilon|((x,y) \times (u,v)) > |\varepsilon|(x,y)]$.

Equilibrium/Non-Equilibrium System. A bipolar dynamic cellular system S is said an *equilibrium system* if the system's total energy $|\varepsilon|S$ remains in an equilibrium state or $d(|\varepsilon|S)/dt=0$ without external disturbance. Otherwise it is said a *non-equilibrium system*. A non-equilibrium system is said a *strengthening system* if $d(|\varepsilon|S)/dt > 0$; it is said a *weakening system* if $d(|\varepsilon|S)/dt < 0$.

Law 2. Energy Transfer Equilibrium Law. Given an $n \times n$ input bipolar matrix $E=[e_{ik}]=[e_{ik}^-, e_{ik}^+]$, $0 < i, k \leq n$, an $n \times n$ bipolar connectivity matrix $M=[m_{kj}]=[m_{kj}^-, m_{kj}^+]$, $0 < k, j \leq n$, and $V=E \times M=[V_j]=[v_j^-, v_j^+]$, $\forall k, j$, let $|\varepsilon|(M_{k*})$ be the k -th row energy subtotal and let $|\varepsilon|(M_{*j})$ be the j -th column energy subtotal, we have, $\forall k, j$,

- (a) $[|\varepsilon|(M_{k*}) \equiv |\varepsilon|(M_{*j}) \equiv 1.0] \Rightarrow [|\varepsilon|(V) \equiv |\varepsilon|(E)]$;
- (b) $[|\varepsilon|(M_{k*}) \equiv |\varepsilon|(M_{*j}) < 1.0] \Rightarrow [|\varepsilon|(V) < |\varepsilon|(E)]$;
- (c) $[|\varepsilon|(M_{k*}) \equiv |\varepsilon|(M_{*j}) > 1.0] \Rightarrow [|\varepsilon|(V) > |\varepsilon|(E)]$.

From the above definitions and laws it is clear that without YinYang bipolarity, classical linear algebra cannot deal with the coexistence of the Yin and the Yang of bipolar elements and their interactions and quantum entanglement.

Law 3. Law of Energy Symmetry (Zhang et al. 2009). Let $t=0, 1, 2, \dots$, $Y(t+1)=Y(t) \times M(t)$, $|\varepsilon|Y(t)$ be the total energy of an YinYang-N-Element vector $Y(t)$, $|\varepsilon|M(t)$ be the total energy of the connectivity matrix $M(t)$, $|\varepsilon|M_{i*}(t)$ be the energy subtotal of row i of $M(t)$, $|\varepsilon|M_{*j}(t)$ be the energy subtotal of column j of $M(t)$.

- 1) Regardless of the local YinYang balance/imbalance of the elements at any time point t , the system will remain a global energy equilibrium if, $\forall t$, $d(|\varepsilon|Y(t))/dt \equiv 0$, or

- (a) $\forall i, j, [|\varepsilon|(M_{i^*}) \equiv |\varepsilon|(M_{j^*}) \equiv 1.0]$ and (b) no external disturbance to the system after the initial vector $Y(0)$ is given.
- 2) Under the same conditions of (1), if, $\forall t, |\varepsilon^-(M_{i^*})| > 0$ and $|\varepsilon^+(M_{j^*})| > 0$, all bipolar elements connected by M will eventually reach a local YinYang balance $(-|\varepsilon|Y(t)/(2N), |\varepsilon|Y(t)/(2N))$ at time t .

Law 4. Law of Broken Symmetry (Growing) (Zhang *et al.* 2009). For the same system as for Law 3, if, $\forall i, j, |\varepsilon|(M_{i^*}) \equiv |\varepsilon|(M_{j^*}) > 1.0$, regardless of the local YinYang balance/imbalance of the elements at any time point t , the system energy will increase and eventually reach a bipolar infinite $(-\infty, \infty)$ state without external disturbance or we have, $\forall t, d(|\varepsilon|Y(t))/dt > 0$.

Law 5. Law of Broken Symmetry (Weakening) (Zhang *et al.* 2009). For the same system as for Law 3, if, $\forall i, j, |\varepsilon|(M_{i^*}) \equiv |\varepsilon|(M_{j^*}) < 1.0$, regardless of the local YinYang balance/imbalance of the elements at any time point t , the system energy will decrease and eventually reach a $(0,0)$ state without external disturbance or we have, $\forall t, d(|\varepsilon|Y(t))/dt < 0$, until $|\varepsilon|Y(t) = 0$.

3 Bipolar Quantum Cellular Combinatorics

3.1 An Equilibrium/Non-Equilibrium Approach

Combinatorics is a branch of mathematics concerning the study of finite or countable discrete structures. Aspects of combinatorics include counting the structures of a given kind and size, deciding when certain criteria can be met, and constructing and analyzing objects meeting the criteria, finding "largest", "smallest", or "optimal" objects, and studying combinatorial structures arising in an algebraic context, or applying algebraic techniques to combinatorial problems (algebraic combinatorics).

Combinatorial problems arise in many areas of pure mathematics and also have many applications. One of the oldest and most accessible parts of combinatorics is graph theory, which also has numerous natural connections to other areas. Combinatorics is used frequently in computer science to obtain formulas and estimates in the analysis of algorithms.

BDL and BQLA provide a new mathematical basis for bipolar quantum combinatorics. While existing combinatorics is truth-based, the new approach is equilibrium or non-equilibrium based focused on the negative or positive energies [Hawking and Mlodinow 2010] or the Yin and Yang of nature [Gore & van Oudenaarden, 2009] [Shi *et al.* 1991] [Zhang and Chen 2008][Zhang *et al.* 2009][Zhang 2011]. Based on BDL and BQLA, in this work we discuss the graphical aspects and their quantum cellular properties.

3.2 Combinatorial YinYang-1-Element Graph

Fig. 5 shows the structure of a YinYang-1-element as the most basic structure of QCC. This element seems to be rather simple. But a closer examination reveals its

quintessential role as the smallest and, at the same time, the largest structure in the new type of combinatorics for quantum cellular computing.

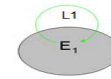


Fig. 5. YinYang-1-Element

First, we consider it as a smallest equilibrium or non-equilibrium structure. In this case, it can be used as a model for a particle-antiparticle pair variable $E_1 = (e^-, e^+)$ or an energy input-output variable. For instance, if $(e^-, e^+) = (-1, 0)$ it can represent an electron or non-equilibrium; if $(e^-, e^+) = (0, +1)$ it can represent a positron or another non-equilibrium; if $(e^-, e^+) = (-1, +1)$ it can represent an electron-positron pair or an energy equilibrium; if $(e^-, e^+) = (0, 0)$ it can represent an annihilation of the pair or eternal equilibrium.

Interestingly, the reflexive link $L1$ is also bipolar that can add dynamic change or mutation to the basic structure. For instance, when n is odd we have $(-1, 0) \otimes (-1, 0) \otimes \dots \otimes (-1, 0) = (-1, 0)^n = (-1, 0)$ and when n is even $(-1, 0)^n = (0, +1)$. This property seems rather bizarre but it can represent the most fundamental natural or biological processes in microscopic as well as macroscopic worlds. For instances, a subatomic particle can change polarity trillion times per second [Fermilab 2006]; some genetic agent exhibits YinYang bipolar repression-activation $(-, +)$ abilities in gene expression regulation [Shi *et al.* 1991].

Secondly, we consider YinYang-1-element as the largest equilibrium or non-equilibrium. Evidently, our universe can switch from big bang $(0, +1)$ state to a black hole state $(-1, 0)$. Interestingly, it may also switch from a black hole state to a big bang state. In that case, our universe would be a cyclic process where space and time would both be curved. Remarkably, we have $(-1, 0) \oplus (-1, 0)^2 = (-1, 0) \oplus (0, +1) = (-1, +1)$ which shows a self-adaptation to equilibrium.

Thirdly, we consider YinYang-1-element as a medium-sized equilibrium or non-equilibrium. This may sound impossible. But, evidently, a person's mind can be depression, mania, equilibrium, eternal equilibrium or between. Actually, all human beings have to be in either mental equilibrium or non-equilibrium or between. To certain extent, we are all mentally bipolar, either in equilibrium or disorder or between because no one's mind can escape equilibrium or non-equilibrium and bipolar equilibrium/non-equilibrium is most fundamental.

3.3 Combinatorial YinYang-2-Element Graph

Fig. 6 shows the structures of a YinYang-2-element as the 2nd most basic structure of QCC for bipolar interaction. Based on YinYang-1-element these structures added two more bipolar links between the two bipolar elements in equilibrium or non-equilibrium (green: harmonic; red: positive; blue: negative). The two more directed links can be simplified to L_3 as shown in Fig. 6(d). The link weight of L_3 can be any (x, y) in B_∞ . For instance, where $(0, 0)$ shows no interaction; $(-1, 0)$ shows conflict or inhibition to each other;

(0,+1) shows coalition or excitation to each other; (-1,+1) shows harmonic interaction. As a basic combinatorial structure for equilibrium or non-equilibrium bipolar interaction, YinYang-2-element is critical in characterizing bipolar quantum entanglement for building larger combinatorial networks.

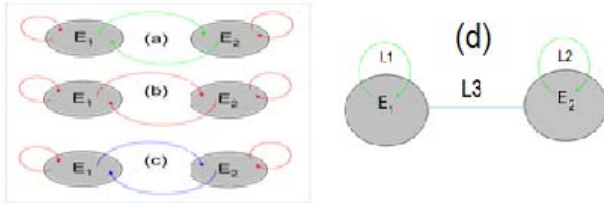


Fig. 6. YinYang-2-Element

3.4 YinYang-N-Element Structures

Fig. 7 shows YinYang-3-element structures; Fig. 8 shows YinYang-4-element structures; Fig. 9 shows an YinYang-5-element structure. Fig. 10 shows YinYang-n-element structures. The 3-element and 4-element structures both show some interesting properties that deserve further investigation. The YinYang-5-element structure is historically prominent in Chinese cosmology and traditional Chinese medicine [Zhang and Chen, 2008; Zhang 2011]. The YinYang-n-element structure is central in QCC and further discussed in the next section.

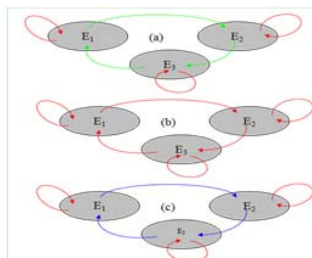


Fig. 7. YinYang-3-element

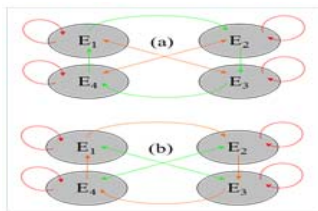


Fig. 8. YinYang-4-element

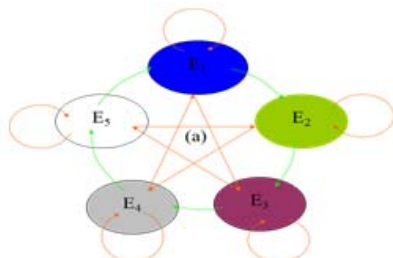


Fig. 9. YinYang-5-element

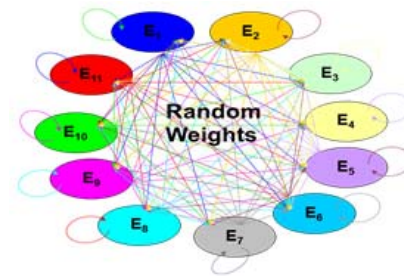


Fig. 9. YinYang-n-element

4 Properties of Bipolar Quantum Cellular Combinatorics

4.1 Quantum Cellular Properties

The YinYang-n-element structure is essential and general in QCC. First, it is quantum in nature due to bipolar quantum entanglement. Secondly, it is well defined based on BQLA and Laws 1-5. Thirdly, it forms a basis for an equilibrium or non-equilibrium-based computing paradigm or a theory of automata. Fourthly, it is scalable for upward integration. Fig. 10 shows such an integration with well-defined properties. The random link weights can be optimized for different applications [Jaeger, Chen & Zhang 2009] [Zhang 2011a].

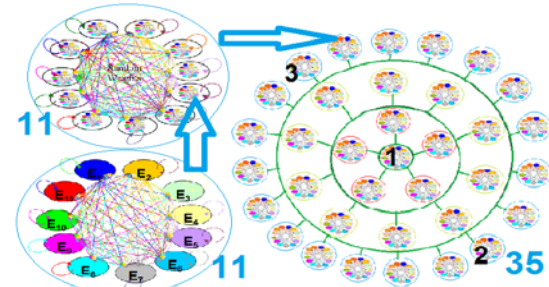


Fig. 10. Integration of YinYang-N-Elements

QCC shows a number of unifying properties. These include particle-wave unification, matter-antimatter unification and quantum-cellular unification [Zhang 2011a,b]. Fig. 11 shows a combinatorial unification of matter and antimatter atoms with YinYang-n-element bipolar quantum cellular automata [Zhang 2011b].

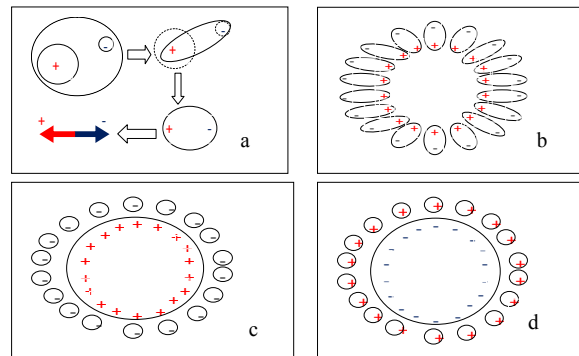


Fig. 11 (a) Bipolar representation of a hydrogen; (b) YinYang-N-Elements; (c) Matter atom; (d) Antimatter atom

Stimulation of disturbance can be visualized as quantum entangled particle-waves with energy symmetry (e.g. Fig. 12). It can be shown that any bipolar harmonic wave can be generated with BQLA. This property makes it possible for system modeling of biological and neural networks to achieve bipolar equilibrium, non-equilibrium, harmony or disharmony [Zhang, Chen & Bezdek 1989] [Zhant et. al 1992] [Zhang 2003] [Zhang, Pandurangi and Peace 2007] [Jaeger, Chen & Zhang 2009] [Zhang et al 2009] [Shi et. al 1991] [Jacobsen & Skalnik 1999] [Ai, Narahari & Roman 2000] [Palko et. al 2004][Wilkinson, Park & Atchison 2006] [Liu et. al 2007] [Vasudevan, Tong and Steitz 2007]

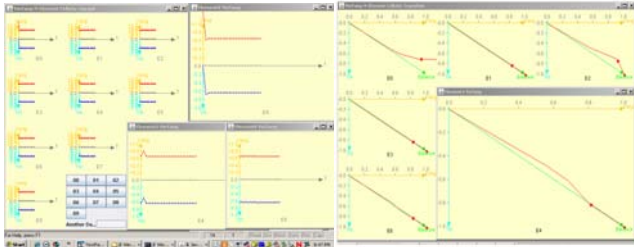


Fig. 12. YinYang particle-wave forms of energy rebalance after a disturbance to one element

It sounds like an unbelievable hype but it is not because nothing in the universe including the universe itself can escape from equilibrium or non-equilibrium whose bipolar forms are the most forms that lead to logically definable causality. For instances, (1) in a black hole all truth will be gone but particle-antiparticle bipolarity will miraculously survive due to Hawking radiation or particle/antiparticle emission; (2) every living being must have input and output energy; (3) without bipolar mental equilibrium we would all be in bipolar disorder and there would be no truth [Zhang et. al 2011]. From (3) it can be asserted that, to a certain extent, equilibrium or non-equilibrium is a unifying property of mind-body.

4.2 Philosophical Distinctions

Despite the continuing debate among scientists and philosophers on various theories regarding the meaning of truth, Western philosophy is being-centered and truth-based. Now, the truth-based philosophy, the oldest field of study, is faced with the crisis of extinction.

After German Philosopher Hegel pronounced the end of philosophy about two centuries ago, many famous philosophers such as Nietzsche and Heidegger concurred with him. Following Heidegger, most philosophers believe that the modern world is a blind-sighted society dominated by science/technology. They believe in what Heidegger claimed: *although philosophy as metaphysics still thinks, science does not think*.

While the end of philosophy was meant to be “the top” or “apex” by Hegel, some philosophers and scientists went further to announce the death of philosophy. For instance, in their influential 2010 book titled *The Grand Design*, two world renowned physicists, namely, Stephen Hawking and Leonard Mlodinow declared: “philosophy is dead”,

“*philosophy has not kept up with developments in modern science, particularly physics*”, “*scientists have become the bearers of the torch in humans’ quest for knowledge*”, “*M-theory predicts that a great many universes were created out of nothing*” and “*Their creation does not require the intervention of some supernatural being or god.*” But can we solve any problem without philosophical thinking? Why has the truth-based and being-centered intensive search for ether, strings and monopoles either failed or got no result so far? Why dipoles are everywhere?

When Hawking and Mlodinow advocated M-theory in their book, they also promoted the concept of negative and positive energies. However, they stopped short of pointing out the unavoidable consequence that the negative and positive energies are respectively the Yin and Yang of nature [Gore & van Oudenaarden, 2009] and, based on YinYang bipolar equilibrium or non-equilibrium, the many dimensions of M-theory can be unified with a YinYang bipolar geometry for supersymmetric bipolar interaction and quantum entanglement. As a result, leaving God alone, we still need to answer the following two immediate follow-up deeper questions:

- (1) Do the many universes in M-Theory need to follow the same equilibrium or non-equilibrium conditions as manifested by the 2nd law of thermodynamics?
- (2) Can all the truth-based and being-centered universes be unified under a single equilibrium-based and harmony-centered universe?
- (3) Why is relativity and quantum theory still not unified?
- (4) What is the driving force of mutation, natural selection and evolution? Could it be equilibrium or non-equilibrium?

Evidently, the equilibrium or non-equilibrium approach presents a new philosophical thinking and a new computational paradigm that cannot be dismissed because it is both scientific and philosophical. Even after the universe disappeared in a black hole, bipolar equilibrium or non-equilibrium of negative-positive energies would still be there due to particle or antiparticle emission [Hawking 1974, 1975]. Actually, the most fundamental property of the universe is not being and truth but equilibrium-based YinYang bipolarity. Thus, the equilibrium or non-equilibrium thinking has the potential of leading to a scientific reincarnation of philosophy. The dynamic nature of the new philosophy can lead to new approaches for problem solving especially for bio-quantum computation.

5 Conclusions

YinYang bipolar dynamic logic and bipolar quantum linear algebra have been introduced in a completely background independent YinYang bipolar Geometry. Basic structures of the new type of combinatorics have been presented and discussed. The potential utilities of the new structures in biological computing have been outlined. Philosophical distinctions of the new approach have been drawn from existing approaches.

It is noted in bioinformatics that, despite one insightful surprise after another the genome has yielded to biologists, the primary goal of the Human Genome Project – to ferret out the genetic roots of common diseases like cancer and Alzheimer's and then generate treatments – has been largely elusive [Wade 2010]. Bipolar quantum cellular combinatorics has been presented as a complementary or alternative approach to bioinformatics.

Reference

- [1] Ai, W., Narahari, J. & Roman A. (2000). Yin Yang 1 Negatively Regulates the Differentiation-Specific E1 Promoter of Human Papillomavirus Type 6. *J. of Virology*, vol. 74, no. 11, 5198-5205.
- [2] Fermi National Accelerator Laboratory, *Press Release 06-19*, Sept. 25, 2006.
- [3] Gore, J & van Oudenaarden, A. (2009). Synthetic biology: The yin and yang of nature. *Nature*, Vol. 457, No. 7227, 2009, 271-272, doi:10.1038/457271a.
- [4] Hawking, S (1974). Black-hole evaporation, *Nature* 248, 30–31 (1974).
- [5] Hawking, S (1975). Particle creation by black holes'. *Communications in Mathematical Physics*, Volume 43, Number 3 (1975), 199-220.
- [6] Hawking, S. and Mlodinow, L. (2010). *The Grand Design*. Random House Digital, Inc., 2010.
- [7] Jacobsen, B. M. & Skalnik, D. G. (1999). YY1 Binds Five cis-Elements and Trans-activates the Myeloid Cell-restricted gp91phox Promoter. *J. Biol. Chem.* 274: 29984-29993 (1999).
- [8] Jaeger, S., Chen, S. -S., & Zhang, W. -R. (2009). TCM in Innate Immunity. *Proc. of International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS)*. Shanghai, China, Aug. 2009. 397-401
- [9] Liu, H., Schmidt-Supprian, M., Shi, Y., Hobeika, E., Barteneva, N., Jumaa, H., Pelanda, R., Reth, M., Skok, J., Rajewsky, K. & Shi, Y (2007). Yin Yang 1 is a critical regulator of B-cell development. *Genes Dev.* 21:1179-1189 (2007)
- [10] Palko, L., Bass, H. W., Beyrouthy, M. J., & Hurt, M. M. (2004). The Yin Yang-1 (YY1) protein undergoes a DNA-replication-associated switch in localization from the cytoplasm to the nucleus at the onset of S phase. *J of Cell Science*, 117, 465-476.
- [11] Shi, Y., Seto, E., Chang, L.-S. & Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenovirus E1A protein. *Cell*, vol. 67, no. 2, 377-388.
- [12] Vasudevan, S., Tong, Y., Steitz, J. A. (2007). Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. *Science*, Vol. 318. no. 5858, 2007, 1931 – 1934
- [13] Wade, N. (2010). A Decade Later, Human Gene Map Yields Few New Cures. *New York Times*, 12 June 2010.
- [14] Wilkinson, F. H. Park, K. & Atchison, M. L. (2006). Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proc. Natl. Acad. Sci. USA*, 103:19296-19301.
- [15] Zhang, W. -R., Chen, S. & Bezdek, J. C. (1989). POOL2: A Generic System for Cognitive Map Development and Decision Analysis. *IEEE Trans. on SMC.*, Vol. 19, No. 1, Jan./Feb. 1989, 31-39.
- [16] Zhang, W. -R., Chen, S., Wang, W., & King, R. (1992). A Cognitive Map Based Approach to the Coordination of Distributed Cooperative Agents. *IEEE Trans. on SMC, Vol. 22, No. 1*, 1992, 103-114.
- [17] Zhang, W. -R. (2003). Equilibrium Relations and Bipolar Cognitive Mapping for Online Analytical Processing with Applications in International Relations and Strategic Decision Support. *IEEE Trans. on SMC, Part B, Vol. 33. No. 2*, 2003, 295-307.
- [18] Zhang, W. -R. and Zhang, L. (2004), "YinYang Bipolar Logic and Bipolar Fuzzy Logic." *Information Sciences*. Vol. 165, No. 3-4, 2004, pp265-287.
- [19] Zhang, W. -R. (2005), "YinYang Bipolar Lattices and L-Sets for Bipolar Knowledge Fusion, Visualization, and Decision." *Int'l J. of Inf. Tech. and Decision Making*, Vol. 4, No. 4: 621-645.
- [20] Zhang, W. -R., A. Pandurangi, and K. Peace (2007), "YinYang Dynamic Neurobiological Modeling and Diagnostic Analysis of Major Depressive and Bipolar Disorders." *IEEE Trans. on Biomedical Engineering*, Oct. 2007 54(10):1729-39.
- [21] Zhang, W. -R and Chen, S. S. (2009), "Equilibrium and Non-Equilibrium Modeling of YinYang Wuxing for Diagnostic Decision Analysis in Traditional Chinese Medicine." *Int'l J. of Infor. Tech. and Decision Making*. Vol. 8, No. 3, 2009. pp529-548
- [22] Zhang, W. -R, H. J. Zhang, Y. Shi & S. S. Chen (2009), "Bipolar Linear Algebra and YinYang-N-Element Cellular Networks for Equilibrium-Based Biosystem Simulation and Regulation." *Journal of Biological Systems*, Volume: 17, Issue: 4 (2009) pp. 547-576.
- [23] Zhang, W. -R., Pandurangi, K. A., Peace, K., E., Zhang, Y. & Zhao, Z. (2011), MentalSquares – A Generic Bipolar Support Vector Machine for Psychiatric Disorder Classification, Diagnostic Analysis and Neurobiological Data Mining. *Int'l J. on Data Mining and Bioinformatics. Vol. 17, No. 4, 2011, 547-576.*
- [24] Zhang, W.-R. (2011a) . *YinYang Bipolar Relativity: A Unifying Theory of Nature, Agents and Causality with Applications in Quantum Computing, Cognitive Informatics and Life Sciences*. IGI Global, 2011.
- [25] Zhang, W.-R. (2011b). YinYang Bipolar Atom and Quantum Cellular Automation. *BIBE Workshops, IEEE BIBE 2011, Atlanta*, pp 791 - 797

Simulated Docking of Laninamivir with the 2009 Pandemic Strain Influenza A/H1N1 Neuraminidase Active Site

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA
email: jhorner@cybermesa.com

Abstract

Influenza neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Laninamivir is a neuraminidase inhibiting drug approved for general use in Japan in 2010 for the treatment of influenza, and for emergency use in the US in 2011. Here I provide a computational docking analysis of laninamivir with the active site of the neuraminidase of the 2009 Influenza A/H1N1 strain. The computed inhibitor/receptor binding energy suggests that laninamivir would be effective against that strain.

Keywords: Influenza, H1N1, neuraminidase, laninamivir

1.0 Introduction

Influenza neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Laninamivir (4S,5R,6R)-5-acetamido-4-carbamimidamido-6-[(1R,2R)-3-hydroxy-2-methoxypropyl]-5,6-dihydro-4H-pyran-2-carboxylic acid; [14]) is a neuraminidase inhibitor approved in Japan in 2010 for general use in the treatment of influenza and for emergency use in the US in 2011.

In the World Health Organization serotype-based influenza taxonomy, influenza type A has nine neuraminidase-related sero-subtypes, and these subtypes correspond at least roughly to differences in the active-site structures of the flu neuraminidases. The subtypes fall into two groups ([3]): group-1 contains the subtypes N1, N4, N5 and N8; group-2 contains the subtypes N2, N3, N6, N7 and N9. Laninamivir was designed to target the group-2 neuraminidases.

The available crystal structures of the group-1 N1, N4 and N8 neuraminidases ([1]) reveal that the active sites of these enzymes have a very different three-dimensional structure from that of group-2 enzymes. The differences lie in a loop of amino acids known as the "150-loop", which in the group-1 neuraminidases has a conformation that opens a cavity not present in the group-2 neuraminidases. The 150-loop contains an amino acid designated Asp 151; the side chain of this amino acid has a carboxylic acid that, in group-1 enzymes, points away from the active site as a result of the 'open' conformation of the 150-loop. The side chain of another active-site amino acid, Glu 119, also has a different conformation in group-1 enzymes compared with the group-2 neuraminidases (8)).

The Asp 151 and Glu 119 amino acid side chains form critical interactions with neuraminidase inhibitors. For neuraminidase subtypes with the "open conformation" 150-loop, the side chains

of these amino acids might not have the precise alignment required to bind inhibitors tightly ([8]). The active site of the 1918 H1N1 strain has the 150-loop configuration.

The difference in the active-site conformations of the two groups of neuraminidases may also be caused by differences in amino acids that lie outside the active site. This means that an enzyme inhibitor for one target will not necessarily have the same activity against another with the same active-site amino acids and the same overall three-dimensional structure.

Influenza

A/California/04/2009(H1N1) is an atypical group 1 NA with some group 2-like features in its active site (lack of a 150-cavity) ([4]).

2.0 Method

The general objective of this study is straightforward: to computationally assess the binding energy of the active site of crystallized A/California/04/2009(H1N1) neuraminidase with laninamivir. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment.

Protein Data Bank (PDB) 3TI3 ([6]) is a structural description of most of the crystallized neuraminidase of Influenza

A/H1N1 3TI3 consists of two identical chains, designated Chain A and Chain B.

3TI3 was downloaded from PDB on 22 February 2011. A PDB description of laninamivir was extracted from PDB 3TI8 ([4]) using *AutoDock Tools* v 4.2 (ADT, [9]). ADT was then used to perform the docking of laninamivir to the receptor. More specifically, in ADT, approximately following the rubric documented in [12]

-- Chain B, and the water in Chain A, of 3TI3 were deleted

-- Chain A's active-site was extracted. (3TI3 identifies the active site of Chain A as 15 amides: ARG118, GLU119, ASP151, ARG152, ARG156, TRP178, ARG224, GLU227, SER246, GLU276, GLU277, ARG292, ASN294, ARG371, and TYR406.)

-- the hydrogens, charges, and torsions in the ligand and active site were adjusted using the ADT-recommended defaults

-- and finally, the ligand, assumed to be flexible wherever that assumption is physically possible, was auto-docked to the active site, assumed to be rigid, using the Lamarckian genetic algorithm implemented in ADT. The best-fit (lowest-energy) configuration from the analysis was saved, and the distances between the receptor and ligand in 3TI3, and those computed here, were compared.

The ADT parameters for the docking are shown in Figure 1. Most values are, or are a consequence of, ADT defaults.

```

autodock_parameter_version 4.2      # used by autodock to validate parameter set
outlev 1                             # diagnostic output level
intelec                              # calculate internal electrostatics
seed pid time                        # seeds for random generator
ligand_types C HD OA N              # atoms types in ligand
fld 3TI3_active.maps.fld            # grid data file
map 3TI3_active.C.map               # atom-specific affinity map
map 3TI3_active.HD.map              # atom-specific affinity map
map 3TI3_active.OA.map              # atom-specific affinity map
map 3TI3_active.N.map               # atom-specific affinity map
elecmap 3TI3_active.e.map           # electrostatics map
desolvmap 3TI3_active.d.map         # desolvation map

```

```

move laninamivirA.pdbqt          # small molecule
about 22.7762 -20.7805 -52.3029  # small molecule center
tran0 random                    # initial coordinates/A or random
axisangle0 random              # initial orientation
dihe0 random                    # initial dihedrals (relative) or random
tstep 2.0                       # translation step/A
qstep 50.0                     # quaternion step/deg
dstep 50.0                     # torsion step/deg
torsdof 9                      # torsional degrees of freedom
rmstol 2.0                     # cluster tolerance/A
extnrg 1000.0                  # external grid energy
e0max 0.0 10000                # max initial energy; max number of retries
ga_pop_size 150                # number of individuals in population
ga_num_evals 2500000           # maximum number of energy evaluations
ga_num_generations 27000       # maximum number of generations
ga_elitism 1                   # number of top individuals to survive to next
generation
ga_mutation_rate 0.02          # rate of gene mutation
ga_crossover_rate 0.8          # rate of crossover
ga_window_size 10              #
ga_cauchy_alpha 0.0            # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0            # Beta parameter Cauchy distribution
set_ga                          # set the above parameters for GA or LGA
sw_max_its 300                 # iterations of Solis & Wets local search
sw_max_succ 4                  # consecutive successes before changing rho
sw_max_fail 4                  # consecutive failures before changing rho
sw_rho 1.0                     # size of local search space to sample
sw_lb_rho 0.01                # lower bound on rho
ls_search_freq 0.06           # probability of performing local search on
individual
set_pswl                        # set the above pseudo-Solis & Wets parameters
unbound_model bound           # state of unbound ligand
ga_run 10                      # do this many hybrid GA-LS runs
analysis                       # perform a ranked cluster analysis

```

Figure 1. ADT parameters for the docking in this study

3.0 Results

The interactive problem setup, which assumes familiarity with the general neuraminidase "landscape", took about 20 minutes in ADT; the docking proper, about 28 minutes on the platform described in Section 2.0. The platform's performance monitor suggested that the calculation was more or less uniformly distributed across the four processors at ~25% of peak per

processor (with occasional bursts to 40% of peak), and required a constant 2.9 GB of memory.

Figure 2 shows the best-fit laninamivir/receptor energy and position summary produced by ADT under the setup shown in Figure 1. The estimated free energy of binding under these conditions is ~ -9.2 kcal/mol; the estimated inhibition constant, ~179 nanoMolar at 298 K.

```

MODEL          1
USER           Run = 1
USER           Cluster Rank = 1
USER           Number of conformations in this cluster = 10
USER
USER           RMSD from reference structure          = 45.375 A
USER
USER           Estimated Free Energy of Binding       = -9.21 kcal/mol  [(1)+(2)+(3)-(4)]
USER           Estimated Inhibition Constant, Ki     = 178.50 nM (nanomolar)  [Temperature =
298.15 K]
USER
USER           (1) Final Intermolecular Energy      = -11.89 kcal/mol
USER           vdW + Hbond + desolv Energy          = -8.64 kcal/mol
USER           Electrostatic Energy                 = -3.25 kcal/mol
USER           (2) Final Total Internal Energy      = -1.55 kcal/mol
USER           (3) Torsional Free Energy            = +2.68 kcal/mol
USER           (4) Unbound System's Energy  [(2)]   = -1.55 kcal/mol
USER
USER
USER           DPF = 3TI3_active.dpf
USER           NEWDPF move      laninamivirA.pdbqt
USER           NEWDPF about     22.776199 -20.780500 -52.302898
USER           NEWDPF tran0    29.995228 14.716856 -20.257575
USER           NEWDPF axisangle0 0.364842 0.651628 0.665035 136.195393
USER           NEWDPF quaternion0 0.338509 0.604594 0.617033 0.373025
USER           NEWDPF dihe0    -37.44 42.58 -154.69 9.42 75.29 -0.69 -63.75 -2.17 -22.15
USER
USER           x      y      z      vdW      Elec      q      RMS
ATOM          1  CAA LNV A 901      29.490 13.269 -22.647 -0.15 +0.15      +0.235 45.375
ATOM          2  CAB LNV A 901      30.930 13.773 -22.576 -0.31 +0.01      +0.103 45.375
ATOM          3  CAC LNV A 901      31.331 14.532 -21.306 -0.31 -0.00      +0.059 45.375
ATOM          4  CAD LNV A 901      30.201 14.801 -20.271 -0.20 +0.03      +0.090 45.375
ATOM          5  CAE LNV A 901      28.733 14.540 -20.712 -0.17 +0.05      +0.107 45.375
ATOM          6  OAF LNV A 901      28.472 13.928 -21.961 -0.16 -0.23      -0.334 45.375
ATOM          7  NAZ LNV A 901      32.666 14.349 -20.823 -0.24 +0.04      -0.194 45.375
ATOM          8  HAZ LNV A 901      33.075 13.416 -20.882 -0.31 -0.15      +0.184 45.375
ATOM          9  CBA LNV A 901      33.464 15.400 -20.262 +0.01 +0.08      +0.669 45.375
ATOM         10  NBC LNV A 901      33.177 16.681 -20.494 -0.24 +0.06      -0.235 45.375
ATOM         11  NBB LNV A 901      34.569 15.045 -19.644 -0.31 -0.14      -0.235 45.375
ATOM         12  1HBC LNV A 901      33.745 17.429 -20.095 -0.33 -0.10      +0.174 45.375
ATOM         13  2HBC LNV A 901      32.320 16.956 -20.973 +0.07 -0.07      +0.174 45.375
ATOM         14  2HBB LNV A 901      34.788 14.065 -19.467 -0.41 +0.17      +0.174 45.375
ATOM         15  1HBB LNV A 901      35.137 15.793 -19.245 -0.45 +0.08      +0.174 45.375
ATOM         16  NBG LNV A 901      30.418 15.692 -19.167 -0.04 -0.20      -0.324 45.375
ATOM         17  HBG LNV A 901      30.088 16.652 -19.263 +0.10 +0.09      +0.169 45.375
ATOM         18  CBD LNV A 901      31.060 15.336 -17.944 -0.25 +0.24      +0.218 45.375
ATOM         19  OBF LNV A 901      31.328 14.168 -17.680 -0.68 -0.43      -0.274 45.375
ATOM         20  CBE LNV A 901      31.334 16.452 -16.985 -0.30 +0.14      +0.117 45.375
ATOM         21  CAG LNV A 901      29.038 12.521 -23.857 -0.23 +0.31      +0.204 45.375
ATOM         22  OAH LNV A 901      29.994 11.981 -24.676 -1.04 -1.48      -0.646 45.375
ATOM         23  OAI LNV A 901      27.913 12.739 -24.274 -1.07 -1.55      -0.646 45.375
ATOM         24  CAJ LNV A 901      27.685 14.292 -19.673 -0.06 +0.16      +0.210 45.375
ATOM         25  OAW LNV A 901      27.479 12.937 -19.582 +0.01 -0.30      -0.381 45.375
ATOM         26  CAX LNV A 901      28.329 12.184 -18.788 -0.06 +0.13      +0.202 45.375
ATOM         27  CAK LNV A 901      26.441 15.111 -19.950 -0.21 +0.12      +0.177 45.375
ATOM         28  OAY LNV A 901      26.236 16.106 -19.010 -0.19 -0.29      -0.390 45.375
ATOM         29  HAY LNV A 901      25.326 16.177 -18.744 -0.27 +0.06      +0.210 45.375
ATOM         30  CAL LNV A 901      25.224 14.253 -20.211 -0.25 +0.15      +0.198 45.375
ATOM         31  OAM LNV A 901      24.033 14.668 -19.586 -0.18 -0.21      -0.398 45.375
ATOM         32  HAM LNV A 901      23.974 15.593 -19.378 -0.40 -0.17      +0.209 45.375
TER
ENDMDL

```

Figure 2. ADT's laninamivir energy and position predictions.

Figure 3 is a rendering of the active-site/inhibitor configuration computed in this study.

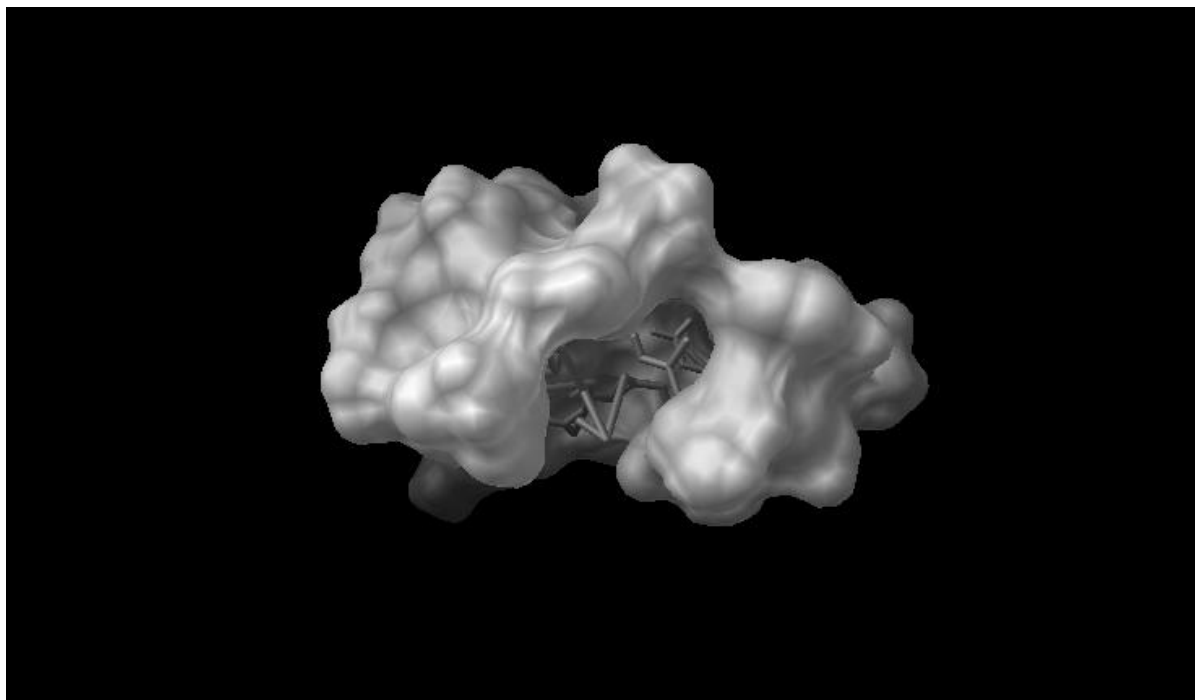


Figure 3. Rendering of laninamivir computationally docked with the active site of PDB 3TI3. The molecular surface of the receptor is shown in white; the inhibitor, in stick form in grey. Only the interior, inhibitor-containing region of the molecular surface of the active site can be compared to *in situ* data: the surface distal to the interior is a computational artifact, generated by the assumption that active site is detached from the rest of the receptor.

The distances between ligand and receptor atoms in 3TI3, and the corresponding distances in the present computation were within 10% of each other.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The inhibition constant computed in this study (~179 nanoMolar at ~298 K) is much smaller than the inhibition constant of neuraminidase inhibitors that are not

clinically effective ([10], [11], [13], [14], [15]) against several H1N1 genotypes. This suggests that laninamivir would be more effective against Influenza A/California/04/2009(H1N1)) than either oseltamivir or zanamivir.

2. The docking study reported here assumes that the receptor is rigid. This assumption is appropriate for the binding energy computation for PDB 3TI3 per se. However, the calculation does not reflect what receptor "flexing" could contribute to the interaction of the ligand with native unliganded receptor.

3. The analysis described in Sections 2.0 and 3.0 assumes receptor is in a crystallized form. *In situ*, at physiologically normal temperatures (~310 K), the receptor is not in crystallized form. The ligand/receptor conformation *in situ*, therefore, may not be identical to their conformation in the crystallized form.

4. Minimum-energy search algorithms other than the Lamarckian genetic algorithm used in this work could be applied to this docking problem. Future work will use Monte Carlo/simulated annealing algorithms.

5. A variety of torsion and charge models could be applied to this problem, and future work will do so.

6. 3TI3 has two chains, each with its own active site. The work described in this paper was performed on Chain A only. Chain B appears to have an active site highly similar to the Chain A active site. Future work will assess the ligand/receptor binding energies of Chains B.

5.0 Acknowledgements

This work benefited from discussions with Tony Pawlicki. For any problems that remain, I am solely responsible.

6.0 References.

[1] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49.

[2] Johnson NP and Mueller J. Updating the accounts: global mortality of the 1918-1920 "Spanish " influenza pandemic. *Bulletin of the History of Medicine* 76 (2002), 105-115.

[3] World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization* 58 (1980), 585-591.

[4] Vavricka CF, Li Q, Wu Y, Qi J, Wang M, Liu Y, Gao F, Liu J, Feng E, He J, Wang J, Liu H, Jiang H, and Gao GF. Structural and functional analysis of laninamivir and its octanoate prodrug reveals group specific mechanisms for Influenza NA inhibition. *PLoS Pathogens* 7 (October 2011): e1002249.

doi:10.1371/journal.ppat.1002249.

[5] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.

[6] PDB ID = 10.2210/pdb3ti3/pdb. See also [4].

[7] US Centers for Disease Control. *Summary: Interim Recommendations for the Use of Influenza Antiviral Medications in the Setting of Laninamivir Resistance among Circulating Influenza A (H1N1) Viruses, 2008-09 Influenza Season.* 19 December 2008. URL <http://www.cdc.gov/flu/professionals/antivirals/summary.htm>.

[8] Luo M. Structural biology: antiviral drugs fit for a purpose. *Nature* 443 (7 September 2006), 37-38. doi:10.1038/443037a, published online 6 September 2006.

[9] Morris GM, Goodsell DS, Huey R, Lindstrom W, Hart WE, Kurowski S, Halliday S, Belew R, and Olson AJ. *AutoDock* v4.2. <http://autodock.scripps.edu/>. 2010.

[10] Drug Bank. *Zanamivir*. <http://www.drugbank.ca/drugs/APRD00378>.

[11] Govorkova EA et al. Comparison of efficacies of RWJ-270201, zanamivir, and oseltamivir against H5N1, H9N2, and other avian influenza viruses. *Antimicrobial Agents and Chemotherapy* 45 (2001), 2723-2732.

- [12] Huey R and Morris GM. *Using AutoDock 4 with AutoDock Tools: A Tutorial*. 8 January 2008. <http://autodock.scripps.edu/>.
- [13] Cheng Y and Prusoff WH. Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. *Biochemical Pharmacology* 22 (December 1973), 3099–3108. doi:10.1016/0006-2952(73)90196-2.
- [14] Horner JK. Simulated docking of oseltamivir with the 1918 pandemic strain Influenza A/H1N1 zanamivir-conformed neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. 2011. pp. 130-135.
- [15] Horner JK. Simulated docking of zanamivir with the 1918 pandemic strain Influenza A/H1N1 neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. pp. 136-142.

Towards a Motor Ability Training Table for Rehabilitating Children with Obstetric Brachial Plexus Lesion

Alberti, Eduardo J.¹, Pichorim, Sérgio F.¹, and Brawerman, Alessandro²

¹School of Electrical Engineering and Industrial Computer Science, Federal University of Technology of Paraná

²Computer Engineering Department, University of Positivo
Curitiba, Paraná, Brasil

Abstract—*Obstetric Brachial Plexus Lesion (OBPL) is an injury of the cervical spine and chest characterized by blockage of one or more brachial plexus nerves, usually during the child delivery procedure. Research indicates that the number of infant OBPL cases has been growing in a much faster rate than the population growth. Despite that, most of the equipment and electronic devices employed to help and accelerate the OBPL treatment are designed for adult use, treating kids as a miniaturized adult. This work proposes a simple yet efficient motor ability training table, specifically designed for infant use. The training table uses games, with light, sound and several complexity levels to arouse the child interesting and to make the treatment more challenging. On the top of that, a computer system that presents patient progress through graphical reports helps the professional to further analyzed the treatment result.*

Keywords: OBPL, rehabilitation engineering, motor ability training table

1. Introduction

The OBPL - Obstetric Brachial Plexus Lesion - is caused by excessive traction of the neck, head and arm during the delivery procedure, exceeding the tolerance thresholds of the nerves [1], [2].

The rate of OBPL cases has been growing along with population growth, but in a much more considerable proportion, about 76 %, in Brazil. According to [3], children rehabilitation technology did not follow this growth, the technologies developed are still based on adults characteristics.

This project proposes the development of a motor ability training table to aid in the treatment of OBPL having as its main focus children rehabilitation. In the following sections, topics concerning the formation of OBPL lesion, treatment techniques, types of injury, the project development and its specifications are discussed.

2. Obstetrical Brachial Lesion

This section presents fundamental concepts such as the formation of Obstetric Brachial Plexus Lesion, its causes, residual deformities, and current treatment possibilities.

2.1 Lesion Formation

The Obstetric Brachial Lesion or Obstetric Brachial Plexus Lesion is an injury of the cervical spine and chest characterized by blockage of one or more nerves of the brachial plexus [4]. The lesion is usually the result of direct trauma caused during delivery. [2].

Research conducted in 2000 and 2010 shows that the rate of tocotraumatism cases (one should take into account the aggregation of all cases of various types of traumas) occurring during delivery, increased approximately 75.6 %, and mortality involving this type of injury accounts only to 0.6 cases per 100,000 births [5], [6]. It should be taken into account, when interpreting such data that, according to the Demographic Census of 2000 and 2010, the Brazilian population grew by 12.3 % during this period [7].

Research indicates that the number of cases is incremented as the infant approaches the range between 4 and 5 kg, which can be seen from the graph of Figure 1.

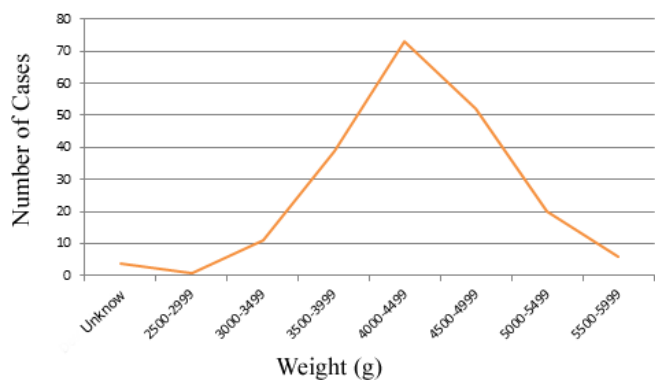


Fig. 1: Relationship between number of cases of OBPL and weight (in grams) of the fetus at birth. Based on [8].

A study conducted by [9] relates, in 311 cases, the number of individuals affected by OBPL versus the type of delivery. This is depicted by Table 1. Note that the sum of births by forceps and suction exceeds the number of assisted births, this is due to the fact that births with the aid of forceps were performed after failure of Ventouse use and Caesarean section.

Table 1: Types of birth *versus* number of cases. Adapted from [9].

Type of Birth	Studed Group		England 1994-1995
	N.	%	%
Spontaneous vertex delivery	183	59	73
Assisted delivery	113	36 *	10,6
Ventouse	87	28	5
Forceps	45	14,5	7
Breech	10	3	1
Caesarean section	5	1,5	15,5

2.2 Types of Lesion

The OBPL is classified taking into account the gravity and the components involved in the lesion. Thus, the lesion can be split into three types: Erb-Duchenne Palsy, Total Brachial Plexus Lesion and Klumpke's Palsy [4].

- **Erb-Duchene Palsy:** in this modality, the injury occurs between the C5 and C7 vertebrae. The arm is in a position called "waiter's tip", with extension and pronation at the elbow and wrist and fingers flexed, as shown in Figure 2. In this case exists decrement of sensitivity, but the movement of grip is intact [6]. Note, when comparing photo A and B, how the arm position and the fingers flexion are characteristics of this type of injury.

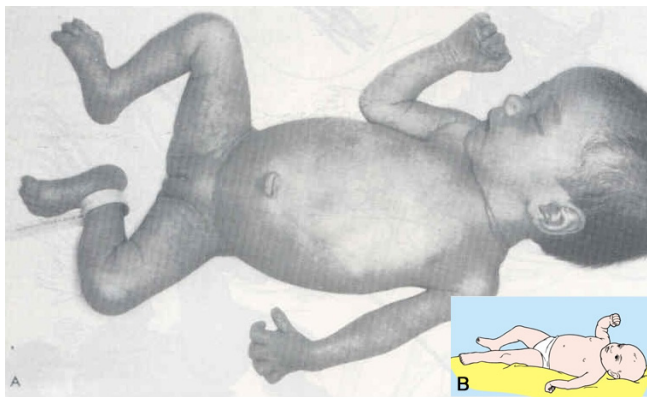


Fig. 2: Typical appearance of the newborn with Erb-Duchenne brachial plexus lesion. Adapted from [10], [11].

- **Total Brachial Plexus Lesion:** in this case, all vertebrae from the C5 to T1 have their roots affected. The sensibility and all reflections are absent, the children does not move or lift the arm [6], [11], this can be observed by the Figure 3, note how the left arm of the patient is shown still.
- **Klumpke's Palsy:** in this modality the vertebrae C7, C8 and T1 are involved, paralyzing the hand muscles, arm flexors and wrist and fingers flexors [13]. The Klumpke's Palsy is the most rare of all types of brachial plexus lesions and response to less than 1 % of the cases [5]. This type of paralysis may affect the



Fig. 3: Total Brachial Plexus Lesion, Clinical Case. Based in [12].

cervical sympathetic fibers, taking ipsilaterally Horner's syndrome, present on the same side of the affected limb [2], [14].

2.3 Residual Deformities

The limitations showed by the subjects affected by the OBPL may vary according with the type of the lesion. According to [15], the patient may present inability to understand and execute tasks requiring bilateral motor skills such as catching a ball or a large object. The residual deformities, according to [16], may be classified in 4 distinct types, taking into account the physiological/anatomical point of this deformity, being:

- **Shoulder:** the subjects affected by OBPL usually present difficulties in the movements of adduction and abduction of the arm, the main motor function of the shoulder [11]. Individuals may also present limitations of active abduction and lateral rotation, which can be seen from Figure 4. Note how the right upper limb movements are limited when compared to the left upper limb [11].
- **Elbow:** the residual deformity at elbow is often developed as a flexion of 45-90 degrees, which may be aesthetically disturbing. There is also a muscle imbalance summed with forced flexion of the elbow, resulting in abnormal bone growth caused by the use of very rigid immobilizing or splints during the recovery process [11].
- **Forearm and Hand:** residual deformities in the forearm and hand are determined by lot distribution, extension and type of the OBPL. While hand paralysis due to Klumpke lesion may be continuous, in the forearm, is common the appearance of deformities of pronation, as shown in Figure 5. Note the contraction of the forearm and elbow flexion [11].

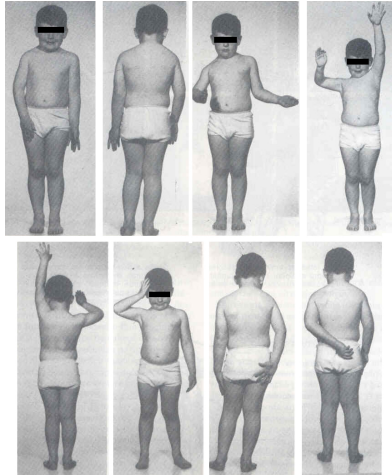


Fig. 4: Residual deformities in the right shoulder of a 6 years old child with obstetric brachial plexus palsy.

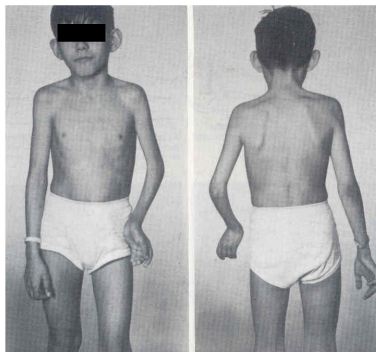


Fig. 5: Residual deformities of the forearm and hand in a child affected by OBPL.

2.4 Treatment

The Orthopedic Management or physical therapy, is the most appropriate treatment for the ones affected by OBPL and has as objective the early treatment in the newborn and child, in order to avoid deformities during the period of spontaneous recovery [15].

According to [11], the passive movements made during the exercise promotes the extension and flexion of the complete arc of all articulations.

Below are shown some of the exercises used during OBPL treatment. According to [15], each of the exercises should be performed repeatedly, several times per day.

2.4.1 Sensorial Development Exercises

The sensorial stimulus, in order to increase awareness that breastfeeding has on its own member, can be performed using a soft towel, gently massaging the arm, or using his/her own arm, massaging his/her own body [17].

Be aware of the affected limb is essential for the good progress of treatment as a patient who has no sensibility to

the member can neglect it and continue to perform tasks only with the "normal" member.

2.4.2 Motor Ability Training Exercises

It is necessary to train grip and manipulate objects with both hands and also just the affected limb. To this end, the therapist can use objects of any kind. To encourage the active use of the atrophied member exercises that are used in everyday situations, such as tying shoes, draw and pick up objects may be strategies for the refinement of activities and for developing more accurate coordination for specific activities [17].

Exercises to gain motion amplitude or motion range are also important since by gaining motion amplitude one reduces the risk of contractures, mental and physical stress and improves blood circulation [18].

3. The Motor Ability Training Table: Specification

The training table proposed is a therapeutic device for the purpose of bringing the physiotherapist and the bearer of OBPL an alternative tool for qualitative and quantitative analysis of brachial plexus injury treatment.

According to [18], the therapist has tools for application of the exercises, but none of these are devoted exclusively or focuses on the OBPL treatment. The tools used by professionals are usually improvised materials such as toys, weights, pulleys and balls.

The development of rehabilitation technologies is occurring at a fast pace and the devices are becoming more individualized, when taken into consideration the type of disability or inability to move. In [3], published in 1996, the authors were concerned with enabling technologies exclusively for children. Current technologies are aimed at adults, eventually considering children as miniaturized adults. This is not enough, since the motor and cognitive functions of a child are in constant change during his/her growth.

The training table constructed consists of a module made of wood and glass, and divided into eight segments, as shown by the diagram of Figure 6. Each of the eight segments has an infrared touch sensor and LEDs in two colors, green and red, and play a note of a eight-note musical scale, from C to C.

These features together form a tool for dynamic exercises. The patient exercises by triggering the segments through touch, as soon as the training table asks for it. The physiotherapist has the ability to choose what will be the exercise performed. An example is to perform a sequence of ordered movements, i.e., the patient should operate with the OBPL affected member by all segments, one at a time, progressively. The table will temporarily turn on the leds of Red color in one segment to tell which should be activated, emits a musical note (for the segment) and wait for the

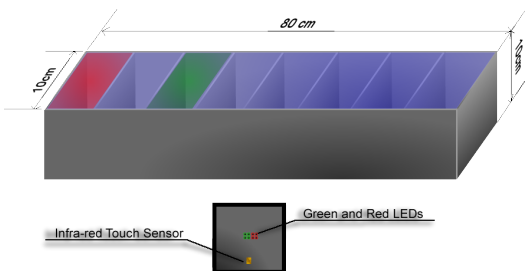


Fig. 6: Diagram of the training table.

patient. If the patient triggers the correct thread it turns green again, otherwise the thread will flash symbolizing the error. Figure 7 shows step-by-step the training table basic operations.

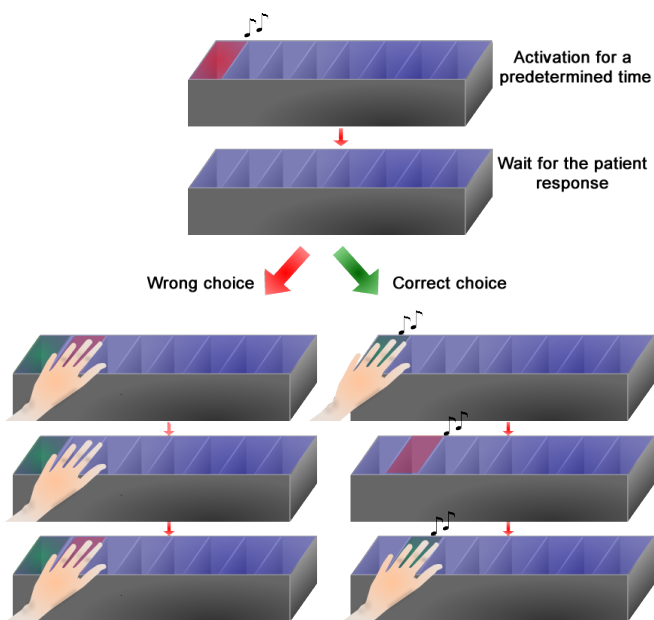


Fig. 7: Basic operation of the training table.

During the exercise, the training table system sends data, related to the exercise, to the computer to which it is connected. The table is able to count time and errors during the exercise execution. Whenever the table system asks the patient to trigger a segment, the time between the actuation by the table itself and by the patient is accounted, this is called arrival time. Whenever a segment is triggered by the patient, the table accounts the duration of this actuation, this is called actuation time. Such measures are important to further analyze the progress of the patient response time according to the complexity of the exercise.

The system allows the physiotherapist to select the complexity of the exercise and stores the progress of each patient in each exercise performed. The exercise complexity is related to how fast each of the segments is trigger by the system, the shorter the time of drive, more attention and

flexibility are required by the patient. There are 10 levels of complexity, which can be used in all exercises, ranging from one second apart (easy level) to 1/10th of a second (hardest level).

To monitor the exercise, a computer system was developed. The system controls/monitors the exercises, receives data regarding the patient movement and presents a graphical analysis of such information. The system is capable of registering patients and storing personal data progress of each registered patient, as shown in Figure 8. The physical therapist can also monitor and control the execution of the exercises through the computer system, as shown in Figure 9.

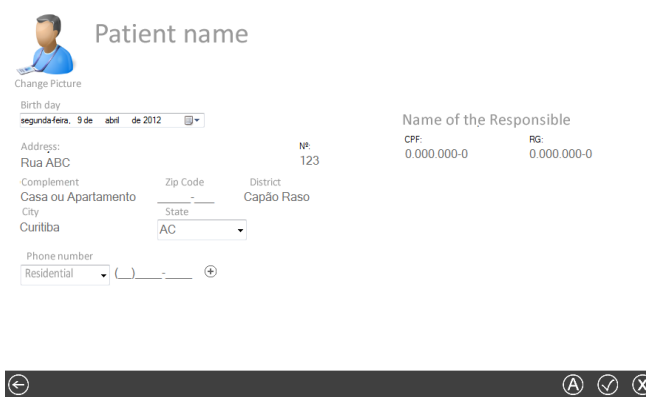


Fig. 8: Registration Screen of the Controlling System.



Fig. 9: Exercises Execution Screen.

The system allows a temporal analysis of the exercise, through the construction of graphs with the data collected during the execution of the exercise. Figure 10 shows the history exercise screen. It is important highlight that the system, by itself, does not have any intelligence to perform data analysis at this moment. It only generates graphics to be analyzed by the responsible professional.

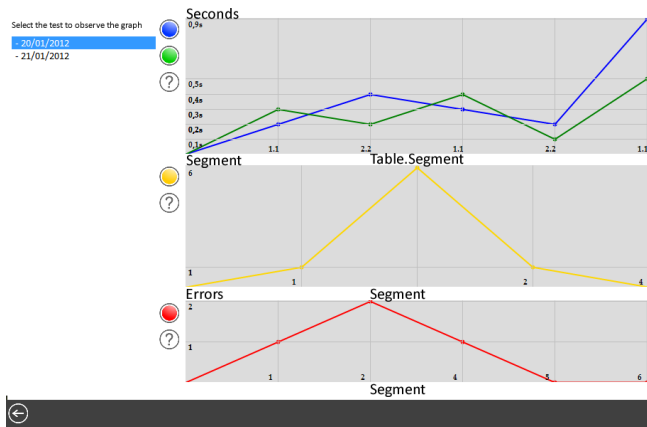


Fig. 10: Exercise history.

3.1 Related Work

The search for related work did not pointed out many items. As mentioned before, there is a bad habit of considering children as miniaturized adults, thus there is not many training devices for kids. This section presents the more relevant related work found.

The authors in [19] describe the development of a generic programmable platform to aid in patient care with physical disabilities, based on a set of non-invasive sensors that can track movements, touches and eye poking. The sensor signals are conditioned and processed in a computer system.

[20] describes the development of a multimedia workstation for children rehabilitation. Based on a cognitive/sensory system, one of the first ever developed, that works on neuromuscular functions. The proposed system uses EMG signals captured to study muscular information.

Finally, although there is not an equipment or device developed, it is important to mention that the authors in [21], [22] discuss about exercise techniques that can be applied during a motor dysfunction treatment.

3.2 Discussion and Results

Is it possible to train a carrier of obstetric paralysis through physical therapy methods? The literature states that it is. According to [15], [11], [23], the orthopedic management is highly recommended as an instrument of neuromuscular recovery and surgical treatment is indicated only in cases of delayed recovery or when there is no response to physiotherapy treatment.

Although all the concepts discussed in the section 2.4, 2 and research using the methodology discussed in section 3.1, there were no positive results for the development of technologies for OBPL treatment, which allows the conclusion that professionals do not have specific devices for performing a focused OBPL physical therapy and that the development of these technologies could result in abbreviation of patient's recovering time and also the reduction of

residual deformities, these are the objectives that we hope to achieve by using the training table proposed in this work.

4. Conclusions

The training table, proposed in this project, is an alternative to conventional treatments that brings to a physical therapist and patient a ludic technique capable of arousing the interest of the child using a game as a treatment model. The presence of light and sound and the presence of several complexity levels make the exercise more interesting and challenging. By the other hand, the computerized system allows a better analysis of the patient's progress throughout the exercise sessions. Thus, the professional can accurately assess the development of motor skills.

References

- [1] G. H. Borschel and H. M. Clarke, "Obstetrical brachial plexus palsy." *Plastic and reconstructive surgery*, vol. 124, no. 1 Suppl, pp. 144e–155e, July 2009.
- [2] R. M. Kliegman, R. E. Behrman, H. B. Jenson, and B. F. Stanton, *Nelson, tratado de pediatria*, 18th ed. Rio de Janeiro: Elsevier Inc., 2009.
- [3] S. Sudarsan, R. Seliktar, P. Benvenuto, and R. Rao, "A method of evaluation of upper limb reaching and keying function in children with motor disability," in *Proceedings of the 1996 Fifteenth Southern Biomedical Engineering Conference*. IEEE, pp. 39–42. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=493108>
- [4] A. E. Bialocerkowski and M. Galea, "Comparison of visual and objective quantification of elbow and shoulder movement in children with obstetric brachial plexus palsy." *Journal of brachial plexus and peripheral nerve injury*, vol. 1, p. 5, Jan. 2006.
- [5] J. P. Cloherty and A. R. Stark, *Manual de Neonatologia*, 4th ed. Rio de Janeiro: MEDSI Editora Médica e Científica Ltda., 2000.
- [6] J. P. Cloherty, A. R. Stark, and E. C. Eichenwald, *Manual de Neonatologia*, 6th ed. Rio de Janeiro: Guanabara Koogan, 2010.
- [7] I. B. d. G. e. E. IBGE, *Sinopse do Censo demográfico 2010*. Rio de Janeiro: Governo Federal Brasileiro, 2010. [Online]. Available: <http://www.ibge.gov.br/home/estatistica/populacao/ceenso2010/sinopse.pdf>
- [8] W. Pondaag, R. Allen, and M. Malessy, "Correlating birthweight with neurological severity of obstetric brachial plexus lesions." *BJOG : an international journal of obstetrics and gynaecology*, pp. 1098–1103, Apr. 2011. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/21481148>
- [9] G. Evans-Jones, S. P. J. Kay, a. M. Weindling, G. Cranny, A. Ward, A. Bradshaw, and C. HERNON, "Congenital brachial palsy: incidence, causes, and outcome in the United Kingdom and Republic of Ireland." *Archives of disease in childhood. Fetal and neonatal edition*, vol. 88, no. 3, pp. F185–9, May 2003.
- [10] H. d. R. a. Rede Sarah, "Paralisia Braquial Obstétrica," 2011. [Online]. Available: http://www.sarah.br/paginas/doencas/po/p_10_paralisia_braquial_obst.htm
- [11] M. O. Tachdjian, *Ortopedia Pediátrica*, 1st ed. São Paulo: Manole Ltda., 1995.
- [12] B. P. P. C. St. Louis Children's Hospital, "Paralisia do plexo braquial, Apresentação clínica dos sintomas," 2011. [Online]. Available: <http://brachialplexus.wustl.edu/portuguese/Presentation0.htm>
- [13] P. C. Chaves, R. R. Albuquerque, and A. L. Moreira, "Reflexos Osteotendinosos, Texto de Apoio," Porto. [Online]. Available: http://www.unirio.br/farmacologia/aulas_fisiologia/2_sistema_nervoso/REFLEXO_E_MOTILIDADE/Reflexos_osteotendinosos_UNIV_DO_PORTO.pdf

- [14] C. A. Shiratori, R. C. Preti, S. A. Schellini, P. Ferraz, and M. Lima, "Síndrome de Horner na infância - Relato de caso," *Arquivos Brasileiros de Oftalmologia*, vol. 67, no. 2, pp. 329–331, 2004. [Online]. Available: http://www.scielo.br/scielo.php?pid=S0004-27492004000200025&script=sci_arttext
- [15] S. K. Campbell, D. W. V. Linden, and R. J. Palisano, *Physical Therapy for Children*, 2nd ed. Philadelphia: W. B. Saunders Company, 2000.
- [16] P. Alves, D. E. S. Oliveira, J. V. Pires, J. Maria, and D. E. M. Borges, "Traumatismos da coluna cervical torácica e lombar: Avaliação epidemiológica," *Revista Brasileira De Ortopedia*, vol. 1995, no. tabela 1, pp. 1–8, 2011.
- [17] K. T. Ratliffe, *Fisioterapia na Clínica Pediátrica*, 1st ed. São Paulo: Livraria Santos Editora Comp. Imp. Ltda, 2001.
- [18] W. D. Bandy and B. Sanders, *Exercício Terapêutico: Técnicas para Intervenção*, 1st ed. Rio de Janeiro: Guanabara Koogan, 2001.
- [19] F. Senatore, D. M. Rubin, and G. J. Gibbon, "Development of a Generic Assistive Platform to Aid Patients with Motor Disabilities," in *14TH Nordic-Baltic Conference on Biomedical Engineering and Medical Physics*. IFMBE Proceedings, 2008, pp. 168–171.
- [20] R. H. Eckhouse and R. A. Maulucci, "A multimedia system for augmented sensory assessment and treatment of motor disabilities," *Telematics and Informatics*, vol. 14, no. 1, pp. 67–82, Feb. 1997. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0736585396000196>
- [21] S. Ostensjo, "Assistive Devices for Children with Disabilities," *International Handbook of Occupational Therapy Interventions*, vol. 2, no. 141-146, 2009.
- [22] M. J. Guralnick, "The system of Early Intervention for Children with Developmental Disabilities: Current Status and Challenges for the Future," *Issues in Clinical Child Psychology*, vol. IV, pp. 465–480, 2007.
- [23] J. Bahm, C. Ocampo-Pavez, and H. Noaman, "Microsurgical technique in obstetric brachial plexus repair: a personal experience in 200 cases over 10 years." *Journal of brachial plexus and peripheral nerve injury*, vol. 2, p. 1, Jan. 2007.

A Simulated Docking of Abiraterone with Cytochrome P450 17A1

Jack K. Horner
PO Box 266
Los Alamos NM 87544
jhorner@cybermesa.com

BIOCOMP 2012

Abstract

Cytochrome P450 17A1 (also known as CYP17A1) catalyses the biosynthesis of androgens in humans. Because prostate cancer cells proliferate in response to androgen steroids, CYP17A1 inhibition can help to prevent androgen synthesis and treat lethal metastatic prostate cancer. Here I report the results of a computational docking of abiraterone, a steroidal inhibitor of CYP17A1 recently approved by the FDA, with the CYP17A1 active site, based on recent X-ray crystallography of the receptor/ligand complex.

Keywords: cytochrome P450, CYP17A1, abiraterone, computational docking, prostate cancer

1.0 Introduction

Cytochrome P450 17A1 (also known as CYP17A1 and cytochrome P450c17) is a membrane-bound monooxygenase that plays a fundamental role in the synthesis of several human steroid hormones ([5]). The 17 α -hydroxylase activity of CYP17A1 is required for the generation of glucocorticoids such as cortisol; the hydroxylase and 17,20-lyase activities of CYP17A1 are required for the production of androgenic and oestrogenic sex steroids. CYP17A1 is thus an important target for the treatment of breast and prostate cancers that proliferate in response to oestrogens and androgens ([6],[7]).

Until recently, steroidal CYP17A1 inhibitors were thought to bind the cytochrome P450 haem iron, more or less parallel to the plane of the haem group in the active site ([8]).

Abiraterone is the active form of a steroidal prodrug recently approved by the US Food and Drug Administration for metastatic prostate cancer ([9],[10]); it is also under investigation

for breast cancer ([11]). Recent X-ray crystallography of abiraterone complexed with the active site of CYP17A1 shows the drug binds the haem iron in the receptor active site, forming a 60° angle above the haem plane and packing against the central I helix with the 3 β -OH interacting with asparagine 202 in the F helix ([1],[3]). This conformation differs substantially from those that are predicted by homology models and from steroids in other cytochrome P450 enzymes with known structures; some features of this conformation are more similar to steroid receptors ([1]).

2.0 Method

The general objective of this study is straightforward: to computationally assess the binding energy of the active site of crystallized cytochrome p450 17A1 with abiraterone. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment.

Protein Data Bank (PDB) 3RUK is a structural description of a crystallized cytochrome p450 17A1 bound to abiraterone. 3RUK has 4 chains, designated A-D.

3RUK was downloaded from PDB ([6]) on 30 January 2012. The ligand and receptor-active-site portions of 3RUK Chain A were extracted to separate files, one each for the ligand and the

receptor, using *AutoDock Tools* (ADT, [2]). ADT was then used to perform the docking of the ligand to the receptor. More specifically, in ADT, approximately following the rubric documented in [4]

-- all waters, and Chains B-D of 3RUK were deleted

-- the ligand (abiraterone) and Chain A's active-site were extracted (3RUK identifies the active site of Chain A as 7 residues: ALA113, ASN202, ILE205, ASP298, ALA302, THR306, and HEM600.)

-- the hydrogens, charges, and torsions in the ligand and active site were adjusted using ADT default recommendations

and finally, the ligand, assumed to be flexible wherever that assumption is physically possible, was auto-docked to the active site, assumed to be rigid, using the Lamarckian genetic algorithm implemented in ADT.

```

autodock_parameter_version 4.2      # used by autodock to validate parameter
                                     set
outlev 1                             # diagnostic output level
intelec                              # calculate internal electrostatics
seed pid time                        # seeds for random generator
ligand_types A C OA HD N            # atoms types in ligand
fld 3RUK_A_active_receptor.maps.fld # grid_data_file
map 3RUK_A_active_receptor.A.map    # atom-specific affinity map
map 3RUK_A_active_receptor.C.map    # atom-specific affinity map
map 3RUK_A_active_receptor.OA.map   # atom-specific affinity map
map 3RUK_A_active_receptor.HD.map   # atom-specific affinity map
map 3RUK_A_active_receptor.N.map    # atom-specific affinity map
elecmap 3RUK_A_active_receptor.e.map # electrostatics map
desolvmap 3RUK_A_active_receptor.d.map # desolvation map
move 3RUK_A_ligand.pdbqt            # small molecule
about 27.936 -1.9813 32.3924        # small molecule center
tran0 random                        # initial coordinates/A or random
axisangle0 random                  # initial orientation
dihe0 random                       # initial dihedrals (relative) or random
tstep 2.0                           # translation step/A
qstep 50.0                          # quaternion step/deg
dstep 50.0                          # torsion step/deg
torsdof 2                           # torsional degrees of freedom
rmstol 2.0                          # cluster_tolerance/A
extnrg 1000.0                       # external grid energy

```

```

e0max 0.0 10000                                # max initial energy; max number of
                                                # retries
ga_pop_size 150                                  # number of individuals in population
ga_num_evals 2500000                             # maximum number of energy evaluations
ga_num_generations 27000                         # maximum number of generations
ga_elitism 1                                      # number of top individuals to survive
                                                # to next generation
ga_mutation_rate 0.02                           # rate of gene mutation
ga_crossover_rate 0.8                           # rate of crossover
ga_window_size 10                                #
ga_cauchy_alpha 0.0                              # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0                              # Beta parameter Cauchy distribution
set_ga                                            # set the above parameters for GA or LGA
sw_max_its 300                                    # iterations of Solis & Wets local
                                                # search
sw_max_succ 4                                    # consecutive successes before changing
                                                # rho
sw_max_fail 4                                    # consecutive failures before changing
                                                # rho
sw_rho 1.0                                       # size of local search space to sample
sw_lb_rho 0.01                                   # lower bound on rho
ls_search_freq 0.06                              # probability of performing local search
                                                # on individual
set_psw1                                         # set the above pseudo-Solis & Wets
                                                # parameters
unbound_model bound                             # state of unbound ligand
ga_run 10                                        # do this many hybrid GA-LS runs
analysis                                         # perform a ranked cluster analysis

```

Figure 1. ADT parameters used in this study. The setup uses a Lamarckian genetic algorithm minimum-energy search; all other ADT parameters are defaulted.

The minimum-energy configuration among those configurations sampled was saved. Interatomic distances between ligand and receptor in the computed form were compared to those in [3].

3.0 Results

The interactive problem setup, which assumes familiarity with the general CYP17A1 "landscape", took about 20 minutes in ADT; the docking proper, about 24 minutes on the platform described in Section 2.0. The platform's performance monitor suggested that the calculation was more or less uniformly

distributed across the four processors at ~25% of peak per processor (with occasional bursts to 40% of peak), and required a constant 2.9 GB of memory.

Figure 2 shows the ligand/receptor energy and position summary produced by ADT for the best-fit conformation obtained under the conditions described in Figure 2.0. The estimated free energy of binding is ~ -6.7 kcal/mol; the estimated inhibition constant, ~13.4 microMolar at 298 K. All distances between receptor and ligand atoms in the computed ligand position lie within 10% of the distances of the corresponding atoms in 3RUK.

 LOWEST ENERGY DOCKED CONFORMATION from EACH CLUSTER

Keeping original residue number (specified in the input PDBQ file) for outputting.

```

MODEL      10
USER      Run = 10
USER      Cluster Rank = 1
USER      Number of conformations in this cluster = 4
USER
USER      RMSD from reference structure      = 7.035 A
USER
USER      Estimated Free Energy of Binding   = -6.65 kcal/mol  [(1)+(2)+(3)-(4)]
USER      Estimated Inhibition Constant, Ki  = 13.35 uM (micromolar)  [Temperature = 298.15 K]
USER
USER      (1) Final Intermolecular Energy    = -7.25 kcal/mol
USER      vdW + Hbond + desolv Energy        = -7.21 kcal/mol
USER      Electrostatic Energy               = -0.03 kcal/mol
USER      (2) Final Total Internal Energy    = -0.21 kcal/mol
USER      (3) Torsional Free Energy          = +0.60 kcal/mol
USER      (4) Unbound System's Energy  [(2)] = -0.21 kcal/mol
USER
USER
USER      DPF = 3RUK_A.dpf
USER      NEWDPF move      3RUK_A_ligand.pdbqt
USER      NEWDPF about     27.936001 -1.981300 32.392399
USER      NEWDPF tran0     24.047148 -7.767324 34.456233
USER      NEWDPF axisangle0 -0.914130 0.398130 -0.076547 106.589001
USER      NEWDPF quaternion0 -0.732874 0.319188 -0.061369 0.597702
USER      NEWDPF dihe0     -168.56 4.04
USER
USER
USER      x      y      z      vdW      Elec      q      RMS
ATOM      1  C1  AER A 601      21.948  -8.070  37.535  -0.23  +0.00  +0.016  7.035
ATOM      2  C2  AER A 601      21.541  -7.891  38.991  -0.16  +0.00  +0.033  7.035
ATOM      3  C3  AER A 601      22.742  -7.712  39.883  -0.16  +0.03  +0.122  7.035
ATOM      4  C4  AER A 601      23.549  -6.485  39.449  -0.28  +0.02  +0.066  7.035
ATOM      5  C5  AER A 601      23.830  -6.469  37.972  -0.34  -0.02  -0.072  7.035
ATOM      6  C6  AER A 601      25.043  -6.039  37.636  -0.40  -0.01  -0.023  7.035
ATOM      7  C7  AER A 601      25.658  -6.228  36.271  -0.50  +0.01  +0.033  7.035
ATOM      8  C8  AER A 601      24.627  -6.516  35.184  -0.38  -0.00  -0.001  7.035
ATOM      9  C9  AER A 601      23.541  -7.465  35.716  -0.35  +0.00  +0.003  7.035
ATOM     10  C10 AER A 601      22.792  -6.934  36.948  -0.25  -0.00  -0.017  7.035
ATOM     11  C11 AER A 601      22.612  -8.006  34.626  -0.27  -0.00  +0.007  7.035
ATOM     12  C12 AER A 601      23.364  -8.531  33.397  -0.20  -0.00  +0.014  7.035
ATOM     13  C13 AER A 601      24.287  -7.478  32.849  -0.31  +0.00  -0.016  7.035
ATOM     14  C14 AER A 601      25.256  -7.237  33.987  -0.41  +0.00  +0.003  7.035
ATOM     15  C15 AER A 601      26.525  -6.656  33.368  -0.48  +0.00  +0.010  7.035
ATOM     16  C16 AER A 601      26.619  -7.229  31.965  -0.35  +0.00  +0.036  7.035
ATOM     17  C17 AER A 601      25.239  -7.844  31.760  -0.36  +0.01  -0.060  7.035
ATOM     18  C18 AER A 601      23.513  -6.267  32.298  -0.11  +0.00  +0.020  7.035
ATOM     19  C19 AER A 601      21.807  -5.832  36.583  -0.08  +0.00  +0.020  7.035
ATOM     20  C20 AER A 601      24.880  -8.690  30.556  -0.35  +0.01  -0.018  7.035
ATOM     21  C25 AER A 601      23.658  -8.502  29.887  -0.18  -0.01  +0.014  7.035
ATOM     22  C24 AER A 601      23.349  -9.288  28.789  -0.14  -0.01  +0.018  7.035
ATOM     23  C23 AER A 601      24.280 -10.236  28.386  -0.12  -0.08  +0.087  7.035
ATOM     24  N22 AER A 601      25.451 -10.402  29.044  -0.15  +0.46  -0.375  7.035
ATOM     25  H22 AER A 601      26.106 -11.116  28.722  -0.34  -0.34  +0.164  7.035
ATOM     26  C21 AER A 601      25.771  -9.655  30.107  -0.30  -0.07  +0.099  7.035
ATOM     27  O3  AER A 601      22.257  -7.463  41.198  -0.06  -0.08  -0.395  7.035
ATOM     28  H3  AER A 601      21.326  -7.653  41.195  +0.03  +0.03  +0.210  7.035
TER
ENDMDL
  
```

Figure 2. Coordinates of abiraterone generated by this study.

Figure 3 is a rendering produced in ADT of the CYP17A1/abiraterone docking described in Section 2.0.

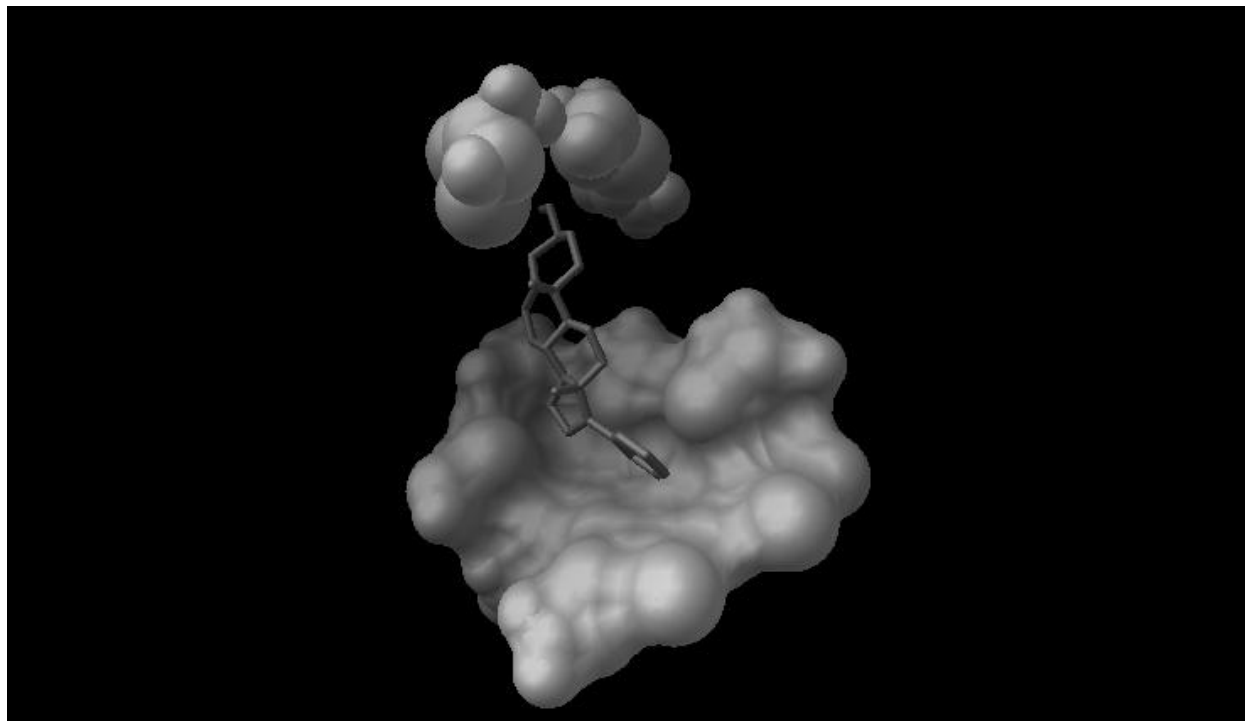


Figure 3. AutoDock Tools (ADT,[2]) rendering of a computational docking of abiraterone (the ligand, shown in stick-and-ball form in darker grey) with molecular surface of the active site of Chain A of cytochrome p450 17A1 (shown in lighter grey), derived from PDB 3RUK ([1],[3]). The lower right end of the ligand lies directly above the center of the haem group in the active site.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The inhibition constant computed in this study (~13.4 microMolar at ~298 K) is comparable to the inhibition constant of cancer-therapeutic ligand/receptor interactions that are clinically effective.

2. All distances between receptor and ligand atoms in the computed ligand position lie within 10% of the distances of the corresponding atoms in 3RUK. (For electrostatic forces, a 10% distance difference would correspond to a ~20% difference in electrostatic force and potential energy, in the worst case. One could of course apply other statistics to the coordinate sets and provide a more comprehensive comparison of other forces/energies. Future work will address those issues.)

3. The docking study reported here assumes that the receptor is rigid. This assumption is appropriate for the binding energy computation for PDB 3RUK per se. However, the calculation does not reflect what receptor "flexing" could contribute to the interaction of the ligand with native unliganded receptor.

4. The analysis described in Sections 2.0 and 3.0 assumes receptor is in a crystallized form. *In situ*, at physiologically normal temperatures (~310 K), the receptor is not in crystallized form. The ligand/receptor conformation *in situ*, therefore, may not be identical to their conformation in the crystallized form.

5. Minimum-energy search algorithms other than the Lamarckian genetic algorithm used in this work could be applied to this docking problem. Future work will use Monte Carlo/simulated annealing algorithms.

6. A variety of torsion and charge models could be applied to this problem, and future work will do so.

7. 3RUK has four chains, each with its own active site. The work described in this paper was performed on Chain A only. Chains B-D appear to have active sites highly similar to the Chain A active site. Future work will assess the ligand/receptor binding energies of Chains B-D.

8. CYP17A1 is a membrane-bound protein; 3RUK describes a conformation that is not bound to a membrane. The membrane-bound conformation of CYP17A1 may differ from the conformation in 3RUK.

5.0 References

[1] DeVore NM and Scott EE. Structure of cytochrome P450 17A1 with prostate cancer drugs abiraterone and TOK-001. *Nature* online pre-publication doi:10.1038/nature10743.

[2] Morris GM, Goodsell DS, Huey R, Lindstrom W, Hart WE, Kurowski S, Halliday

S, Belew R, and Olson AJ. *AutoDock Tools* v4.2. <http://autodock.scripps.edu/>. 2011.

[3] Protein Data Bank. PDB ID: 3RUK. DeVore NM and Scott EE. Structure of cytochrome P450 17A1 with prostate cancer drugs abiraterone and TOK-001. *Nature* online pre-publication doi:10.1038/nature10743.

[4] Huey R and Morris GM. *Using Autodock4 with AutoDock Tools: A Tutorial*. 8 January 2008.

[5] Miller WK and Auchus RJ. The molecular biology, biochemistry, and physiology of human steroidogenesis and its disorders. *Endocrine Reviews* 32 (2011), 81–151.

[6] Attard G, Reid, AH, Olmos D, and de Bono JS. Antitumor activity with CYP17 blockade indicates that castration-resistant prostate cancer frequently remains hormone driven. *Cancer Research* 69 (2009), 4937–4940.

[7] Yap TA, Carden CP, Attard G, and de Bono JS. Targeting CYP17: Established and novel approaches in prostate cancer. *Current Opinion in Pharmacology* 8 (2008), 449–457.

[8] Vasaitis TS, Bruno RD and Njar VC. CYP17 inhibitors for prostate cancer therapy. *Journal of Steroid Biochemistry and Molecular Biology* 125 (2011), 23–31.

[9] de Bono JS et al. Abiraterone and increased survival in metastatic prostate cancer. *New England Journal of Medicine* 364 (2011), 1995–2005.

[10] Attard G et al. Phase I clinical trial of a selective inhibitor of CYP17, abiraterone acetate, confirms that castration-resistant prostate cancer commonly remains hormone driven. *Journal of Clinical Oncology* 26 (2008), 4563–4571.

[11] Brodie A, Njar V, Macedo LF, Vasitis TS and Sabnis G. The Coffey Lecture: steroidogenic enzyme inhibitors and hormone dependent cancer. *Urologic Oncology* 27 (2009), 53–63.

The Lymph Node Lymphocytes First Humoral Immune Response as an AnyLogic Agent-Based Model

B. Khaldi¹ and F. Cherif¹

¹Department of Computer Sciences, University M^{ed} Khider, Biskra, Algeria

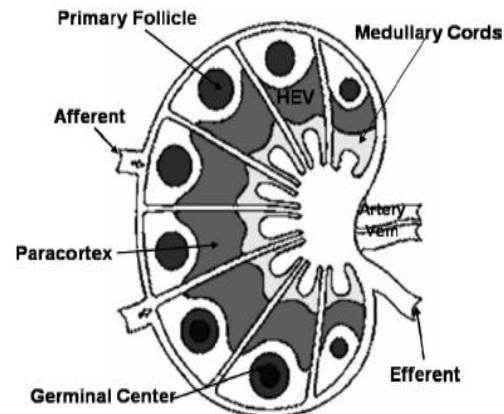
Abstract - In this paper we present a computational model for the first humoral immune response initiated in the Lymph Nodes organs against both T-Independent and T-Dependent antigens. The model is an AnyLogic Agent based model in which the behavior of the constitute agents are modeled using the Statecharts formalism. Using AnyLogic as an implementation platform wherein Statecharts can be programmed very conveniently; offers a great advantages especially final models can be modified, extended and handled in an elegant way. The results issued from our AnyLogic simulation respect several immunology experimentations (B-Cell activation, proliferation, differentiation and antibody generation).

Keywords: Simulation, AnyLogic, Multi-Agent system, Statecharts, First humoral immune response, Lymph Node.

1 Introduction

Lymph Nodes (LNs) [Figure 1] are a part of our secondary lymphoid organs that are filtering lymphatic fluid from bacteria, viruses, and foreign particles [1]. They are distributed at various points in the lymphatic system of our bodies and they are considered as sites that initiate and orchestrate the humoral immune response which refers to the production of antibodies and the accessory process that accompany it in response to antigens [2],[3]. These antigens are classified either [4] into T-Independent antigens, that can instantly mount a humoral immune response without the implication of T-Helper Cells, or T-Dependent antigens, that must implicate T-Helper cells to mount an humoral immune response.

The first humoral immune response results from the first exposure of an antigen. This last once it's recognized, it leads to the activation of unstimulated naïve B lymphocytes [2],[4],[5] that enter so on into the *clonal expansion* phase where large clones of identical cells are produced; the proliferating cells will then differentiate either into antibody-producing plasma cells or memory cells. Some of the antibody-producing cells migrate to the bone marrow and live in this site for several years; the others circulate in the blood and participate in the process of destructing or neutralizing antigens.



- Afferent lymphatic: drain lymph fluid from tissues, including antigen presenting cells (APC) and antigen from infected sites to the lymph node (LN).
- HEV (High Endothelial Venules): the capillary walls where T and B cells enter the LN from the blood. Paracortex: the T cell zone.
- Primary Follicles (PF): where B cells are localized, includes Follicular Dendritic Cells (FDC's).
- Germinal Center (GC): is formed when activated B cells proliferate in the PF.
- Medullary Cords: where plasma cells secrete Antibodies.
- Efferent lymphatics: the only exit from the LN, where activated or re-circulating T and B cells, as well as antibodies (Ab's) leave the LN and join the blood circulation.

Figure 1: Lymph Node schematic structure [1].

2 Modeling

The simulation of such immune system is extremely complex due to the high mechanisms and interactions being behind these systems; however great efforts are taking place to better understand these mechanisms and interactions. The researches that have been issued during the last years to simulate the immune system as a whole system or as a part of it such as LNs varies from mathematical simulation models [6],[7],[8] to Cellular Automata models (CA) [9],[10],[11] to Multi-Agent based models (ABM) [12],[13] and finally to the Reactive Animation (RA) models [1],[14],[15]. The RA models, which aim to couple between: state-of-the-art reactivity and

state-of-the-art animation [14],[16], are the recent modeling methods having used in biology simulation. They are based on two combined techniques: the Statecharts formalism [17] to model the system's behavior and the front-end animation tool to visualize the animation simulation with enabling natural-looking. The most well-known immunology works based on this technique is the David Harel's works: modeling the maturation of T-cells in the thymus [14] and the development of the lymph node [1]. This last studies the dynamic development of the LN with a focus on the behavior of a subset of immune cells that enter a single 2-dimentionel LN with immunogenic antigens.

2.1 Model development

We have focused heir on modeling the first humoral immune response initiated in the LN as an encountered antigen (either a T-Independent or a T-Dependent one) is recognized. Our model is an ABM [18],[19],[20],[21] that is defined as a computational model aiming to offer a manner on how to build complex systems composed of autonomous interacting computing elements called agents.

The behavior of our constitute agents; that are modeled with regard to their behavior, movement, location, and interactions; is modeled via the Statecharts formalism which is an essentially Finite State Machines diagram extended into a modular, highly structured, and economical description language [22],[23],[24].

The modeled agents are on continuous movement between the different LN zones wherein the agents are initially situated and wherefrom they enter or live them, a simplified process of the whole modeled system from the time that B-Cells enter the LN to the moment that they exit it is illustrated in [Figure 2].

The presented model is developed under the AnyLogic [25] simulation tool which is a Java based multi-approach simulation modeling tool based on advanced technologies such as UML, hybrid systems theory, and best numerical methods. The AnyLogic simulation tool provides great features such: reducing development cost and time, developing more models with one tool and improving the visual Impact of models.

Our AnyLogic simulation takes into account the three correlated modeling activities suggested in [19]: “*the behavior module*”, concerns modeling the agent behaviors; “*the environment module*”, defines the virtual place wherein the agents evolve and interact; and finally, “*the scheduling module*” which is related to the definition of how the two above modules are coupled and managed with taking into account the time factor.

2.1.1 Modeling the Time

Time is an important factor in every biologic system and in certain cases it's too long to model directly the entire processes that take place during these systems. Immune

response is one of these biologic systems for which we should be careful when simulating the time of its processes. In our simulation, in addition to attempt calculating the corresponding time values, we have tried to keep the relative times between its different processes correct; for that we have used the AnyLogic simulated time unit (TU) which is fixed to (0.001) and which in our model corresponds to 1 second so an hour is evaluated to (0.36 UT). In immunology many processes, for which time is an important factor, were examined; for example: a typical lymphocyte circulation cycle takes 12–24 hours [1] (so: 4.32 – 8.64 UT); normal proliferation takes 8–12 hours [1] (2.88 – 4.32 UT); etc.

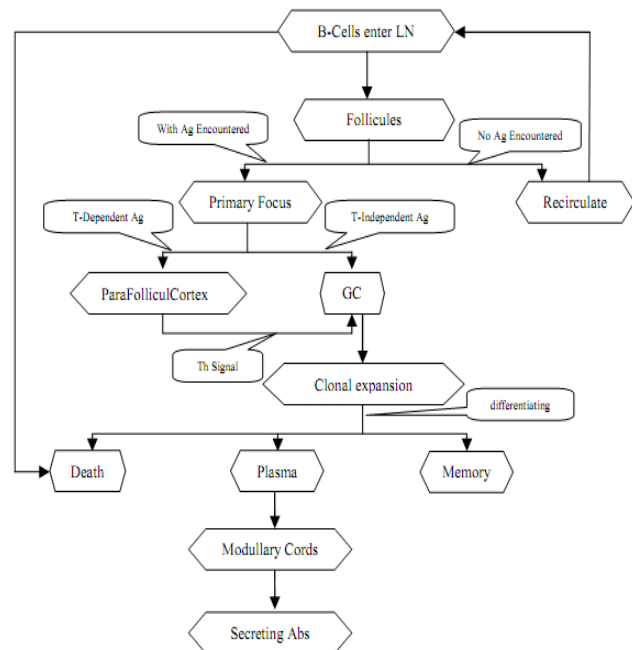


Figure 2: Simplified view of the process launched in the LN.

2.1.2 Modeling the LN environment

In our AnyLogic model, we have tried to let the simulation realistic. For that it's suitable to use a real image of a LN (taken from [26]) which can show its different constitute regions. To model these regions we have used a set of closed curves each represents a special LN areas. The agents can move continuously between these areas with regards to the biologic experiments.

In the main ActiveObject class of the simulation, we have developed a set of functions that is shared by all the agents indicating the movement to one zone to another, for example the function *moveToGC* (*AgentContinuous2D cell*) move a given cell from its current location to one of the modeled GCs. the path followed to reach the target location is specified automatically by the AnyLogic move API.

2.1.3 Modeling the Immune Cells agents

As we are simulating a part of the humoral immune response being initiated in the LN against both T-Dependent and T-Independent antigens, the captured developed agents include:

1. The Lymphocyte Agent: represents B-Cells, B-Memory Cells and T-Helper Cells.
2. The Plasma Agent,
3. The Antibody Agent,
4. And the Antigen Agent.

The behavior of each modeled agent is specified by the use of the AnyLogic integrated Statecharts formalism. Each behavior is divided into two main Statecharts: one to model the life cycle of the agent, the other model the location cycle of the agent.

2.1.3.1 Representative Example of our AnyLogic Agent Model: The Lymphocyte Agent

In this section we present the Lymphocyte agent which is an example of one of our Anylogic agents giving all its most detailed properties, methods and behaviors. The lymphocyte agent models different lymphocytes including B-Cells, Memory B-Cells, and T-Helper Cells. For each of the B-Cells and Memory B-Cells two kinds of cells are taken into account: ones are matured to recognize T-Dependent antigens, the others are matured to recognize T-Independent antigens; the T-Helper Cells are implicated only in the humoral immune response to T-Dependent antigens.

The biologic experimental [1],[2],[4],[5] illustrates that B-Cells enter permanently into the LN via HEVs and migrate to its zone area (Follicles) where they may meet antigens. If a B-Cell recognizes the encountered antigen,

the humoral immune response process will differ belongs to the presented antigen type:

If the antigen is a T-Independent one the B-Cell becomes an activated cell and migrates to the Follicle Center (FC) where it begins what the immunologist called colonal expansion phase. At this moment a large number of B-Cell clones are generated; some of them become a plasma cells, others become a memory B-Cells and the others are died.

Whereas if the type of presented antigen is a T-Dependent: the stimulated B-Cell will be completely activated when it migrates to the Paracortex zone and waits for an activated T-Helper Cell to interact with it, the B-Cell is then activated after a set of stimulating events interactions between them. The activated B-Cell will instantly take the same process of colonal expansion followed by a matured B-Cell activated by a T-Independent antigen.

All the lymphocyte cells which either recognize or do not recognize antigen live the LN via efferent zone than restart the recirculation process. They are death after an expiration of their life duration.

In our AnyLogic model we have developed a Lymphocyte class which extends the AnyLogicAgentContinuous2D subclass. The behavior of the Lymphocyte agent is specified alike the other modeled agents by using the AnyLogic integrated Statecharts formalism. We have used two main parallel Statecharts to model the behavior of the agent: one is for modeling its life cycle; the second is for modeling its location cycle. The Lymphocyte properties, methods and Statecharts behavior are illustrated in [Figure 3]. The whole AnyLogic models can be checked in [27].

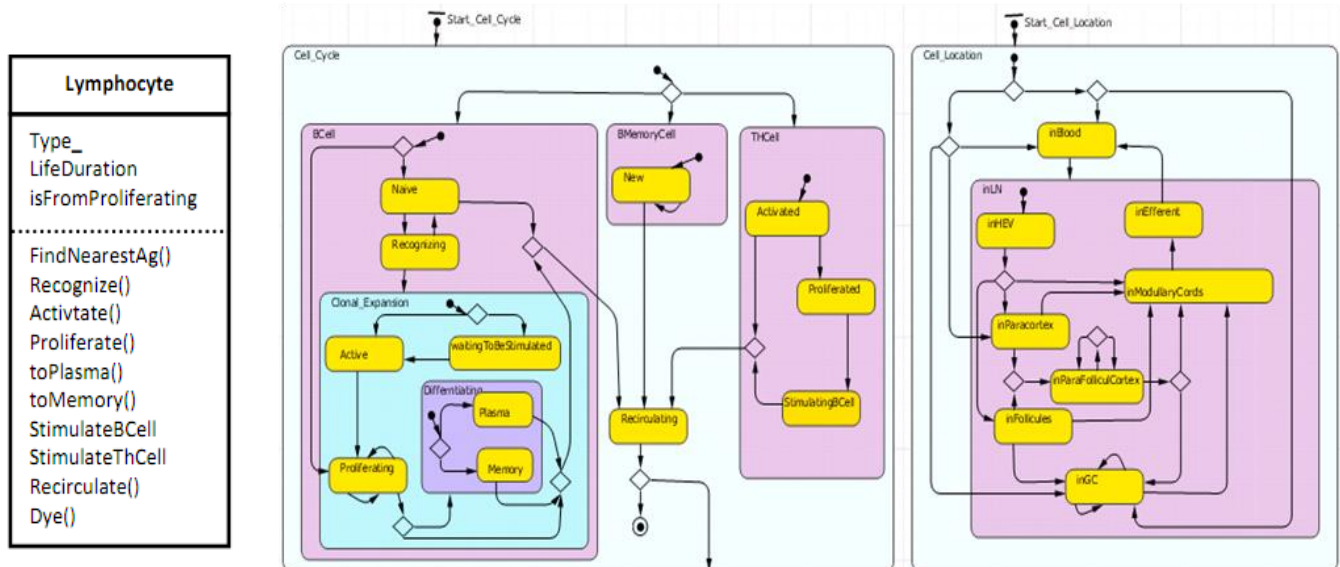


Figure 3: The Lymphocyte agent: properties, methods and Statecharts behavior.

3 Results

In this section we describe firstly the behavior of the simulator, and then show the type of results it generates. During a typical run of the simulator, a number of emergent behaviors can be seen that result from the rules of the model described in the previous section. At the beginning, the user defines the initial number of different immune agents that are implicated in the simulation, the agents include: Th-Cells, two kinds of antigens (T-Dependent antigens and T-Independent antigens), two kinds of B-Cells each kind is matured to recognize an antigen type, and as also by the same two kinds of B-Memory Cells, two kinds of plasma and two kinds of secreting antibodies. After all these parameters were specified, the user can switch to the root simulation that shows the initial allotment of the entire implicated agents in their LN zones. The running

simulation illustrates firstly the random distribution of the concerned defined agents, then the simulation begins showing the movement of each agent from its current location to its target zone with regards to the movement rules defined in its Statecharts location behavior. The user can interact whenever he wants with the simulation interface by a set of available controls: for example he can inject T-Independent antigens to the LN environment where they enter it from the afferent zone. Each antigen starts subsequently moving and if it's recognized by any specified B-Cell that is matured to recognize this kind of antigen, the humoral immune response will instantly begin processing from the activation of the specified B-Cells to the clonal expansion phase finished by plasma secreting antibody phase. In our model the details behind the antigen recognition phase isn't taken into account due to the extreme chemical interconnection signals known in this

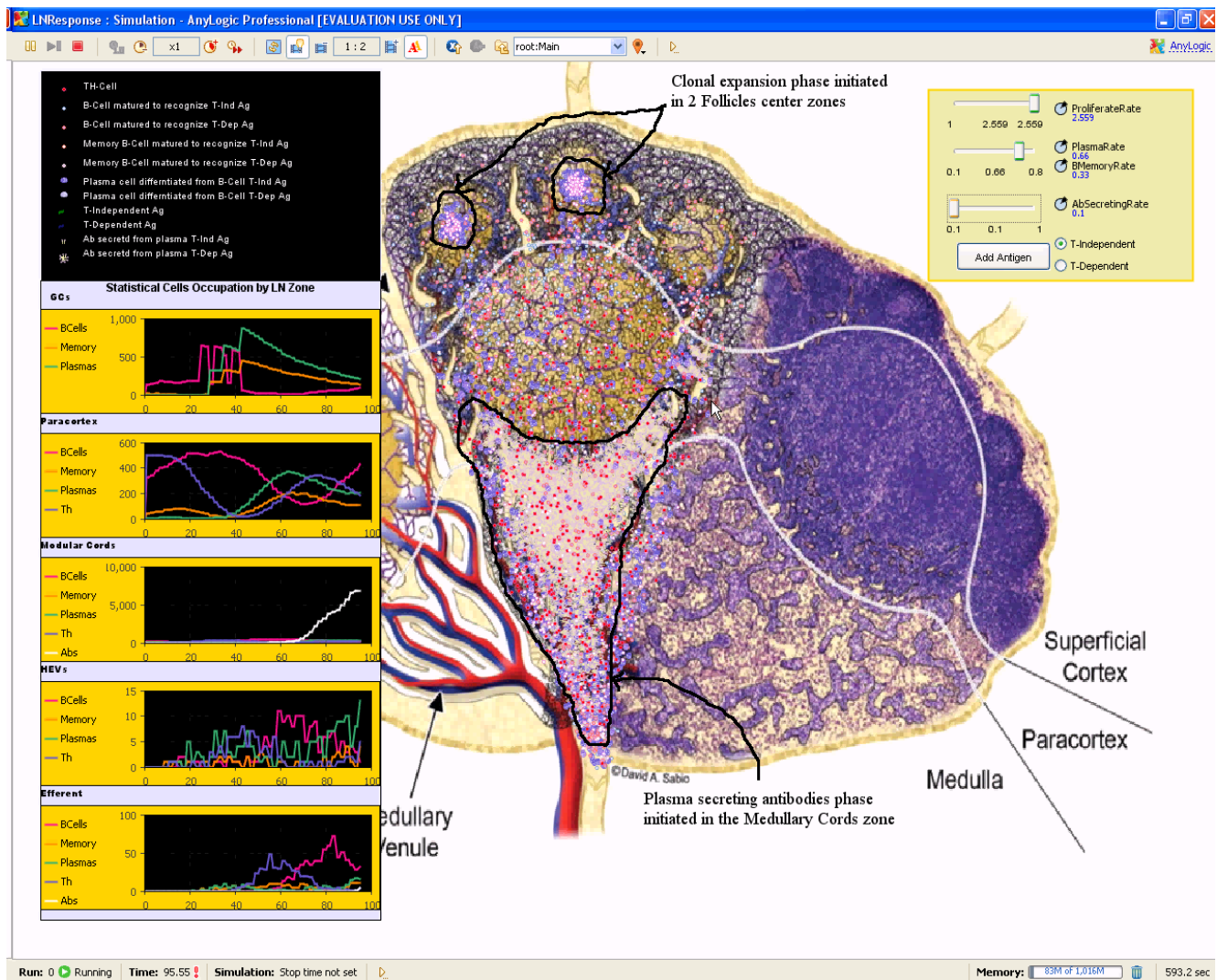


Figure 4: Simulation of a plasma secreting antibodies phase.

case between an antigen and a B-Cell; in consequence we have only develop an event that periodically calculate the distance between the current B-Cell and all the antigens cells; if the calculated distance is less or equal to two (2), the concerned B-Cell is then becoming in the active state of the *Clonal_Expansion* composite state that invokes an immediate migration of the concerned B-Cell to one of the modeled FCs and starts proliferating with a specific modified proliferate rate initially has the value $\rho=(24/6.24)*Ln(2)$ [28]. The proliferation process, which itself involves the creation of additional instances of the same object, is stopped when the number of the total proliferate B-cells exceeds the allowed total proliferated number which in our model assigned the value ($T_{prolif} = (2^p - 1) * 100$). The proliferated cells will then either dye with a probability of 1% [1] or differentiate either to Plasma cells or Memory B-Cells; the probability of this differentiating phase is defined by the user (the initial used probabilities are [1]: ($P_{plasma}=66\%$) to become a plasma cell and ($P_{mem}=33\%$) to become a Memory Cell). After that the generating cells migrate to the Medullary Cords zone where each plasma cell has a user defined probability initiated to 25% to begin secreting a huge number of antibodies (it's around 2000 antibodies are secreted every second for a few days [1]); in our model the total number of secreted antibodies is fixed to ($T_{ab}=(T_{prolif}*T_{plasma})/2$) per plasma cell for the reason of

the limitations of the computer' resources (memory and processor frequency) we have used for the simulation.

The result illustrated in [Figure 4] mentions a running simulation situation of complete plasma secreting antibodies phase launched after a previous initiation of two complete clonal expansion phases in response to encountered T-Independent antigens; these clonal expansion phases are taking place in two of the Follicles center zones of our modeled LN.

During the simulation the user can also modify the parameters that control the proliferation rate, differentiation rate and antibody secreting rate. Our model offers also to the user statistical analyses showing him in every time unit the occupation of the total number of each Cell per LN zone; the graphs viewed in the left side of the shown snapshot illustrate the occupation of: the both types of B-Cells, Memory Cells, Plasma Cells, Th-Cells and antibodies for each of: the GCs zone, the Paracortex zone, the Medullary cords zone, the HEVs zone and the Efferent zone.

The user has moreover a possibility to know the current state of any agent via the tools offered by the AnyLogic toolbar. For instance the snapshot mentioned in [Figure 5] highlights the current state of a given generated B-Cell matured to recognize a T-Independent antigen (remember that a B-Cell is a Lymphocyte instance class);

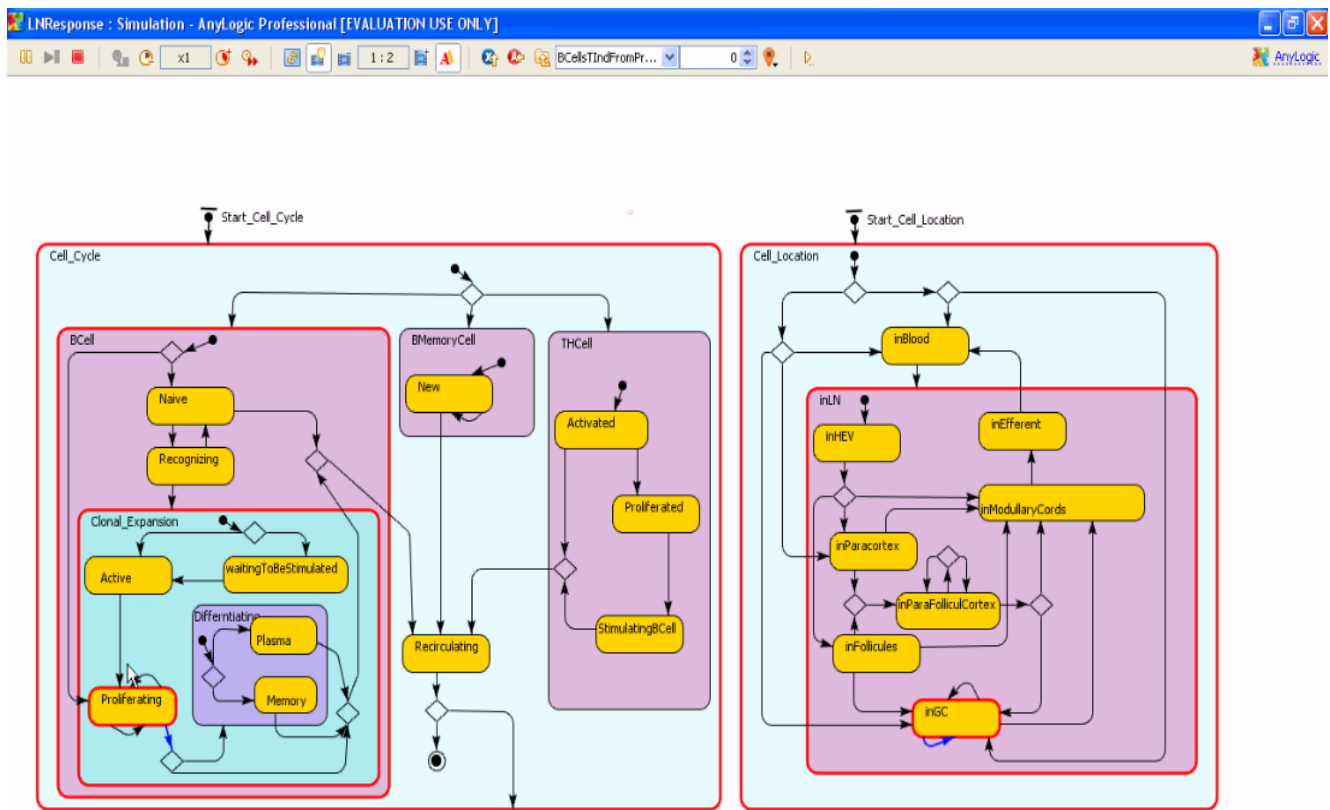


Figure 5: The current highlighted active state for a T-Independent B-Cell at run time.

the figure shows that the concerned Lymphocyte is actually on parallel composite states: the *Cell_cycle* one and the *Cell_Location* one. Inside the *Cell_Cycle* Statechart the Lymphocyte is currently in its *proliferating* state that belongs to the *Clonal_Expansion* state which is a sub-state of the *BCell* composite state; whereas inside the *Cell_Location* statechart, the figure shows that the specified Lymphocyte is actually in the *inLN* composite state wherein the current active state inside it is the *inGC* state.

4 Discussion

The study described in this issue demonstrates how we can use an Agent-Based approach for which every agent behavior is controlled completely by the Statecharts formalism to simulate a part of the first LN humoral immune response against antigens. The use of the Statecharts technique; that has been used together with a front-end-animation tool in the work of [1] to serve as an enlightenment of the manner on how a LN computational simulation can be translated into realistic animation; proves that it's a suitable and powerful visual modeling technique to be applied in biologic systems as they are considered as reactive systems.

The work presented heir; which is developed with the AnyLogic simulation tool; is considered as the first attempt in our Laboratory to model and simulate such biologic system using the AnyLogic environment.

The approach adopted heir is looking to profit from the work done in [1]; in which the Statecharts technique has used as a state-of-the-art reactivity to model the development of a LN; we have remodeled completely the LN using the AnyLogic simulation tool with regards to the immunological experimentations. Although we haven't model all the experimental details that are issued from the immunology researches due to its immense complexity, however our simulation results those are compared at run time with a real LN image are closes to the reality.

The results issued during the execution of our AnyLogic simulation model show that the process of mounting a LN humoral immune response against both T-Dependent and T-Independent antigens is well-fitted to the biologic experiments; all the phenomenon emerged from the application of the behavior rules defined in the Statecharts of each implicated cell are compared with the real images issued from the immunology experimentations. The obtained results demonstrate also that we were able to transform a part of these static experimental data into dynamical behavior including: cell migration from LN zone to another, cell proliferation, cell differentiation into memory or plasma cells and generation of antibody-producing plasma cells; statistical analyses of the dynamic occupancy of the different LN zones are also given to the final users in order to illustrate statistics about the total numbers of cells that are actually residing in each LN zone.

As a deep analyze of our LN AnyLogic model which has much been simplified due to the immense complexity of some immune mechanisms, and with regards to the LN model established in [1]; our model haven't detailed the cell interactions signals that can be viewed during an immune response. For example: the antigen-BCells interaction signals, the antigen-ThCell interaction signals and BCell-ThCell ones aren't carried out in our model. The model also doesn't take into account the orthogonal states feature used in the work of [1] for the reason that the AnyLogic simulation tool doesn't support in its professional used current release (6.5.1) this powerful Statecharts features; nevertheless we have simulated this feature on profiting from the ability of the AnyLogic simulation tool to create multi-statecharts for the same agent. These multi-statecharts can be executed on parallel manner with the same execution fashion of orthogonal states.

A positive view point of our model is that it deals with two kinds of antigens: T-Dependent and T-Independent, it's also developed with one simulation tool that can combine different modeling approaches at once, and which integrates also a 2D and 3D render engine that can animate the simulation in two or three dimensions. Contrary to the model of [1] which dials only with the T-Dependent antigens and it's developed using two different tools: the IBM Rhapsody developer tool to model the cells behavior and the Adobe Flash tool as a render engine to animate the cells behavior.

5 Conclusion & Future Works

In the AnyLogic model proposed in this paper we have focused on modeling the first humoral immune response initiated in the LN as an encountered antigen is recognized. The model dial with two kinds of antigens: the T-Independent antigens and the T-Dependent ones.

Although the simulation of such an immune response is very highly complex due to the high complexity of the mechanism behind it; we believe that we have succeeded to build a simplified AnyLogic model that models the first LN humoral immune response with taking into account a part of the immunology experimentations. Our results obtained during the execution simulation of the modeled system show that the model respects several immunology experimentations (B-Cell activation, proliferation, differentiation and antibody generation) even that some behaviors such as cell signal interactions, activation of Th-cells, etc., aren't carried on for which a perspective future work can be initiated to extend the model. The model also can be extended to tack into account the secondary humoral immune response that results from the second exposure or more of an antigen; in this case Memory B-Cells play the major factors in mounting such humoral immune response.

We hope also that other AnyLogic immune researches works can be initiated to involve the other immune organs such Spleen, Bone Marrow and Thymus for the aim to model the entire immune response by gathering piece to piece the models of each immune organ. This also can initiate a collaboration work between computing laboratories as a computing simulation research side with hospital immunology laboratories as biology research side.

Finally it would be a great pleasure for us that our AnyLogic model is the first attempt in our laboratory even in the entire world to initiate a simulation of a first LN humoral immune response against antigens using the AnyLogic simulation tool; as consequence we hope that we have enriched the existing immune models that have taken place and we have opened a research windows for the future extension of our work and for other researches area.

6 References

- [1] Naamah S., Irun R. C., and David H. "The Lymph Node B Cell Immune Response: Dynamic Analysis In-Silico", Proc. IEEE, 96(8), pp 1421-1443, 2008.
- [2] Abbas, A. K. and A. Lichtman H. "Basic Immunology: Functions and Disorders of the Immune system", Saunders Elsevier, Philadelphia, 8-20, 2004.
- [3] Dipankar D. & Luis F. Niño. "Immunological Computation Theory and Applications", CRC Press, Boca Raton, 8-22, 2009.
- [4] Sridhar R., "B cell activation and humoral immunity", URL:www.microrao.com/micronotes/pg/humoral_immunity.pdf.
- [5] Kitchen G., Horton-Szar D. , "Crash course: Immunology and hematology", Mosby Ltd, 1, 2007.
- [6] Nuno F.& al. "Agent Based Modeling and Simulation of the Immune System: a review", Evolutionary System and Biomedical Engineering Lab Systems and Robotics Institute Portugal. Unpublished
- [7] Stephanie F. & Catherine B. "Computer immunology", Immunological Reviews, 216(1), pp 176-197, 2007.
- [8] Vorgelegt Von Martin. "The Immune system as complex System: Description and simulation the interactions of its constituents", PhD dissertation, Dept Physic, University of Hamburg, Germany, 2001.
- [9] Celada F. and Seiden P. E.. "A Computer Model of Cellular Interactions in the Immune System", Immunol. Today, 13, pp 56-62, 1992.
- [10] Bernachi M., Castiglione F. & Succor S. "A parallel algorithm for the simulation of immune response". CytSeerx, pp 198-208, 1997.
- [11] <http://www.immunogrid.org/>, Accessed: Sept, 2011.
- [12] Meier-Schellersheim M. & Mack G., "Simune, a tool for simulating and analyzing immune system behavior", URL:<http://www.citebase.org/abstract?id=oai:arXiv.org:cs/9903017>.
- [13] Christina Elizabeth W. "Modeling intercellular interactions in the peripheral immune system", PhD dissertation, University of New Mexico, 2004.
- [14] Sol E., David H., and Irun R. Cohen. "Toward Rigorous Comprehension of Biological Complexity: Modeling, Execution, and Visualization of Thymic T-Cell Maturation", Genome Research, 13, pp 2485-2497, 2003.
- [15] Yaki S, Irun R. Cohen and David H. "Four-Dimensional Reactive Animation Model for the Early Stages of Pancreatic Organogenesis", National Academy of Sciences, 105(51), pp 20374-20379, 2008.
- [16] Irun R. Cohen and David H. "Explaining a complex living system: dynamics, multi-scaling and emergence", J. R. Soc. Interface. 4(13), pp 175-182, 2007.
- [17] Hila A., Avital S., Irun R. Cohen and David H. "GemCell A generic platform for modeling multi-cellular biological systems", Theoretical Computer Science, 391(3), pp 276-290, 2008.
- [18] José M Vidal: "Fundamentals of Multiagent Systems with NetLogo Examples", p 155, 2010.
- [19] Adeline M. Uhrmacher Danny Weyns. "Multi-Agent Systems Simulation and Applications", New York, USA: CRC Press, Taylor and Francis Group, 2009.
- [20] Michael W. and Nicholas R. J "Intelligent Agents: theory and practice", The knowledge Engineering Review, 10(2), pp 115-152, 1995.
- [21] Ferber, J. "Les systèmes Multi-Agents: Vers une intelligence collective", Paris: InterEditions, 1995.
- [22] Russell, S. and Norvig, P. "Artificial Intelligence: A Modern Approach", Prentice Hall (3rd edt), 2009.
- [23] David H. "Statecharts: A Visual Formalism for Complex Systems", Science of Computer Programming. 8(3), pp 231-274, 1987.
- [24] David H. "Statecharts in the Making: A Personal Account", Communications of the ACM 67, 52(3), March 2009.
- [25] <http://www.xjtek.com>, Accessed: Sept 2011.
- [26] Cynthia L. Willard-Mack. "Normal Structure, Function, and Histology of Lymph Nodes", Toxicologic Pathology, 34, pp 409-424, 2006.
- [27] Khaldi B. "Computer Simulation of an Immune Response against Virus Infection using Artificial Life Techniques", Magister thesis, Dept of Computing sciences, University of Med Khider, Biskra Algeria, 2012.
- [28] Paul S. Andrews and Jon T. "A Computational Model of Degeneracy in a Lymph Node", J. Springer, 4163(1), pp 164-177, 2006.

Alternative Two Sample Tests in Bioinformatics

Xiaohui Zhong and Kevin Daimi

Department of Mathematics, Computer Science and Software Engineering
University of Detroit Mercy,
4001 McNichols Road, Detroit, MI 48221
{zhongk, daimikj}@udmercy.edu

Abstract— Bioinformatics is a multidisciplinary field. Statistics is getting immense popularity in bioinformatics research. The goal of this paper is to introduce a survey of two sample tests applied to bioinformatics. The vast majority of these methods do not follow the classical two sample test techniques, which require strict assumptions. Thus, unlike other classical surveys, this paper will emphasize the justifications behind the deviations from the standard approach, and the implementation of such deviations.

Index Terms— Statistical Methods, Sequence Analysis, Microarray, Two-sample Testing, Bootstrap Hypothesis Testing, Non-traditional Hypothesis Testing

I. INTRODUCTION

Bioinformatics is a rapidly growing discipline that has matured from the fields of Molecular Biology, Computer Science, mathematics, and Statistics. It refers to the use of computers to store, compare, retrieve, analyze and predict the sequence or the structure of molecules. According to Cohen [2], “The underlying motivation for many of the bioinformatics approaches is the evolution of organisms and the complexity of working with incomplete and noisy data.” Bioinformatics is a multidisciplinary field in which teams from Biology, Biochemistry, Mathematics, Computer Science, and Statistics work together to stipulate perception into the functions of the cell [3], and [10]. More precisely, Bioinformatics is the marriage between the fields of biology and computer science together in order to analyze biological data and consequently solve biological problems [12].

The need for collaboration in bioinformatics research and teaching is inevitable. “The explosive increase in biological information produced by large-scale genome sequencing and gene/protein expression projects has created a demand that greatly exceeds the demand for researchers trained both in biology and in computer science” [4]. According to the European Bioinformatics Institute [5], “Bioinformatics is an interdisciplinary research area that is the interface between the biological and computational sciences. The ultimate goal of bioinformatics is to uncover the wealth of biological information hidden in the mass of data and obtain a

clearer insight into the fundamental biology of organisms. This new knowledge could have profound impacts on fields as varied as human health, agriculture, environment, energy and biotechnology.”

The field of statistics plays a vital role in bioinformatics. Modified statistical techniques are being constantly evolving. Statistics is the science of collection, organization, presentation, analysis, and interpretation of data. [16], [18]. Statistical methods which summarize and present data is referred to as descriptive statistics. Data modeling methods that account for randomness and uncertainty in the observations and drawing inferences about the population of interest lie within the inferential statistics. When the focus is on the biological and health science information, biostatistics is applicable [18].

The techniques of statistics that have been applied include hypothesis test, ANOVA, Bayesian method, Mann–Whitney test method, and regressions tailored mainly to microarray data sets, which take into account multiple comparisons or cluster analysis and beyond. In bioinformatics, microarrays readily lend themselves to statistics resulting in a number of techniques being applied [15], [22]. The above mentioned methods assess statistical power based on the variation present in the data and the number of experimental replicates. They even help to minimize Type I and type II errors in demanding analysis. While these methods sound familiar to people with statistics background, they might be foreign to researchers in the field of bioinformatics. On the other hand, statisticians will enjoy the benefit of seeing how these techniques are being applied to the field of bioinformatics when getting to know what DNA sequences or protein sequences are.

This paper aims to survey some basic statistical techniques, especially different kinds of hypothesis testing techniques that have been developed lately and used in the context of bioinformatics. The goal of this survey is to pinpoint the motivations for modifying the classical two-sample tests when applied to bioinformatics by researchers. The classical two-sample tests have strict assumptions. The reason that forced researchers to relax or violate some of these assumptions will be explored.

II. CLASSICAL TWO-SAMPLE t -TESTS

The classical two-sample t -test has been applied to only few bioinformatics problems. The reason for that should be clear shortly. An example is the following scenario. When measuring the level of gene expression in a segment of DNA, the process usually requires several repeated experiments in order to obtain the measurements of one cell type. This is due to biological and experimental variability. The objective is to compare the levels of the gene expression between two types of DNA based on the measured levels of gene expressions for these two types of DNA's. Such a procedure is a typical classical two sample t -test. Assuming that $M_{t,i}$ are the measurements from type $t=1,2$ respectively, with $1 \leq i \leq n_t$, the null hypothesis $H_0: \mu_1 = \mu_2$ is tested with alternative hypothesis $\mu_1 \neq \mu_2$. The appropriate test statistics is

$$t = \frac{(\bar{M}_1 - \bar{M}_2) \sqrt{n_1 n_2}}{S \sqrt{n_1 + n_2}}, \quad (1)$$

$$\text{where } S = \left(\frac{\sum_{t=1}^2 \sum_{i=1}^{n_t} (M_{t,i} - \bar{M}_1)^2}{n_1 + n_2 - 2} \right)^{\frac{1}{2}}.$$

Using the assumptions that $M_{t,i}$ are n_t $NID(\mu_t, \sigma_t^2)$ random variables, the statistics t follows a t -distribution with degrees of freedom $n_1 + n_2 - 2$ if the null hypothesis is true. While this test procedure is very simple, it requires very strict assumptions. Some or all of these assumptions cannot be met in real life applications. In some cases, it is either not known or hard to confirm whether the variables $M_{t,i}$ are normally distributed. If they are normally distributed, then the requirement of both normal populations sharing a common variance could be hard to fulfill. Another requirement to be satisfied mandates these variables to be independent, which is generally true in many gene expressions measurements. In practice, some or all of these conditions are not satisfied, but the decision on the equality of two means is still needed. Thus, alternatives to this standard classical t -test are required. In this paper, we will survey several modified tests appearing in recent bioinformatics literature.

III. TWO SAMPLE TEST WITH INTRA-DEPENDENCY

Gilbert et al [9] compared the genetic diversity of the virus between two groups of children who were infected with HIV at birth. The children were classified into a group of 9 slow/non-progressors (group 1) and a group of 12 progressors (group 2). Between 3 to 7 HIV gag P17 sequences were sampled from each child and pair-

wise sequence distances were derived for each child's sample as the measures of diversity within a child. The goal was to assess whether the level of HIV genetic diversity differed between the two groups in order to help identify the role of viral evolution in HIV pathogenesis. In what follows, we will show why the authors have to deviate from the standard two sample test. We will first introduce and explain their statistical model.

Let M_{kij}^g represent the distance between sequence i and j of child k in group g , $g=1$ or 2 . It was found that if a sequence is involved in two distances of a child's sequences, then the two distances are positively correlated. Also the contrasts involving common individual are also positively correlated. Therefore, the conditions for a classical t -test described in section II is violated. This will force the application of this procedure to produce bias results. The natural option is to perform the test based on a subset of independent samples in which not all the information is fully considered. Thus, a new two-sample test that took account of the correlations between samples was proposed. The detail is described as follows:

Assume that there are n_g children from group g , $g=1$ or 2 respectively, and child k has m_k^g sequences sampled. Then there are $N_g = \sum_{k=1}^{n_g} m_k^g (m_k^g - 1) / 2$ many pair-wise distances from each group. There are also $Q^g = \sum_{k=1}^{n_g} 2(m_k^g - 2)$ many covariances between the distances for the individuals in each group. The test statistics is similar to (1) above, $t = \frac{\bar{M}^1 - \bar{M}^2}{\sigma(\bar{M}^1 - \bar{M}^2)}$. The

main idea is to estimate the standard deviation $\sigma(\bar{M}^1 - \bar{M}^2)$ with the correlations between M_{kij}^g , assuming the null hypothesis $H_0: \mu_1 = \mu_2$ is true. Here, the mean distances are defined as $\bar{M}^g = (N_g)^{-1} \sum_{k=1}^{n_g} \sum_{i < j} M_{kij}^g$. It is noticeable that the correlation only occurs within the group and particularly within individuals, so the estimate of the variance within one group can be discussed without indexing on g and k . Since there are $n(n-1)/2$ pair-wise distances, the standard estimate for the variance of \bar{M} is

$$\hat{\sigma}^2(\bar{M}) = (n(n-1)/2 - 1)^{-1} \sum_{i < j} (M_{ij} - \bar{M})^2.$$

But this estimate is too small because it did not account for the positive correlations between distances sharing the same sequences. Another option is

$\hat{\sigma}^2(\bar{M}) = (n-1)^{-1} \sum_{i<j} (M_{ij} - \bar{M})^2$. However, this one is too large unless the correlations between the sequences are perfectly linear. Therefore, something in between these two estimates could be a more accurate estimate of the variance. Because the correlation only occurs between the pair-wise distances sharing the same sequence, this variance can be estimated by calculating the covariance in two parts:

$$\hat{\sigma}^2(\bar{M}) = (n(n-1)/2)^{-1} \{2(n-2)\sigma_1^2 + \sigma_2^2\}$$

where $\sigma_1^2 = \text{cov}(M_{ij}, M_{il})$ is the covariance of the pair-wise distances that share the same sequence, and $\sigma_2^2 = \text{var}(M_{ij})$ is the variance of all pair-wise distances.

The empirical estimates of these two variances are:

$$\hat{\sigma}_1^2 = \frac{\sum_{i<j<l} \{(M_{ij} - \bar{M})(M_{il} - \bar{M}) + (M_{ij} - \bar{M})(M_{jl} - \bar{M})\}}{n(n-1)(n-2)/3-1}, \quad (2)$$

$$\hat{\sigma}_2^2 = (n(n-1)/2-1)^{-1} \sum_{i<j} (M_{ij} - \bar{M})^2. \quad (3)$$

Since there are two groups, the estimate can be modified to

$$\hat{\sigma}^2(\bar{M}^1 - \bar{M}^2) = \sum_{g=1}^2 N_g^{-1} \sum_{k=1}^{n_g} \{2(m_k^g - 2)\hat{\sigma}_{g,1}^2 + \hat{\sigma}_{g,2}^2\} \quad (4)$$

where

$$\begin{aligned} \hat{\sigma}_{g1}^2 &= \sum_{k=1}^{n_g} (m_k^g (m_k^g - 1)(m_k^g - 2)/3 - 1)^{-1} \\ & \left(\sum_{i<j<l} \{(M_{kij}^g - \bar{M}^g)(M_{kil}^g - \bar{M}^g) \right. \\ & \left. + (M_{kij}^g - \bar{M}^g)(M_{kij}^g - \bar{M}^g)\} \right) \end{aligned}$$

and

$$\hat{\sigma}_{g2}^2 = (N_g - 1)^{-1} \sum_{k=1}^{n_g} \sum_{i<j} (M_{kij}^g - \bar{M}^g)^2$$

Modified this way, the test statistics $t = \frac{(\bar{M}^1 - \bar{M}^2)}{\hat{\sigma}(\bar{M}^1 - \bar{M}^2)}$

will have asymptotic normal distribution, provided that

$$\sum_{g=1}^2 \frac{N_g(N_g-1)/2}{2(N_g-2)\rho_g^2+1}$$

is large enough, where ρ_g is the correlation coefficient of the two pair-wise distances sharing the same sequence in group g .

The authors provided the comparative results for the DNA sequences of the 21 children described earlier. Classical two sample t -test was performed on the differences based on synonymous distance with sample means $\bar{D}^1 = -0.0113$ and $\bar{D}^2 = 0.00713$, and sample sizes $N_1 = 387$ and $N_2 = 523$ respectively. The result suggested a difference between the two groups with $p = 2.2 \times 10^{-6}$. However, it was estimated that the correlations of the pair-wise distances within individuals are $\rho_1 = 0.55$ and $\rho_2 = 0.61$ respectively. The classical t -test ignored these positives correlations, which resulted in a smaller estimated variance for the difference of the means. Thus, the newly developed procedure was applied producing $p = 0.56$, which indicates that the difference between the mean distances of the two groups is not significant.

The above two-sample test method provided an alternative to the traditional two-sample t -test to accommodate the situation where data within the group may be correlated. This approach will have significant impact on many areas. First, a new method for the existing statistical tests is introduced. This method not only can be applied in the area of bioinformatics, but can also be applied in other fields, such as finance, engineering, chemistry, and behavior science. Most important, it can have distinct significance in the bioinformatics domain. For example, in the analysis of DNA sequences [6], one of the tasks is to test the similarity or differentially expressed genes of two sequences by matching the subsequences. One of the assumptions for such matching rules is that the occurrences of the nucleotides must be independent. Such an assumption was found to be inaccurate in many DNA sequences. This method provides an alternative formula for the test statistics by calculating the variance of the mean of data that might be dependent on each other. Furthermore, the method for calculating the variance can be extended to building statistical models from data that might be interdependent.

IV. BOOTSTRAP AND PERMUTATION METHODS

The test discussed in last section dealt with comparing means from two samples. With the advancements of biology and other bioscience, collections of microscopic DNA spots attached to a solid surface called microarrays are studied. With the power of computation, scientists use DNA microarrays

to measure the expression levels of large numbers of genes simultaneously. One of their objectives is to detect differentially expressed genes between two types of cells.

Suppose we have two types of cells. Associated with each cell are a number of microarrays. Let the number of microarrays be n_1 and n_2 respectively. The n_1 arrays contain m genes from the first type of cells, and the n_2 arrays have m genes from the second type. Let M_{ij}^c be the expression value of the i th gene in the j th array in cell c , $c = 1, 2$. Let t_i , $i = 1, 2, \dots, m$ be the two sample test statistics calculated using formula (1). The null hypothesis for each test is $H_{i0} : \mu_{i1} = \mu_{i2}$. When this hypothesis is being rejected as a positive result, the two genes will be differentially expressed. Assuming the cumulative distribution function of t_i is $D_i(t)$ when the null hypotheses are true, the p -value of each test can be calculated as $p_i = (1 - D_i(|t_i|) + D_i(-|t_i|))$. These p -values are arranged in ascending order $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$. Any gene tested with a p -value below certain threshold will be rejected (indicating the test is positive). These genes are ranked in the order of p -values with the smallest value as the most significant for further study. The remaining task is to find the distributions $D_i(t)$.

There are many different ways to identify these distribution functions. Under the classical assumptions that all M_{ij}^c are normally identically distributed, the distributions are either student t -distribution or standard normal distribution. As discussed in the last section, such an assumption is either unrealistic or difficult to verify. As a result of increasing computing power, resampling methods, such as permutation and bootstrap methods are being widely used. These methods generate empirical distributions D_i , which are also the distributions of p_i .

The classical bootstrapping/permutation resampling scheme is described as follows.

- Calculate the test statistics from the original sample t_i for each gene using formula (1).
- All $n_1 + n_2$ arrays are put in the same pool. The n_1 arrays are randomly drawn to be assigned to type 1 cell, and n_2 arrays are randomly drawn to be assigned to type 2 cell.
- If the draws are with replacement, the bootstrap method will be used. If the draws are without replacement, the permutation method will be applied.
- Repeat the above steps B times. In the case of permutation, not all possible permutations have to be

considered. In this paper, the two methods will be treated similarly.

- Calculate the t -statistic t_i^b , $i = 1, 2, \dots, B$ using formula (1) for each sample.
- Under the null hypotheses that there is no differentially expressed gene, the t -statistics should have the same distribution regardless of how the arrays are arranged. Hence, the empirical p -values can be calculated by:

$$p_i = \frac{1}{B} \sum_{b=1}^B \frac{\#\{j : |t_j^b| \geq |t_i|, j = 1, 2, \dots, m\}}{m} \quad (5)$$

This scheme was discussed and applied in a number of papers [1], [8], [13], [19]-[21]. It also has another alternative described in [13] as Posterior Mixing Scheme:

- Resample the n_1 arrays from type 1 cell and place on type 1 cell, and resample n_2 arrays from type II cell and place on type II cell.
- Using the data in question, calculate t_{i1}^b , $i = 1, 2, \dots, B$ for each sample using formula (1).
- Repeat the above two steps on the array from type II cell and obtain t_{i2}^b , $i = 1, 2, \dots, B$. Finally, calculate

$$t_i^b = \sqrt{\frac{n_1}{n_1 + n_2}} t_{i1}^b + \sqrt{\frac{n_2}{n_1 + n_2}} t_{i2}^b. \quad (6)$$

Then p_i 's are calculated with formula (5). It was concluded that the Posterior Mixing Scheme will have better power $[1 - P(\text{type II error})]$ than the classical one. To our knowledge, this formula for calculating the test statistics has not been employed in the bioinformatics literature yet. The formula should be appealing to researchers to further investigate and validate it, and obtain more accurate results for identifying differentially expressed genes.

Mukherjee et al [17] took the bootstrap method for calculating these statistics a step further. From the bootstrap schemes described above, the bootstrap, t_i 's are assumed to be normally distributed with empirical mean $\mu_{t_i} = \frac{1}{B} \sum_{b=1}^B t_i^b$ and standard deviation σ . Formula (5) was not used to calculate the p -values. Instead the expected p -value was calculated by

$$\hat{p}_i = E(p_i) = \int_{-\infty}^{+\infty} (1 - D_i(|x|) + D_i(-|x|)) G(x, \mu_{t_i}, \sigma) dx$$

where D_i is the cumulative distribution function of t_i and G is the Gaussian.

This procedure was applied to some widely analyzed microarray data with D_i replaced by t -distribution with degrees of freedom $n_1 + n_2 - 2$, and the variance σ^2 was set between 1 and 3. Results of ranking on the genes by this proposed bootstrap method and classical two sample t -test were compared. It was found that the genes identified to be differentially expressed were subsequently confirmed by further costly test to rank an average of 25.5 places higher than genes ranked by the classical method [17]. This shows that the bootstrap method provides a powerful alternative to the classical method by estimating the p -values more accurately.

Bootstrap two sample test is widely used by many researches in identifying the differentially expressed genes. This method is particularly suitable for the cases when the underlining distributions are unknown. For example, Troyanskaya et al [21] used this procedure to perform 50,000 permutation on a data set comprised of normal lung and squamous cell lung tumor specimens with the Bonferroni correction p -values. The result of this method was compared to the result of rank sum test and ideal discriminator method. It was concluded that the bootstrap two sample test is most appropriate for a high-sensitivity test [21]. Many other researchers, such as Pan [19], Ge [8], and Abul [1] also used this method as an integral part of their more comprehensive study of microarrays.

The procedure of bootstrapping requires intensive computation. Computer packages/algorithms are also developed to tackle the issues related to computation time, storage and efficiency. Li et al [14] developed an algorithm, Fast Pval, to efficiently calculate very low p -values from large number of resampled data. The software package, SAFEGUI, was designed to bootstrap resampling t -tests for testing gene categories [7].

V. MULTIPLE TESTING WITH Q-VALUES

Modified two-sample tests, and bootstrap two-sample tests introduced in the last section concentrated on finding the p -values of the test so that genes can be ranked accordingly. Notice that the p -value is only the probability that the test statistic falls in the critical region controlled by the maximum tolerance of Type I error for one test. In the case of multiple tests, such as the gene expressions in microarrays, the Type I error can be inflated. For example, assume that 1000 genes are represented in each array of the two types of cells, and 20 out of the 1000 genes are differentially expressed. To find these 20 genes, two-sample t -tests are performed among 1000 pairs of genes using a p -

value for 5% Type I error. This will produce 49 [5% of (1000-20)] miss-identified genes, which is even more than the actual differentially expressed genes. Thus, the Type I error for the entire array is greater than 5%, which is undesirable result. A well-known classical procedure to correct this problem is the Bonferroni correction by replacing the cut-off Type I error α by $\frac{\alpha}{m}$ where m is the total number of tests [6], [21]. For $m = 1000$, Type I error becomes 0.0005, which forces the test to miss most of the significantly differentially expressed genes. Actually, the possible outcomes of any multiple tests can be described in the tabular format (Table 1) below. The numbers in parentheses represent the intended scenario.

A variety of measurement schemes in the development of procedures dealing with microarray data were proposed. These include *Per-comparison error rate* (PCER), *Family-wise error rate* (FWER), *False discovery rate* (FDR), and *positive False discovery rate* (pFDR). They are stated as [8]:

$$\begin{aligned} \text{PCER} &= \frac{E(V)}{m} \\ \text{FWER} &= \Pr(V > 0) \\ \text{FDR} &= E\left(\frac{V}{R} \mid R > 0\right) \Pr(R > 0) \\ \text{pFDR} &= E\left(\frac{V}{R} \mid R > 0\right) \end{aligned}$$

TABLE 1
POSSIBLE OUTCOMES FOR 1000 GENES WITH 5% P-VALUE

	Tested Positive	Tested Negative	Total
Genes are significantly different	$R - V$ (19)	$m_s - (R - V)$ (1)	m_s (20)
Genes are not significantly different	V (false positive) 49	$m_v - V$ (931)	m_v (980)
Total	R (68)	$m - R$ (932)	m (1000)

Among these four measures, the most commonly used is the pFDR. Since this quantity is only meaningful and useful when R is positive, this rate is usually written as $\text{FDR} = E\left(\frac{V}{R}\right)$ which is the symbol used here. A

straightforward estimate for FDR is $FDR = \frac{V}{R}$, which represents the ratio of number of false positive and the total tested positive. While the traditional multiple tests have to deal with thousands of test with only one cut-off value for the p -values, the false discovery rate takes into account the joint behavior of all the p -values. The false discovery rate is therefore a useful measure of the overall accuracy of a set of significant tests. We will discuss a method using a q -value developed by Storey et al [20]. The q -value method took into consideration the FDR balancing the identification of as many significant features as possible, while keeping a relatively low proportion of false positives. This method and an important application of this method [1] will be discussed below.

A value similar to the p -value is defined by Storey et al [20] as the q -value corresponding to a particular p -value. Assume the p -values for each test are calculated as p_i by one of the methods introduced in previous sections. Then the q -value is calculated by:

$$q(p_i) = \min_{p_i \leq \lambda \leq 1} FDR(\lambda) = \min_{p_i \leq \lambda \leq 1} \left\{ \frac{V(\lambda)}{S(\lambda)} \right\}$$

where $V(\lambda) = \#\{\text{false positive} \mid p_i \leq \lambda, i = 1, 2, \dots, m\}$, and $S(\lambda) = \#\{p_i \leq \lambda, i = 1, 2, \dots, m\}$. The objective is to simultaneously control the q -value and the p -value so that the FDR will not be out of proportion. A procedure for finding the q -values and the criteria for selecting the threshold in a sequential procedure are described below [20].

1) Assume the test statistics are calculated by (1), with p -values p_i calculated by (5), for $i = 1, 2, \dots, m$.

2) Arrange the p -values in ascending order $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$, which is also the order of genes in terms of their order against the null hypotheses.

3) Use one of the options described below to estimate the value of $\hat{\pi}_0$.

4) Estimate $q(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 m t}{\#\{p_i \leq t\}} = p_{(m)}$

5) For $j = m-1, m-2, \dots, 1$, estimate

$$\begin{aligned} q(p_{(j)}) &= \min_{t \geq p_{(j)}} \left\{ \frac{\hat{\pi}_0 m t}{\#\{p_i \leq t\}}, q(p_{(j+1)}) \right\} \\ &= \min \left\{ \frac{\hat{\pi}_0 m p_{(j)}}{j}, q(p_{(j+1)}) \right\} \end{aligned}$$

Now, two lists for p -values and q -values are simultaneously formed:

$$\begin{aligned} p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)} \\ q(p_{(1)}) \leq q(p_{(2)}) \leq \dots \leq q(p_{(m)}) \end{aligned}$$

One can select the maximum index $1 \leq k \leq m$ in the above lists so that both p -values and q -values up to k th gene will satisfy both thresholds.

The quantity π_0 in step 3 is the proportion of null genes (no differences between the two cell types) of the total number m of genes tested. Despite the fact of having a difficult task to deal with, three different ways have been developed to estimate this quantity [20].

A. Rule of Thumb Method

Let $\pi_0 = \frac{\#\{p_i > \lambda\}}{m(1-\lambda)}$ for some λ , $0 < \lambda < 1$.

The rationale for this estimate is that the null p -values are uniformly distributed after certain value, λ . A simple rule of thumb is choosing $\lambda = 0.5$. This implies that the value of π_0 is estimated by

$$\hat{\pi}_0 = \frac{\#\{p_i > 0.5\}}{0.5m}$$

B. Bootstrap Method

Assumed that all p -values are calculated from the original set of data. Calculate $\pi_0(\lambda_k) = \frac{\#\{p_i > \lambda_k\}}{m(1-\lambda_k)}$

for $\lambda_k = k\Delta\lambda$, $k = 0, 1, \dots, M$, $\Delta = \frac{\lambda_{\max}}{M}$, $0 < \lambda_{\max} < 1$

from these p -values. Here, λ_{\max} is close to 1 and M is the number of desired points. Let $\pi = \min_{0 \leq k \leq M} \{\pi_0(\lambda_k)\}$. Resample the data B times,

calculating $\pi_0^b(\lambda_k) = \frac{\#\{p_i^b > \lambda_k\}}{m(1-\lambda_k)}$ from the

bootstrap p -values for all λ_k each time. Define the mean square error to be:

$$MSE(\lambda_k) = \sqrt{\frac{\sum_{b=1}^B (\pi_0^b(\lambda_k) - \pi)^2}{B}}$$

Then the estimate of the proportion of null genes will be:

$\hat{\pi}_0 = \min_{\lambda_k \in \Gamma} \{\pi_0(\lambda_k), 1\}$, where Γ is the collection of λ_k s such that $MSE(\lambda_k)$ is minimum. A simple algorithm was given in [1].

C. Curve Fitting Method

The ideal estimate for $\pi_0(\lambda)$ is $\pi_0(\lambda_{\max})$, where λ_{\max} is close to 1 since genes should be null in this region. However, the value of $\pi_0(\lambda)$ is very sensitive to change of λ . To obtain a stable estimate, a natural cubic spline $f(\lambda)$ is suggested to be fitted to the points $\{(\lambda_k, \pi(\lambda_k)) \mid \lambda_k \in \{0, \Delta\lambda, \dots, \lambda_{\max}\}\}$, the estimate is $\hat{\pi}_0 = f(1)$. There were two suggestions for fitting the curve. Storey et al [20] suggested that the curve fitting should be weighted by $(1 - \lambda)$ to control the instability near 1. However, Abul et al [1] suggested that the result with no weighting is better to avoid underestimation. For any new set of data, both weighted and un-weighted fitting should be tried and the better estimate used.

The procedure of estimating π_0 was extended to one-sided hypothesis [1] with some adjustments. For example, if the tests are right-sided (up-regulated), the formula for the t -statistics remains the same as (1). The corresponding p -values can also be calculated by the bootstrap process described in last section. However, formula (5) for calculating the p -values should be modified to

$$p_i = \frac{1}{B} \sum_{b=1}^B \frac{\#\{j : t_j^b \geq t_i, j = 1, 2, \dots, m\}}{m}. \quad (5')$$

This change will make $\lim_{\lambda \rightarrow 1} \pi_0(\lambda) > 1$, which is meaningless. The adjustment will be to set λ_{\max} as the upper bound of λ for which $\pi_0(\lambda) \leq 1$. This results in $\lambda_{\max} = \sup\{0 < \lambda < 1 \mid \pi_0(\lambda) \leq 1\}$. Bootstrap or curve fitting will be deployed to estimate $\hat{\pi}_0$, which is needed for finding the q -values.

Experiments on some artificial data demonstrated that this approach could provide very accurate estimates. The procedure described above can guide researchers to fine tune the selection of genes for further experiments. By

bounding false-discoveries, the amount of wasted time and cost can also be bounded with the same rate of false-discoveries beforehand. This procedure has many applications in microarray experiments and gene analysis.

VI. CONCLUSIONS

Bioinformatics is being used in many fields such as molecular medicine, preventative medicine, gene therapy, drug development, and waste cleanup. The interdisciplinary nature of bioinformatics demands close collaboration between biologists, computer scientist, mathematicians, and statisticians. Statistics is playing a significant role in various applications of bioinformatics. One of the important areas of statistics that has been heavily used is two sample tests. These tests classically have rigorous postulations. Researcher involved in bioinformatics concluded that these tests are not readily suitable for their work due to the nature of many of the bioinformatics applications. Consequently, they were forced to weaken some/all of these postulations. This paper surveyed a number of methods that pushed researcher to diminish these constraints. Assumptions that were relaxed and the reasons behind this relaxation were demonstrated.

It is our future goal to introduce studies dealing with variation of formulas for two sample tests, variety methods of controlling the false discovery rate, such as selecting proper sample size, methods taking into account the dependency of sample data, and extension of these techniques to multi-sample testing. While many of these techniques were proposed based on certain set of data or artificial data, work needs to be done on different data sets to validate the results. More importantly, statisticians can help in seeking theoretical justification or support for these methods. Computer scientists can assist in developing more efficient algorithms to implement these techniques. It is hoped that these methods can spark new ideas in the future research in bioinformatics.

REFERENCES

- [1] O. Abul, R Alhaji, and F Polat, "A Powerful Approach for Effective Finding of Significantly Differentially Expressed Genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 3, No. 3, pp. 220-231, 2006.
- [2] J. Cohen, "Bioinformatics: An Introduction for Computer Scientists," *ACM Computing Surveys*, Vol. 36, No. 2, pp. 122-158, 2004.
- [3] J. Cohen, "Computer Science and Bioinformatics," *Communications of the ACM*, Vol. 48, No. 3, pp. 72-78, 2004.

- [4] Editorial, "Training for Bioinformatics and Computational Biology," *Bioinformatics*, Vol. 17, No. 9, pp. 761-762, 2001.
- [5] European Bioinformatics Institute, Available: <http://www.ebi.ac.uk/2can/home.html>.
- [6] W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics: An Introduction*, New York: Springer-Verlag, 2001.
- [7] D. M. Gatti, M. Sypa, I. Rusyn, F. A. Wright, and W. T. Barry, "SAFEGUI: Resampling-Based Tests of Categorical Significance in Gene Expression Data Made Easy," *Bioinformatics*, Vol. 25, No. 4, pp. 541-542, 2009.
- [8] Y. Ge, S. Dudoit, and T. P. Speed, "Resampling-Based Multiple Testing for Microarray Data Analysis," Dept. of Statistics, University of California, Berkeley, Tech. Rep. 633, 2003.
- [9] P. B. Gilbert, A. J. Rossini, and R. Shankarappa, "Two-Sample Tests for Comparing Intra-Individual Genetic Sequence Diversity between Populations," *Biometrics*, Vol. 61, No. 1, pp. 106-117, 2005.
- [10] S. Gopal, A. Haake, R. P. Jones, and P. Tymann, *Bioinformatics: A Computing Perspective*, New York: McGraw Hill, 2009.
- [11] Y. Ji, Y. Lu and G. Mills, "Bayesian Models Based on Test Statistics for Multiple Hypothesis Testing Problems," *Bioinformatics*, Vol. 24, No.7, pp. 943-949, 2008.
- [12] M. LeBlanc, and B. Dyer, "Bioinformatics and Computing Curricula 2001: Why Computer Science is Well Positioned in a Post Genomic World," *ACM SIGCSE Bulletin*, Vol. 36, No. 4, pp. 64-68, 2004.
- [13] S. Lele and E. Carlstein, "Two-Sample Bootstrap Tests: When to Mix?" Department of Statistics, University of Carolina at Chapel Hill, Tech. Rep. 2031.
- [14] M. J. Li, P. C. Sham, and J. Wang, "FastPval: A Fast and Memory Efficient Program to Calculate Very Low P-values from Empirical Distribution," *Bioinformatics*, Vol. 26, No. 22, pp. 2897-2899, 2010.
- [15] P. Liu, J. T. Hwang, "Quick Calculations for Sample Size While Controlling False Discovery Rate with Application to Microarray Analysis," *Bioinformatics*, Vol. 23, No. 6, pp. 739-746, 2007.
- [16] V. Mantzapolis, and X. Zhong, *Probability and Statistics*, Dubuque: Kendall Hunt Publishing Company, 2010.
- [17] S. N. Mukherjee, P. Sykacek, S. J. Roberts, and S. J. Gurr. "Gene Ranking Using Bootstrapped P-Values," *SIGKDD Explorations*, Vol. 5, No. 2, pp. 16-22, 2003.
- [18] M. Pagano, and K. Gauvreau, *Principles of Biostatistics*, Belmont: Brooks/Cole, 2000.
- [19] W. Pan, "A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments," *Bioinformatics*, Vol. 18, No. 4, pp. 546-554, 2002.
- [20] J. Storey and R. Tibshirani, "Statistical Significance for Genome-wide Experiments," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 100, No. 16, pp. 9440-9445, 2003.
- [21] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric Methods for Identifying Differentially Expressed Genes in Microarray Data," *Bioinformatics*, Vol. 18, No. 11, pp.1454-1461, 2002.
- [22] Y. Zhao, and W. Pan, Modified Nonparametric Approaches to Detecting Differently Expressed Genes in Replicated Microarray Experiments, *Bioinformatics*, Vol. 19, No. 9, pp. 1046-1054, 2003.

Simulated Docking of Oseltamivir with the 2009 Pandemic Strain Influenza A/H1N1 Neuraminidase Active Site

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA
email: jhorner@cybermesa.com

Abstract

Influenza neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Oseltamivir is the most widely used neuraminidase inhibitor. Here I provide a computational docking analysis of oseltamivir with the active site of the neuraminidase of the 2009 Influenza A/H1N1 strain. The computed inhibitor/receptor binding energy suggests that oseltamivir would be effective against that strain.

Keywords: Influenza, H1N1, neuraminidase, oseltamivir

1.0 Introduction

Influenza neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Oseltamivir ((3R,4R,5S)-4-(acetilamino)-5-amino-3-(pentan-3-yloxy)cyclohex-1-ene-1-carboxylic acid; [4]) is the most widely used influenza therapeutic.

In the World Health Organization serotype-based influenza taxonomy, influenza type A has nine neuraminidase-related sero-subtypes, and these subtypes correspond at least roughly to differences in the active-site structures of the flu neuraminidases. The subtypes fall into two groups ([3]): group-1 contains the subtypes N1, N4, N5 and N8; group-2 contains the subtypes N2, N3, N6, N7 and N9. Oseltamivir was designed to target the group-2 neuraminidases.

The available crystal structures of the group-1 N1, N4 and N8 neuraminidases ([1]) reveal that the active sites of these enzymes have a very different three-

dimensional structure from that of group-2 enzymes. The differences lie in a loop of amino acids known as the "150-loop", which in the group-1 neuraminidases has a conformation that opens a cavity not present in the group-2 neuraminidases. The 150-loop contains an amino acid designated Asp 151; the side chain of this amino acid has a carboxylic acid that, in group-1 enzymes, points away from the active site as a result of the 'open' conformation of the 150-loop. The side chain of another active-site amino acid, Glu 119, also has a different conformation in group-1 enzymes compared with the group-2 neuraminidases ([8]).

The Asp 151 and Glu 119 amino-acid side chains form critical interactions with neuraminidase inhibitors. For neuraminidase subtypes with the "open conformation" 150-loop, the side chains of these amino acids might not have the precise alignment required to bind inhibitors tightly ([8]). The active site of the 1918 H1N1 strain has the 150-loop configuration.


```

axisangle0 random          # initial orientation
dihe0 random              # initial dihedrals (relative) or random
tstep 2.0                 # translation step/A
qstep 50.0               # quaternion step/deg
dstep 50.0               # torsion step/deg
torsdof 7                # torsional degrees of freedom
rmstol 2.0               # cluster_tolerance/A
extnrg 1000.0            # external grid energy
e0max 0.0 10000          # max initial energy; max number of retries
ga_pop_size 150          # number of individuals in population
ga_num_evals 2500000     # maximum number of energy evaluations
ga_num_generations 27000 # maximum number of generations
ga_elitism 1             # number of top individuals to survive to next
generation
ga_mutation_rate 0.02    # rate of gene mutation
ga_crossover_rate 0.8    # rate of crossover
ga_window_size 10       #
ga_cauchy_alpha 0.0     # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0     # Beta parameter Cauchy distribution
set_ga                  # set the above parameters for GA or LGA
sw_max_its 300          # iterations of Solis & Wets local search
sw_max_succ 4           # consecutive successes before changing rho
sw_max_fail 4           # consecutive failures before changing rho
sw_rho 1.0              # size of local search space to sample
sw_lb_rho 0.01          # lower bound on rho
ls_search_freq 0.06     # probability of performing local search on
individual
set_pswl                # set the above pseudo-Solis & Wets parameters
unbound_model bound     # state of unbound ligand
ga_run 10               # do this many hybrid GA-LS runs
analysis                # perform a ranked cluster analysis

```

Figure 1. ADT parameters for the docking in this study

3.0 Results

The interactive problem setup, which assumes familiarity with the general neuraminidase "landscape", took about 20 minutes in ADT; the docking proper, about 28 minutes on the platform described in Section 2.0. The platform's performance monitor suggested that the calculation was more or less uniformly distributed across the four processors at ~25% of peak per

processor (with occasional bursts to 40% of peak), and required a constant 2.9 GB of memory.

Figure 2 shows the best-fit oseltamivir/receptor energy and position summary produced by ADT under the setup shown in Figure 1. The estimated free energy of binding under these conditions is ~ -8.5 kcal/mol; the estimated inhibition constant, ~603 nanoMolar at 298 K.

```

MODEL          8
USER          Run = 8
USER          Cluster Rank = 1
USER          Number of conformations in this cluster = 10
USER
USER          RMSD from reference structure          = 146.946 A
USER
USER          Estimated Free Energy of Binding      = -8.49 kcal/mol  [(1)+(2)+(3)-(4)]
USER          Estimated Inhibition Constant, Ki    = 603.08 nM (nanomolar)  [Temperature =
298.15 K]
USER
USER          (1) Final Intermolecular Energy      = -10.57 kcal/mol
USER          vdW + Hbond + desolv Energy          = -6.89 kcal/mol
USER          Electrostatic Energy                 = -3.68 kcal/mol
USER          (2) Final Total Internal Energy      = -1.06 kcal/mol
USER          (3) Torsional Free Energy           = +2.09 kcal/mol
USER          (4) Unbound System's Energy  [(2)]   = -1.06 kcal/mol
USER
USER
USER          DPF = 3TI3_oseltamivir.dpf
USER          NEWDPF move      oseltamivir.pdbqt
USER          NEWDPF about     0.529200 81.163696 109.114304
USER          NEWDPF tran0     30.119561 14.578922 -20.694645
USER          NEWDPF axisangle0 0.665691 -0.552718 0.501357 -102.866417
USER          NEWDPF quaternion0 0.520492 -0.432160 0.392002 -0.623427
USER          NEWDPF dihe0     168.80 -157.63 -177.46 -10.23 -55.65 -0.70 28.39
USER
USER
USER          x          y          z          vdW      Elec          q          RMS
ATOM          1  C2  G39  A  800      29.602  13.195 -22.935 -0.22 +0.06      +0.091 146.946
ATOM          2  C3  G39  A  800      31.208  13.233 -22.632 -0.25 +0.00      +0.050 146.946
ATOM          3  C4  G39  A  800      31.725  14.395 -21.669 -0.22 -0.05      +0.209 146.946
ATOM          4  C5  G39  A  800      30.777  14.481 -20.473 -0.19 +0.03      +0.143 146.946
ATOM          5  C6  G39  A  800      29.308  14.818 -20.993 -0.17 +0.05      +0.147 146.946
ATOM          6  C7  G39  A  800      28.741  13.977 -22.105 -0.21 +0.03      +0.049 146.946
ATOM          7  O7  G39  A  800      28.408  14.858 -19.795 -0.01 -0.25      +0.379 146.946
ATOM          8  C8  G39  A  800      27.326  15.852 -19.571 -0.24 +0.09      +0.121 146.946
ATOM          9  C9  G39  A  800      27.103  16.932 -20.666 -0.40 +0.01      +0.027 146.946
ATOM          10 C91 G39  A  800      26.896  18.375 -20.179 -0.45 +0.01      +0.007 146.946
ATOM          11 C81 G39  A  800      26.079  15.023 -19.329 -0.28 +0.02      +0.027 146.946
ATOM          12 C82 G39  A  800      25.448  14.581 -20.611 -0.36 +0.00      +0.007 146.946
ATOM          13 N5  G39  A  800      31.316  15.593 -19.600 -0.06 -0.14      -0.352 146.946
ATOM          14 H5  G39  A  800      31.504  16.508 -20.010 +0.10 +0.01      +0.163 146.946
ATOM          15 C10 G39  A  800      31.552  15.389 -18.289 -0.28 +0.19      +0.214 146.946
ATOM          16 C11 G39  A  800      32.087  16.540 -17.517 -0.30 +0.11      +0.117 146.946
ATOM          17 O10 G39  A  800      31.350  14.297 -17.682 -0.76 -0.42      -0.274 146.946
ATOM          18 N4  G39  A  800      33.088  14.075 -21.248 -0.18 +0.04      -0.073 146.946
ATOM          19 H42 G39  A  800      33.671  13.752 -22.021 -0.11 -0.43      +0.274 146.946
ATOM          20 H41 G39  A  800      33.480  14.890 -20.776 +0.18 -0.03      +0.274 146.946
ATOM          21 H43 G39  A  800      33.133  13.232 -20.676 -0.27 -0.24      +0.274 146.946
ATOM          22 C1  G39  A  800      29.038  12.409 -24.007 -0.24 +0.29      +0.177 146.946
ATOM          23 O1B G39  A  800      27.789  12.427 -24.260 -1.03 -1.58      -0.648 146.946
ATOM          24 O1A G39  A  800      29.818  11.689 -24.695 -0.96 -1.48      -0.648 146.946
TER
ENDMDL

```

Figure 2. ADT's oseltamivir energy and position predictions.

Figure 3 is a rendering of the active-site/inhibitor configuration computed in this study.

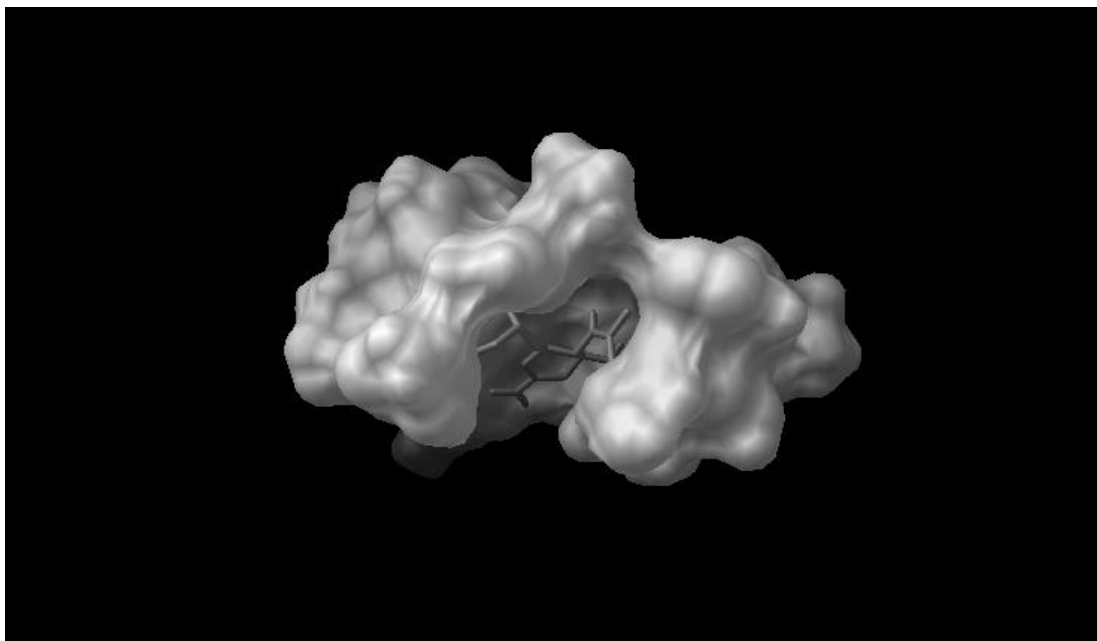


Figure 3. Rendering of oseltamivir computationally docked with the active site of PDB 3TI3. The molecular surface of the receptor is shown in white; the inhibitor, in stick form in grey. Only the interior, inhibitor-containing region of the molecular surface of the active site can be compared to *in situ* data: the surface distal to the interior is a computational artifact, generated by the assumption that active site is detached from the rest of the receptor.

The distances between ligand and receptor atoms in 3TI3, and the corresponding distances in the present computation were within 10% of each other.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The inhibition constant computed in this study (~603 nanoMolar at ~298 K) is much smaller than the inhibition constant of neuraminidase inhibitors that are not clinically effective ([10], [11], [13], [14], [15]) against several H1N1 genotypes. This suggests that oseltamivir would be effective

against Influenza A/California/04/2009(H1N1)).

2. The docking study reported here assumes that the receptor is rigid. This assumption is appropriate for the binding energy computation for PDB 3TI3 per se. However, the calculation does not reflect what receptor "flexing" could contribute to the interaction of the ligand with native unliganded receptor.

3. The analysis described in Sections 2.0 and 3.0 assumes receptor is in a crystallized form. *In situ*, at physiologically normal temperatures (~310 K), the receptor is not in crystallized form. The ligand/receptor conformation *in situ*, therefore, may not be identical to their conformation in the crystallized form.

4. Minimum-energy search algorithms other than the Lamarckian genetic algorithm used in this work could be applied to this docking problem. Future work will use Monte Carlo/simulated annealing algorithms.

5. A variety of torsion and charge models could be applied to this problem, and future work will do so.

6. 3TI3 has two chains, each with its own active site. The work described in this paper was performed on Chain A only. Chain B appears to have an active site highly similar to the Chain A active site. Future work will assess the ligand/receptor binding energies of Chains B.

5.0 Acknowledgements

This work benefited from discussions with Tony Pawlicki. For any problems that remain, I am solely responsible.

6.0 References.

[1] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49.

[2] Johnson NP and Mueller J. Updating the accounts: global mortality of the 1918-1920 "Spanish " influenza pandemic. *Bulletin of the History of Medicine* 76 (2002), 105-115.

[3] World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization* 58 (1980), 585-591.

[4] Vavricka CF, Li Q, Wu Y, Qi J, Wang M, Liu Y, Gao F, Liu J, Feng E, He J, Wang J, Liu H, Jiang H, and Gao GF: Structural

and functional analysis of laninamivir and its octanoate prodrug reveals group specific mechanisms for Influenza NA inhibition. *PLoS Pathogens* 7 (October 2011): e1002249. doi:10.1371/journal.ppat.1002249.

[5] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.

[6] PDB ID = 10.2210/pdb3ti3/pdb. See also [4].

[7] US Centers for Disease Control. *Summary: Interim Recommendations for the Use of Influenza Antiviral Medications in the Setting of Oseltamivir Resistance among Circulating Influenza A (H1N1) Viruses, 2008-09 Influenza Season.* 19 December 2008. URL <http://www.cdc.gov/flu/professionals/antivirals/summary.htm>.

[8] Luo M. Structural biology: antiviral drugs fit for a purpose. *Nature* 443 (7 September 2006), 37-38. doi:10.1038/443037a, published online 6 September 2006.

[9] Morris GM, Goodsell DS, Huey R, Lindstrom W, Hart WE, Kurowski S, Halliday S, Belew R, and Olson AJ. *AutoDock* v4.2. <http://autodock.scripps.edu/>. 2010.

[10] Drug Bank. *Zanamivir*. <http://www.drugbank.ca/drugs/APRD00378>.

[11] Govorkova EA et al. Comparison of efficacies of RWJ-270201, zanamivir, and oseltamivir against H5N1, H9N2, and other avian influenza viruses. *Antimicrobial Agents and Chemotherapy* 45 (2001), 2723-2732.

[12] Huey R and Morris GM. *Using AutoDock 4 with AutoDock Tools: A Tutorial.* 8 January 2008. <http://autodock.scripps.edu/>.

[13] Cheng Y and Prusoff WH. Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. *Biochemical Pharmacology* 22 (December 1973), 3099–3108. doi:10.1016/0006-2952(73)90196-2.

[14] Horner JK. Simulated docking of oseltamivir with the 1918 pandemic strain Influenza A/H1N1 zanamivir-conformed

neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. 2011. pp. 130-135.

[15] Horner JK. Simulated docking of zanamivir with the 1918 pandemic strain Influenza A/H1N1 neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. pp. 136-142.

Simulation of an Implantable Microphone in the Middle Ear Cavity

Sang-Hyo Arman Woo¹, Ji Soon Park², Ji Min Kim³,
Woon Hwan Na², and Byung Seop Song⁴

¹BK21 Team, College of Rehabilitation Science, Daegu University, Republic of Korea,

²Dept. of Rehabilitation Psychology, Daegu University, Republic of Korea,

³Dept. of Vocational rehabilitation, Daegu University, Republic of Korea,

⁴Dept. of Rehabilitation Engineering, Daegu University, Republic of Korea.

Abstract - *With the advent of implantable hearing aids, the implementation and acoustic sensing strategy of the implantable microphone becomes an important issue. Previously, implantable microphones were inserted under the skin, which caused loud noise signals whilst touching or moving the skin. In this paper, a microphone was mounted in a hole drilled in the middle ear cavity and the acoustic signal was measured. Based on experiments, a simple finite element model was conducted to predict the optimal placement of the microphone. From experiments with guinea pigs (n=2), the loss of transmission observed from the proposed microphone was little as was as the total harmonic distortion. Furthermore, the simulation model predicted that there was no significant difference between the placements of the microphones.*

Keywords: Hearing aids, Microphones, Implantable biomedical devices, Acoustic, Transmission loss

1 Introduction

Hearing aids are devices that assist in providing a better understanding of communication for hearing-impaired patients. Over the past few decades, many implantable hearing aid devices have been developed which provide a better sound quality and treat hearing loss more effectively than conventional hearing aids [1-6]. The implantable hearing aids are composed of four parts: a microphone, an amplifier, a transducer and a battery. The microphone is the most challenging part to implement fully implantable because various noises are generated from the human body, and the implantable microphones are usually inserted under the skin meaning that the effects of ear canal resonance and pinna effects are not useable [7]. Furthermore, the skin dramatically decreases the sensitivity of the microphone at high frequencies because the mass of the membrane increases [8]. In order to overcome the sensitivity problem, implantable microphones are designed with a large size to increase sensitivity. However, patients and surgeons often request small microphones to aid in a fast recovery and safety. Jung et al used a resonance technique with the effects of the skin set to the maximum audible range [8]. In spite of above effect, placing the implantable microphone under the skin caused fundamental problems with skin

movement such as chewing food and combing hair can cause large sound noises [9]. This is because of skin movements and motion artifacts that cause large deformations of the membrane when the sound is transmitted into the skin.

In order to overcome the above barriers, methods involving measuring from the ear canal, tympanic membrane, or ossicular chain were proposed. Leysieffer et al. proposed the placement of the microphone inside the ear canal skin [10]. Although this method has the advantages of using pinna and can avoid skin movement noise, the associated surgical operation is very difficult and the sensitivity of the microphone is affected by the skin density and thickness.

Ko et al. proposed attaching a displacement or acceleration sensor onto the ossicular chain and measuring the vibration motions [11]. Surgical operations using this method are difficult because the ossicular chain is very fragile and it is hard to place proper attachments for drilling a hole because of the limited locations. Esteem® proposed to measure the displacement value from the malleus using a piezoelectric transducer (PZT) [12]. Surgical operations using this method are easier than previous methods, but it is still not easy to perform and can't be used for patients with middle ear mechanic problems.

In this paper, sensing the acoustic signals from the middle ear cavity (MEC) is proposed. Unlike previous methods, the proposed method does not require a difficult surgical operation, as the microphone can be directly mounted onto the bone after drilling a hole into the MEC. Simple finite element analysis was conducted to determine the optimal attachment placement for the microphone in the MEC and concluded that there would be no significant difference with the placement. A calibrated commercial microphone was packaged in a titanium screw and attached to the MEC. The transmission loss (TL) total harmonic distortion (THD), and vibration differences between skin implantable microphones and the implemented microphone were measured.

2 Methods

Figure 1 (a) illustrates how a skin implantable microphone is placed under the skin resulting in the deformation of the membrane when it is touched by the hand. Furthermore, the

hairs are displaced at the skin and cause large deformations while combing the hair. Figure 1 (b) illustrates the proposed method that places the microphone at the MEC. All implantable hearing aids require a hole to be drilled in the MEC, in order to place the transducer in various positions [13]. Therefore, the microphone can be placed in an existing hole to sense the sound so that some of the signal from the outside is transmitted into the tympanic membrane and the microphone can sense the signals from inside the MEC.

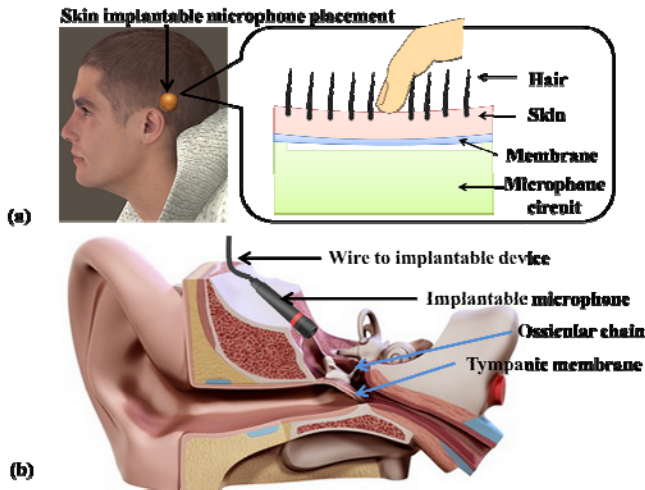


Figure 1. Illustrations of a proposed microphone. (a) The implantable microphone located under the skin and its weaknesses. (b) Proposed placement of the implantable microphone.

2.1 Simple FEM analysis of transmission loss from the tympanic membrane

Figure 2 (a) and (b) shows the generated meshes and tympanic membrane displacement results for the Finite Element Analysis (FEM). The ear canal was ignored because this experiment also does not measure. Although the shape of the MEC is a similar oval, it can be simplified to a plain cylinder. Total numbers of elements were 35,868 and average element quality was 0.93. 90dB SPL plane wave radiation source was applied to the tympanic membrane and all of the walls were set as hard boundaries, because the acoustic impedance of the bone was much higher than that of air. The tympanic membrane was set as a linear elastic solid model and the ossicular chains were ignored. COMSOL was used to calculate the multi physics between solid walls and acoustic transmissions. Figure 3 (b) shows the normalised surface displacement of the membrane and red signifying high displacement and blue indicating values close to zero. From the FEM simulation, the position of the microphone was not found to be important when the microphone is approximately 3mm from the membrane because sound reflects from the wall and becomes similar sound pressures. Figure 3 (a) shows the isosurface of the

sound pressure from a 5 kHz radiation source without a hard boundary, and (b) depicts the same surface with a hard boundary. The sound pressure near the membrane varies depending on the position, but it becomes similar when there is a hard boundary in comparison to without a hard boundary. Simulation results of other frequencies between 0.2 and 10 kHz also show a similar trend. Therefore, the position of the microphone is not critical because the bone of the MEC will work as hard boundary.

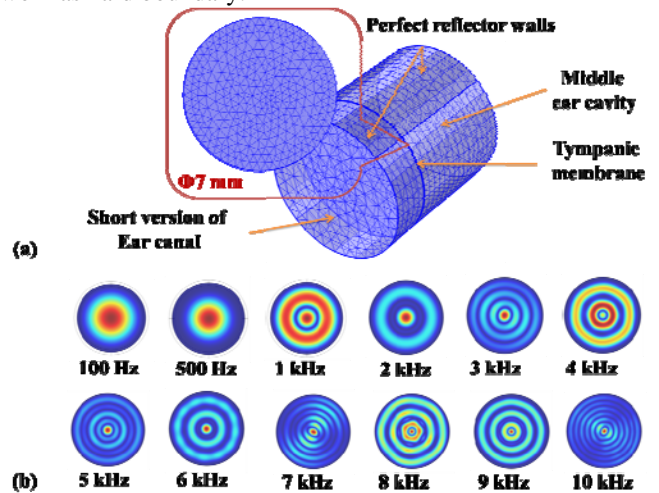


Figure 2. FEM analysis. (a) Simplified diagram of the MEC and generated meshes. (b) Displacement results of the tympanic membrane.

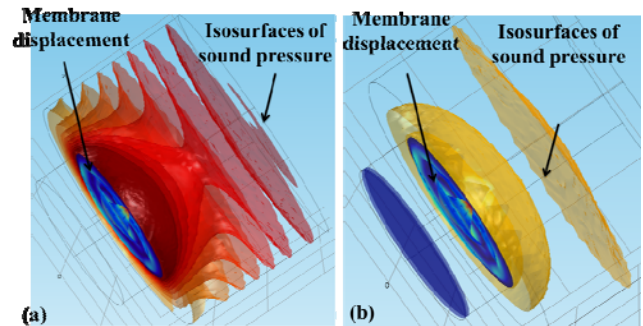


Figure 3. FEM analysis of isosurfaces of sound pressure at 5 kHz source. (a) Without hard boundary. (b) With hard boundary.

3 Experiments

Figure 4 (a) shows a simple block diagram of current experiment. The experiments were performed in a sound chamber (TL: 40dB) using an ER-1 speaker (Etymotic Research), which was calibrated. Although the speaker was calibrated, the actual sound pressure in front of the tympanic membrane was very different [14]. Therefore, a probe microphone (ER-7C, Etymotic Research, sensitivity: 50 mV/Pa) was used to measure the sound pressure in front of the tympanic membrane. An electret condenser microphone (ECM, BSE co., sensitivity: 23 mV/Pa), which was calibrated, was used to test

the implantable microphone. After pre-amplification, the signal was collected using LabVIEW with an AES17 20 kHz low-pass filter and a 200 Hz high pass filter. The high pass filter was set higher than for basic acoustic measurements because of the instability of the ECM sensor signal below 200 Hz. Sampling rate and period were set as 44.1 kHz and 1 second per sinusoidal wave, and exponentially increased 100 points for frequencies between 200 Hz and 10 kHz. A linear weighting and 1/3 octave analysis was used to measure the vibration level or noise level.

Figure 4 (b) shows animal experiment setup. Hartley guinea pigs (n=2) weighing between 130g and 180g were used in this study, and the experiments were performed according to the guidelines of the Committee on Animal Experimentation of Kyungpook National University. The guinea pigs were sacrificed with concentrated potassium chloride injection into their hearts and then the hairs were removed. A hole was drilled into the mastoid about 8mm from the ear canal in order to open the MEC. A microphone was mounted into a hole and fixed using polycarboxylate cement (HY-bond, SHOFU inc.). The membrane was observed to check for any problems at the tympanic membrane.

The microphone was fully covered with the cement to prevent the leakage. In addition, a vacuum sealing compound (HIVAC-G, ShinEtsu) was used to block sound leakage. The experiment was performed within two hours of death of the animal. All experiments were conducted two times and mean and standard deviation values were plotted.

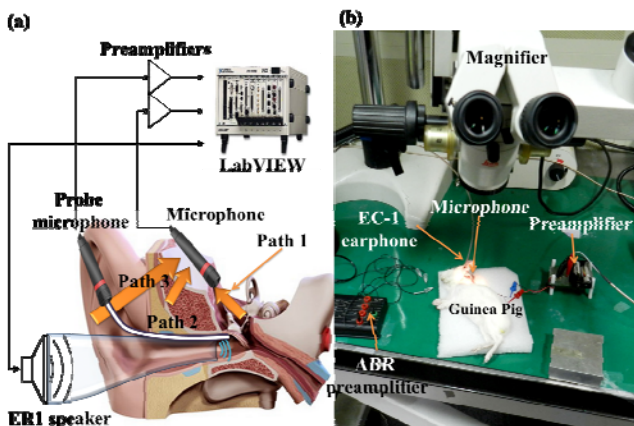


Figure 4. Experiment setup. (a) Block diagram of the experiment. (b) Picture of an animal experiment.

3.1 Experiment results

Figure 5 shows signals measured at the MEC while applying three different types of constant acoustic pressures (70, 80, and 90dB SPL) to the tympanic membrane. Since the reference microphone was placed in front of the tympanic membrane, there was no significant difference between different input sound pressures. Therefore, only data for 80dB SPL was plotted. There were no significant attenuations from low fre-

quencies and small resonance was only seen around 4-5 kHz. In high frequency bands, of around 10-15dB, attenuation was observed. Further, there were no significant differences of placement.

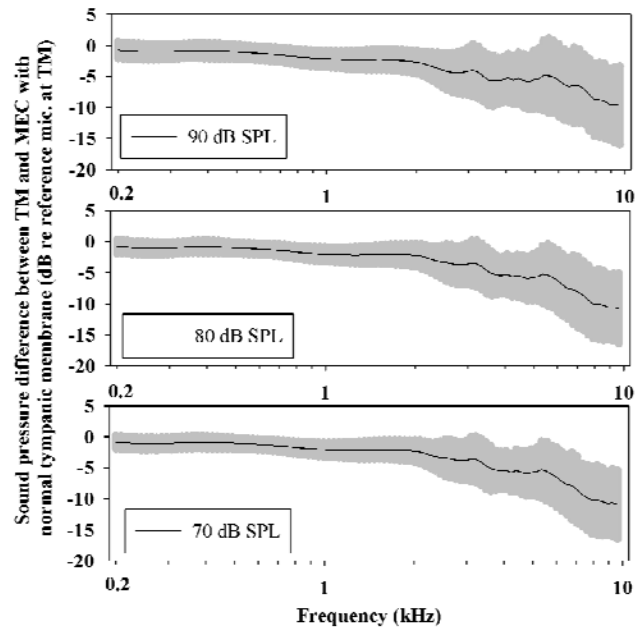


Figure 5. Experiments with the normal tympanic membrane for different input sound levels.

4 Conclusions and Discussion

In this paper, a microphone placed at the MEC was proposed and its characteristics were measured. The proposed method did not cause severe attenuation problems and was less sensitive to motion artifacts. In this paper, simple FEM analysis was performed to determine the proper position for the microphone and showed there is no significant difference.

The greatest problem encountered with attaching the microphone to the tympanic membrane or ossicular chain is feedback, which can easily cause hallowing. Since the proposed method does not attach the microphone to the tympanic membrane, it has the potential to be less sensitive to the feedback problem. The size of the MEC in guinea pigs was too small for the attachment of the available transducers (2x3mm), so this issue will be researched in the near future.

In this experiment, there was no significant sound TL at low frequencies. This could be as a result of the difference between the tympanic membranes in human and guinea pigs, as all of the previous experiments were conducted on the human tympanic membrane. In addition, the experiments outlined here used a reference in front of the tympanic membrane, which is the usual method for hearing aid products.

5 References

- [1] E. P. Hong, et al., "Application of piezoelectric multi-layered actuator to floating mass transducer for implantable middle ear hearing devices," *Journal of Electroceramics*, vol. 23, pp. 335-340, Oct 2009.
- [2] K. W. Seong, et al., "Vibration Analysis of Human Middle Ear with Differential Floating Mass Transducer Using Electrical Model," *IEICE Transactions on Information and Systems*, vol. E92d, pp. 2156-2158, Oct 2009.
- [3] K. W. Seong, et al., "Design of A New Vibration Transducer for Implantable Middle Ear Hearing Devices," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 5, pp. 608-610, Sep 2010.
- [4] I. Y. Park, et al., "Comparisons of electromagnetic and piezoelectric floating-mass transducers in human cadaveric temporal bones," *Hearing Research*, vol. 272, pp. 187-192, Feb 2011.
- [5] M. K. Kim, et al., "Fabrication and optimal design of differential electromagnetic transducer for implantable middle ear hearing device," *Biosensors & Bioelectronics*, vol. 21, pp. 2170-2175, May 15 2006.
- [6] M. K. Kim, et al., "Design of differential electromagnetic transducer for implantable middle ear hearing device using finite element method," *Sensors and Actuators a-Physical*, vol. 130, pp. 234-240, Aug 14 2006.
- [7] S. Tringali, et al., "Fully implantable hearing device with transducer on the round window as a treatment of mixed hearing loss," *Auris Nasus Larynx*, vol. 36, pp. 353-358, Jun 2009.
- [8] E. S. Jung, et al., "Implantable microphone with acoustic tube for fully implantable hearing devices," *IEICE Electronics Express*, vol. 8, pp. 215-219, Feb 25 2011.
- [9] H. A. Jenkins, et al., "Anatomical vibrations that implantable microphones must overcome," *Otology & Neurotology*, vol. 28, pp. 579-588, Aug 2007.
- [10] H. Leysieffer, et al., "Ein implantierbares Mikrofon für elektronische Hörimplantate," *HNO*, vol. 45, pp. 816-827-827, 1997.
- [11] H. Leysieffer, et al., "Ein implantierbares Mikrofon für elektronische Hörimplantate," *HNO*, vol. 45, pp. 816-827-827, 1997.
- [12] W. H. Ko, et al., "Studies of MEMS Acoustic Sensors as Implantable Microphones for Totally Implantable Hearing-Aid Systems," *Ieee Transactions on Biomedical Circuits and Systems*, vol. 3, pp. 277-285, Oct 2009.
- [13] M. Barbara, et al., "The totally implantable middle ear device "Esteem" for rehabilitation of severe sensorineural hearing loss," *Acta Oto-Laryngologica*, vol. 131, pp. 399-404, Apr 2011.
- [14] R. L. Snyder, et al., "Cochlear implant electrode configuration effects on activation threshold and tonotopic selectivity," *Hearing Research*, vol. 235, pp. 23-38, Jan 2008.
- [15] M. L. Whitehead, et al., "The frequency response of the ER-2 speaker at the eardrum," *Journal of the Acoustical Society of America*, vol. 101, pp. 1195-1198, Feb 1997.

Simulated Docking of Zanamivir with the 2009 Pandemic Strain Influenza A/H1N1 Neuraminidase Active Site

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA
email: jhorner@cybermesa.com

Abstract

Influenza neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Zanamivir is a widely used neuraminidase inhibitor. Here I provide a computational docking analysis of zanamivir with the active site of the neuraminidase of the 2009 Influenza A/H1N1 strain. The computed inhibitor/receptor binding energy suggests that zanamivir would be only marginally effective against that strain.

Keywords: Influenza, H1N1, neuraminidase, zanamivir

1.0 Introduction

Influenza neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Zanamivir (5-(acetylamino)-2,6-anhydro-3,4,5-trideoxy-4-[(diaminomethylidene)amino]-D-glycero-D-galacto-non-2-enonic acid (4S,5R,6R)-5-acetamido-4-(diaminomethylideneamino)-6-[(1R,2R)-1,2,3-trihydroxypropyl]-5,6-dihydro-4H-pyran-2-carboxylic acid; [10]) is a widely used influenza therapeutic.

In the World Health Organization serotype-based influenza taxonomy, influenza type A has nine neuraminidase-related sero-subtypes, and these subtypes correspond at least roughly to differences in the active-site structures of the flu neuraminidases. The subtypes fall into two groups ([3]): group-1 contains the subtypes N1, N4, N5 and N8; group-2 contains the subtypes N2, N3, N6, N7 and N9.

Zanamivir was designed to target the group-2 neuraminidases.

The available crystal structures of the group-1 N1, N4 and N8 neuraminidases ([1]) reveal that the active sites of these enzymes have a very different three-dimensional structure from that of group-2 enzymes. The differences lie in a loop of amino acids known as the "150-loop", which in the group-1 neuraminidases has a conformation that opens a cavity not present in the group-2 neuraminidases. The 150-loop contains an amino acid designated Asp 151; the side chain of this amino acid has a carboxylic acid that, in group-1 enzymes, points away from the active site as a result of the 'open' conformation of the 150-loop. The side chain of another active-site amino acid, Glu 119, also has a different conformation in group-1 enzymes compared with the group-2 neuraminidases ([8]).

The Asp 151 and Glu 119 amino acid side chains form critical interactions with neuraminidase inhibitors. For neuraminidase subtypes with the "open

conformation" 150-loop, the side chains of these amino acids might not have the precise alignment required to bind inhibitors tightly ([8]). The active site of the 1918 H1N1 strain has the 150-loop configuration.

The difference in the active-site conformations of the two groups of neuraminidases may also be caused by differences in amino acids that lie outside the active site. This means that an enzyme inhibitor for one target will not necessarily have the same activity against another with the same active-site amino acids and the same overall three-dimensional structure.

Crystallized Influenza A/California/04/2009(H1N1) neuraminidase is an atypical group 1 NA with some group 2-like features in its active site (lack of a 150-cavity) ([4]).

2.0 Method

The general objective of this study is straightforward: to computationally assess the binding energy of the active site of crystallized A/California/04/2009(H1N1) neuraminidase with zanamivir. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment.

Protein Data Bank (PDB) 3TI3 ([6]) is a structural description of most of the crystallized neuraminidase of Influenza A/H1N1 3TI3 consists of two identical chains, designated Chain A and Chain B.

3TI3 was downloaded from PDB on 22 February 2011. A PDB description of zanamivir was extracted from PDB 3B7E ([10]) using *AutoDock Tools* v 4.2 (ADT, [9]). ADT was then used to perform the docking of zanamivir to the receptor. More specifically, in ADT, approximately following the rubric documented in [12]

-- Chain B, and the water in Chain A, of 3TI3 were deleted

-- Chain A's active-site was extracted. (3TI3 identifies the active site of Chain A as 15 amides: ARG118, GLU119, ASP151, ARG152, ARG156, TRP178, ARG224, GLU227, SER246, GLU276, GLU277, ARG292, ASN294, ARG371, and TYR406.)

-- the hydrogens, charges, and torsions in the ligand and active site were adjusted using the ADT-recommended defaults

-- and finally, the ligand, assumed to be flexible wherever that assumption is physically possible, was auto-docked to the active site, assumed to be rigid, using the Lamarckian genetic algorithm implemented in ADT. The best-fit (lowest-energy) configuration from the analysis was saved, and the distances between the receptor and ligand in 3TI3, and those computed here, were compared.

The ADT parameters for the docking are shown in Figure 1. Most values are, or are a consequence of, ADT defaults.

```

autodock_parameter_version 4.2      # used by autodock to validate parameter set
outlev 1                             # diagnostic output level
intelec                             # calculate internal electrostatics
seed pid time                       # seeds for random generator
ligand_types C HD OA N              # atoms types in ligand
fld 3TI3_active.maps.fld            # grid_data_file
map 3TI3_active.C.map               # atom-specific affinity map
map 3TI3_active.HD.map              # atom-specific affinity map
map 3TI3_active.OA.map              # atom-specific affinity map
map 3TI3_active.N.map               # atom-specific affinity map
elecmap 3TI3_active.e.map           # electrostatics map

```



```

desolvmap 3TI3_active.d.map      # desolvation map
move zanamivir.pdbqt           # small molecule
about -29.5772 12.7517 -20.6465 # small molecule center
tran0 random                   # initial coordinates/A or random
axisangle0 random              # initial orientation
dihe0 random                   # initial dihedrals (relative) or random
tstep 2.0                      # translation step/A
qstep 50.0                    # quaternion step/deg
dstep 50.0                    # torsion step/deg
torsdof 9                     # torsional degrees of freedom
rmstol 2.0                    # cluster_tolerance/A
extnrg 1000.0                 # external grid energy
e0max 0.0 10000               # max initial energy; max number of retries
ga_pop_size 150               # number of individuals in population
ga_num_evals 2500000          # maximum number of energy evaluations
ga_num_generations 27000      # maximum number of generations
ga_elitism 1                  # number of top individuals to survive to next
generation                    #
ga_mutation_rate 0.02         # rate of gene mutation
ga_crossover_rate 0.8         # rate of crossover
ga_window_size 10             #
ga_cauchy_alpha 0.0           # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0           # Beta parameter Cauchy distribution
set_ga                        # set the above parameters for GA or LGA
sw_max_its 300                # iterations of Solis & Wets local search
sw_max_succ 4                 # consecutive successes before changing rho
sw_max_fail 4                 # consecutive failures before changing rho
sw_rho 1.0                   # size of local search space to sample
sw_lb_rho 0.01               # lower bound on rho
ls_search_freq 0.06          # probability of performing local search on
individual                    #
set_pswl                      # set the above pseudo-Solis & Wets parameters
unbound_model bound          # state of unbound ligand
ga_run 10                    # do this many hybrid GA-LS runs
analysis                      # perform a ranked cluster analysis

```

Figure 1. ADT parameters for the docking in this study

3.0 Results

The interactive problem setup, which assumes familiarity with the general neuraminidase "landscape", took about 20 minutes in ADT; the docking proper, about 28 minutes on the platform described in Section 2.0. The platform's performance monitor suggested that the calculation was more or less uniformly distributed across the four processors at ~25% of peak per

processor (with occasional bursts to 40% of peak), and required a constant 2.9 GB of memory.

Figure 2 shows the best-fit zanamivir/receptor energy and position summary produced by ADT under the setup shown in Figure 1. The estimated free energy of binding under these conditions is ~ -8.7 kcal/mol; the estimated inhibition constant, ~408 nanoMolar at 298 K.

```

MODEL          1
USER   Run = 1
USER   Cluster Rank = 1
USER   Number of conformations in this cluster = 10
USER
USER   RMSD from reference structure      = 56.144 A
USER
USER   Estimated Free Energy of Binding   = -8.72 kcal/mol  [(1)+(2)+(3)-(4)]
USER   Estimated Inhibition Constant, Ki  = 408.13 nM (nanomolar) [Temperature =
298.15 K]

```

```

USER
USER (1) Final Intermolecular Energy = -11.40 kcal/mol
USER vdW + Hbond + desolv Energy = -8.30 kcal/mol
USER Electrostatic Energy = -3.10 kcal/mol
USER (2) Final Total Internal Energy = -2.75 kcal/mol
USER (3) Torsional Free Energy = +2.68 kcal/mol
USER (4) Unbound System's Energy [= (2)] = -2.75 kcal/mol
USER
USER
USER
USER DPF = 3TI3_zanamivir.dpf
USER NEWDPF move zanamivir.pdbqt
USER NEWDPF about -29.577200 12.751700 -20.646500
USER NEWDPF tran0 29.961176 14.781299 -20.419074
USER NEWDPF axisangle0 -0.004045 -0.391949 0.919978 3.081993
USER NEWDPF quaternion0 -0.000109 -0.010540 0.024740 0.999638
USER NEWDPF dihe0 4.89 175.54 139.90 180.00 67.18 1.07 -179.74 0.58 -36.96
USER
USER
USER x y z vdW Elec q RMS
ATOM 1 C2 ZMR A1001 29.610 13.398 -22.778 -0.14 +0.09 +0.144 56.144
ATOM 2 C3 ZMR A1001 30.901 13.720 -22.564 -0.34 +0.01 +0.045 56.144
ATOM 3 C4 ZMR A1001 31.277 14.664 -21.442 -0.27 -0.00 +0.150 56.144
ATOM 4 C5 ZMR A1001 30.226 14.586 -20.317 -0.17 +0.04 +0.143 56.144
ATOM 5 C6 ZMR A1001 28.817 14.747 -20.891 -0.14 +0.08 +0.185 56.144
ATOM 6 O6 ZMR A1001 28.541 13.810 -21.924 -0.14 -0.22 -0.335 56.144
ATOM 7 NE ZMR A1001 32.576 14.369 -20.810 -0.22 +0.04 -0.217 56.144
ATOM 8 HE ZMR A1001 32.843 13.389 -20.711 -0.26 -0.16 +0.178 56.144
ATOM 9 CZ ZMR A1001 33.401 15.265 -20.371 +0.01 +0.06 +0.665 56.144
ATOM 10 NH1 ZMR A1001 33.240 16.579 -20.493 -0.24 +0.05 -0.235 56.144
ATOM 11 NH2 ZMR A1001 34.493 14.843 -19.724 -0.31 -0.14 -0.235 56.144
ATOM 12 2HH1 ZMR A1001 32.407 16.900 -20.987 +0.08 -0.07 +0.174 56.144
ATOM 13 1HH1 ZMR A1001 33.890 17.285 -20.148 -0.38 -0.08 +0.174 56.144
ATOM 14 2HH2 ZMR A1001 34.617 13.835 -19.630 -0.39 +0.16 +0.174 56.144
ATOM 15 1HH2 ZMR A1001 35.144 15.549 -19.378 -0.44 +0.11 +0.174 56.144
ATOM 16 N5 ZMR A1001 30.437 15.627 -19.309 -0.02 -0.20 -0.352 56.144
ATOM 17 H5 ZMR A1001 30.130 16.576 -19.525 +0.10 +0.07 +0.163 56.144
ATOM 18 C10 ZMR A1001 31.013 15.406 -18.112 -0.24 +0.22 +0.214 56.144
ATOM 19 C11 ZMR A1001 31.268 16.657 -17.329 -0.34 +0.13 +0.117 56.144
ATOM 20 O10 ZMR A1001 31.344 14.278 -17.729 -0.74 -0.41 -0.274 56.144
ATOM 21 C1 ZMR A1001 29.129 12.658 -23.951 -0.19 +0.35 +0.233 56.144
ATOM 22 O1A ZMR A1001 30.010 12.129 -24.683 -1.05 -1.46 -0.642 56.144
ATOM 23 O1B ZMR A1001 27.908 12.571 -24.177 -1.03 -1.48 -0.642 56.144
ATOM 24 C7 ZMR A1001 27.690 14.594 -19.863 -0.09 +0.13 +0.180 56.144
ATOM 25 C8 ZMR A1001 26.561 15.617 -20.084 -0.25 +0.09 +0.173 56.144
ATOM 26 O8 ZMR A1001 25.343 14.887 -20.303 -0.20 -0.19 -0.391 56.144
ATOM 27 H8 ZMR A1001 24.662 15.515 -20.514 -0.40 -0.11 +0.210 56.144
ATOM 28 C9 ZMR A1001 26.902 16.556 -21.266 -0.21 +0.02 +0.198 56.144
ATOM 29 O9 ZMR A1001 25.780 16.637 -22.140 -0.01 -0.06 -0.398 56.144
ATOM 30 H9 ZMR A1001 25.104 16.044 -21.835 -0.35 -0.03 +0.209 56.144
ATOM 31 O7 ZMR A1001 27.148 13.287 -19.968 +0.01 -0.32 -0.390 56.144
ATOM 32 H7 ZMR A1001 27.094 13.052 -20.887 +0.08 +0.19 +0.210 56.144
TER
ENDMDL

```

Figure 2. ADT's zanamivir energy and position predictions.

Figure 3 is a rendering of the active-site/inhibitor configuration computed in this study.

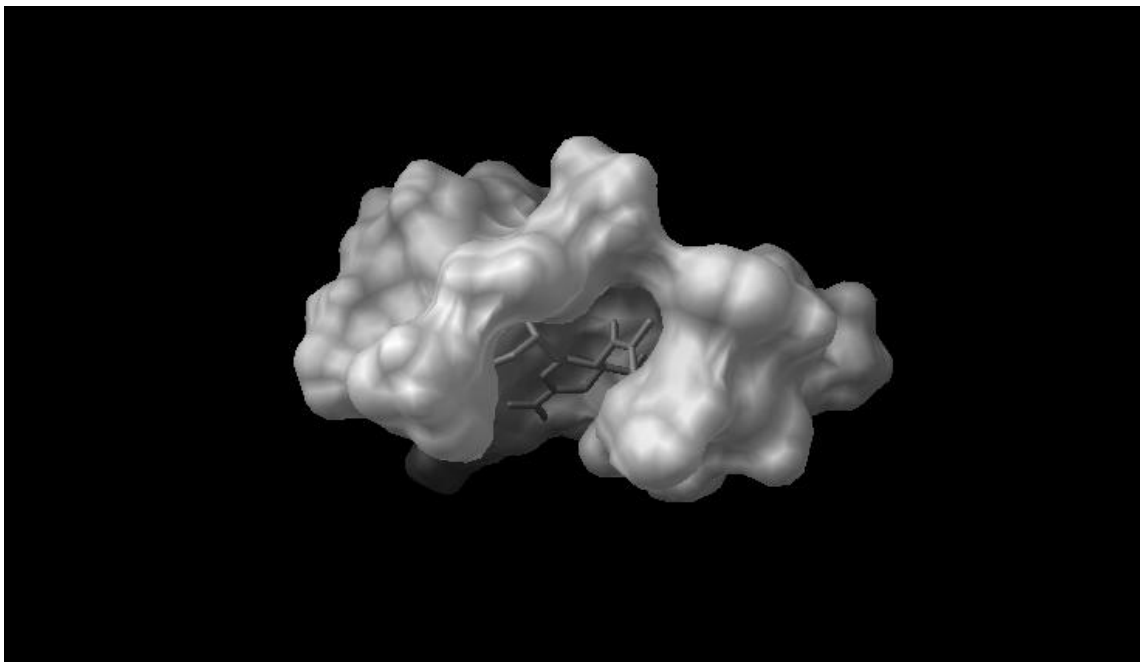


Figure 3. Rendering of zanamivir computationally docked with the active site of PDB 3TI3. The molecular surface of the receptor is shown in white; the inhibitor, in stick form in grey. Only the interior, inhibitor-containing region of the molecular surface of the active site can be compared to *in situ* data: the surface distal to the interior is a computational artifact, generated by the assumption that active site is detached from the rest of the receptor.

The distances between ligand and receptor atoms in 3TI3, and the corresponding distances in the present computation were within 10% of each other.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The inhibition constant computed in this study (~408 nanoMolar at ~298 K) is comparable inhibition constant of neuraminidase inhibitors that are not clinically effective ([10], [11], [13], [14], [15]) against several H1N1 genotypes. This suggests that zanamivir would be only marginally effective against Influenza A/California/04/2009(H1N1)). It would,

however, be more effective than oseltamivir (Tamiflu®) against that strain.

2. The docking study reported here assumes that the receptor is rigid. This assumption is appropriate for the binding energy computation for PDB 3TI3 per se. However, the calculation does not reflect what receptor "flexing" could contribute to the interaction of the ligand with native unliganded receptor.

3. The analysis described in Sections 2.0 and 3.0 assumes receptor is in a crystallized form. *In situ*, at physiologically normal temperatures (~310 K), the receptor is not in crystallized form. The ligand/receptor conformation *in situ*, therefore, may not be identical to their conformation in the crystallized form.

4. Minimum-energy search algorithms other than the Lamarckian

genetic algorithm used in this work could be applied to this docking problem. Future work will use Monte Carlo/simulated annealing algorithms.

5. A variety of torsion and charge models could be applied to this problem, and future work will do so.

6. 3TI3 has two chains, each with its own active site. The work described in this paper was performed on Chain A only. Chain B appears to have an active site highly similar to the Chain A active site. Future work will assess the ligand/receptor binding energies of Chains B.

5.0 Acknowledgements

This work benefited from discussions with Tony Pawlicki. For any problems that remain, I am solely responsible.

6.0 References.

- [1] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49.
- [2] Johnson NP and Mueller J. Updating the accounts: global mortality of the 1918-1920 "Spanish " influenza pandemic. *Bulletin of the History of Medicine* 76 (2002), 105-115.
- [3] World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization* 58 (1980), 585-591.
- [4] Vavricka CF, Li Q, Wu Y, Qi J, Wang M, Liu Y, Gao F, Liu J, Feng E, He J, Wang J, Liu H, Jiang H, and Gao GF. Structural and functional analysis of laninamivir and its octanoate prodrug reveals group specific mechanisms for Influenza NA inhibition. *PLoS Pathogens* 7 (October 2011): e1002249. doi:10.1371/journal.ppat.1002249.
- [5] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.
- [6] PDB ID = [10.2210/pdb3ti3/pdb](http://www.rcsb.org/pdb/entry.do?entry=10.2210/pdb3ti3/pdb). See also [4].
- [7] US Centers for Disease Control. *Summary: Interim Recommendations for the Use of Influenza Antiviral Medications in the Setting of Oseltamivir Resistance among Circulating Influenza A (H1N1) Viruses, 2008-09 Influenza Season*. 19 December 2008. URL <http://www.cdc.gov/flu/professionals/antivirals/summary.htm>.
- [8] Luo M. Structural biology: antiviral drugs fit for a purpose. *Nature* 443 (7 September 2006), 37-38. doi:10.1038/443037a, published online 6 September 2006.
- [9] Morris GM, Goodsell DS, Huey R, Lindstrom W, Hart WE, Kurowski S, Halliday S, Belew R, and Olson AJ. *AutoDock* v4.2. <http://autodock.scripps.edu/>. 2010.
- [10] PDB ID = [10.2210/pdb3b7e/pdb](http://www.rcsb.org/pdb/entry.do?entry=10.2210/pdb3b7e/pdb). Xu X, Zhu X, Dwek RA, Stevens J, Wilson IA. Structural characterization of the 1918 influenza virus H1N1 neuraminidase. *Journal of Virology* 82 (2008), 10493-10501.
- [11] Govorkova EA et al. Comparison of efficacies of RWJ-270201, oseltamivir, and zanamivir against H5N1, H9N2, and other avian influenza viruses. *Antimicrobial Agents and Chemotherapy* 45 (2001), 2723-2732.
- [12] Huey R and Morris GM. *Using AutoDock 4 with AutoDock Tools: A*

Tutorial. 8 January 2008.
<http://autodock.scripps.edu/>.

[13] Cheng Y and Prusoff WH. Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. *Biochemical Pharmacology* 22 (December 1973), 3099–3108. doi:10.1016/0006-2952(73)90196-2.

[14] Horner JK. Simulated docking of oseltamivir with the 1918 pandemic strain

Influenza A/H1N1 zanamivir-conformed neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. 2011. pp. 130-135.

[15] Horner JK. Simulated docking of zanamivir with the 1918 pandemic strain Influenza A/H1N1 neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. pp. 136-142.

Developing the Information Architecture for the Outsourcing Physical Examination

P.Y Hung¹, P.Y Lee¹, C.H Hsiao², A.J Lee³, and S.T Tang¹

¹ Department of Biomedical Engineering, Ming Chuan University, Taoyuan, Taiwan

² Department of Medical Informatics, Tzu Chi University, Hualien, Taiwan

³ Department of Healthcrae Information and Management, Ming Chuan University, Taoyuan, Taiwan

Abstract - In the developed country, it is necessary that the enterprise should arrange the occupational physical examination yearly, there are millions cases, results in heavy loading to the hospital. Because of the limited scale, the hospital usually needs outsourcing physical examination institutes. But the introduction of the outsourcing institute results in problems of examination information sharing. Although the information systems for hospital have been developed for decades, and now there are various successful systems. But the information system for outsourcing institute integration is not yet developed. This study is to develop the information architecture for outsourcing institute integration, which is basing on the relational medical information standards, includes DICOM, HL7 CDA, and IHE XDS. The proposed architecture would provide the information sharing in heterogeneous systems for different hospitals.

Keywords: IHE XDS; HL7 CDA; DICOM

1 Introduction

In the developed countries, it is usually necessary that the enterprise should arrange the physical examination for the employees yearly. As a result, there are million employees should undergo examination every year, which makes the heavy loading to the hospitals and lots of traffic time to the employees. Additionally, basing on the cost considerations, the enterprise always requests the hospital to provide on-site examination service. The general hospital is originally designed for treatment, not for physical examination. But physical examination should be done by the hospital is necessary by law. Then the hospital needs the outsourcing institute for on-site physical examination.

Along with the development of the healthcare process that has involved complicated information flow [1]. Nowadays there are various information systems for the operation of the hospital, and due to the heterogeneity, the hospital information systems can not share data between different hospitals. As a result, when a hospital requests the outsourcing institute, which would be problems in examination information sharing. The outsourcing institute is impossible to design different system for different contract hospital. Although the information systems for hospital have developed for decades,

and now there are various successful systems, e.g. HIS, RIS, LIS. But the information system for outsourcing physical examination is not yet developed. As a result, the outsourcing institute usually adopts the modified commercial MIS (Management information system) for institute management, and the associated paper-forms for data sharing with contract hospital. The general paper-form [2] workflow is shown in Fig. 1. In the workflow of the on-site physical examination, before the examination the employee should get a blank form firstly, and then fill his/her personal information, and then the healthcare provider note the results of physical examination on the form. Then the paper form would deliver to clerks for data key-in for the information system of contract hospital, additional double checking process is necessary. Finally the examination report is generated, and then sends to the employee.

Paper-form based workflow is a error-prone process, and consuming lots of time and manpower. With more and more outsourcing institutes being introduced into hospital, a key issue is how to design the necessary information architecture, which can reduce data errors, time consuming and even the data loss risk. This study is to develop the information architecture for outsourcing institute, which transfers the current process from paper-form to electronic-form, would well reduce lots of processing time and manual resources. Additionally, the developed architecture is basing on the relational medical information standards, which provides the information sharing in heterogeneous systems for different contract hospitals.

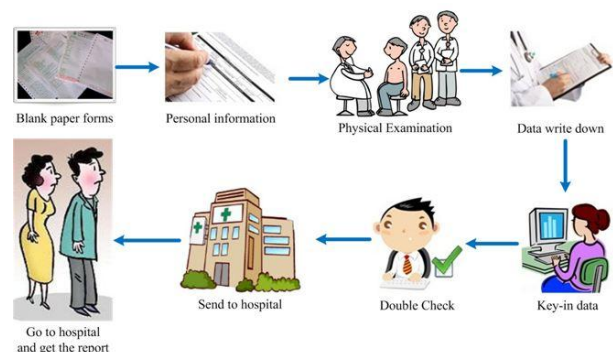


Figure 1. Traditional paper-form based on-site physical examination flow.

2 Methods

For the outsourcing physical examination, the development of the information architecture is to re-design the workflow firstly. Then the key information modules are developed for replacing the current manual operations, which includes electronic form, ultrasonography encapsulation. Additionally, the architecture development is basing on the medical information standards for data exchange with the current healthcare information systems.

2.1 Redesign workflow

The workflow is redesigned, which referred the current examination procedure of Taipei Veterans General Hospital, and the current workflow of outsourcing institute.

2.2 Electronic form

We use Android SDK to develop the electron-form system. Android is a Linux-based operating system developed by Google. In addition to the operating system, it also provides Android SDK/NDK application software development kit that allows embedded systems developers to develop Android platform applications. There are lots of facilities involved in the study, include Eclipse, Java Development Kit (JDK), Android Development Tools (ADT), Android SDK, and ASUS TF101.

2.3 Ultrasonography acquisition

The most common medical image modalities are ultrasonography for its non-radical. But the most general ultrasonography is non-dicom compatible for its cost. The proposed ultrasonography acquisition module includes three parts: the driver for image acquisition device, DICOMDIR generator, and the shared dynamic link libraries (DLLs), which as shown in Fig. 2. For inherent low resolution of ultrasonography and the ease of future maintenance, the off-the-shelf image acquisition device was adopted. As a result, the driver should be developed for controlling the acquisition device. The acquired image is JPEG format, and then a DICOMDIR generator is required for image converting. Additionally, the acquired image could be also import to other DICOM applications in local clinics for further study, which is necessary to develop shared dynamic link libraries for cooperating with other software applications.

National Instruments (NI) LabVIEW is a graphical based programming language, which support rich libraries to facilitate the development of device driver or instrument control console [3]. NI LabVIEW Plug and Play Instrument Drivers is deployed to develop the driver for the off-the-shelf image acquisition device. The LabVIEW Instrument Driver Finder (IDFinder) is firstly applied to find, download, and start using the similar instrument driver. The DICOMDIR is a directory object, which is to serve as an index for organizing and finding DICOM files inside a physical storage media [4].

The DICOMDIR object formal definition and its structure are in part 3-annex F and part 10-section 7 of the DICOM standard document. The DICOMDIR file contains hierarchically sorted registers with the information related to objects stored into a DICOM files set. [5] In most DICOM storage media, a set of DICOM information is described by an index file, DICOMDIR, which accompanies the files that it references. The main function of the DICOM encoder is to convert the NTSC video signal into a DICOM-compatible digital file. Besides NI LabVIEW, the Intel JPEG Library and DicomObjects are also included for developing the DICOM encoder. The NI LabVIEW Application Builder to pack the developed application and the shared dynamic link libraries.

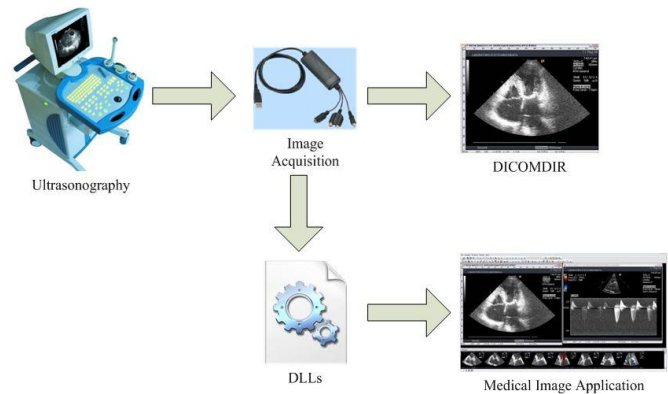


Figure 2. Ultrasonography acquisition module.

2.4 Report generation

Currently, the text and image reports of physical examination are separately. We use NI LabVIEW Report Generation Toolkit for Microsoft Office in report generation, which integrates the results of text and image.

2.5 Medical information standards

We referred the IHE XDS standards [6] and the requirements for healthcare institute. The system framework is based on HL7 [7] and DICOM [8] standards. The report system is basing on the HL7 [9] CDA [10], which shares the medical text files in different platforms.

The XML (eXtensible Markup Language) is the most popular format, which supports global data exchange. Additionally, XML could be embedded in Web service. As a result, the XML document is easy shared via internet. Health level 7 (HL7) clinical document architecture (CDA) provides a standard form for digitizing a series of medical documents, and cross-discipline data exchange [11, 12]. The CDA is a XML based format, which is constituted by medical objects, including text, image, and voice. As a result, the CDA document could be accessed via Web browser. In this study, we demonstrate the electronic forms of the physical

examination basing on HL7 CDA standard. The examined data and image are described by CDA level 3.

3 Results

The developed electron-form based architecture is shown in Fig. 3. The developed ultrasonography acquisition application is shown in Fig. 4, which had been confirmed in hospital. The GUI of ultrasonography DICOM transformer is shown in Fig. 5. The GUI of report generation is shown in Fig. 6. The remolded information architecture is shown in Fig. 7. During the on-site physical examination, healthcare staff confirms the examination items with employee firstly. Then collect specimens and medical images (X-ray or ultrasonography). After investigation, transmit all data to the database. Then output report from database and deliver via Internet.

4 Conclusions

In order to exchange data between outsourcing laboratory and hospital. We provide the information architecture for outsourcing medical laboratory, which referred the standards of IHE and HL7. The research aims is not only to improve the information exchange, but also effectively reduce incident errors. We will develop personal health record system in the future, so the subjects can search personal healthy situation in real-time at home and provide long-term caring services.

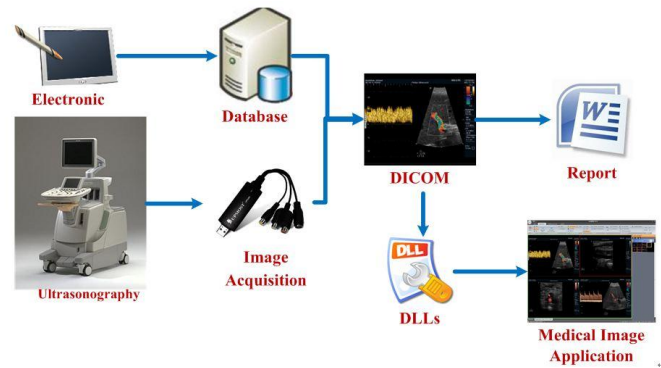


Figure 4. The developed ultrasonography acquisition application.



Figure 5. The GUI of ultrasonography DICOM transformer.



Figure 3. The GUI of electronic form.

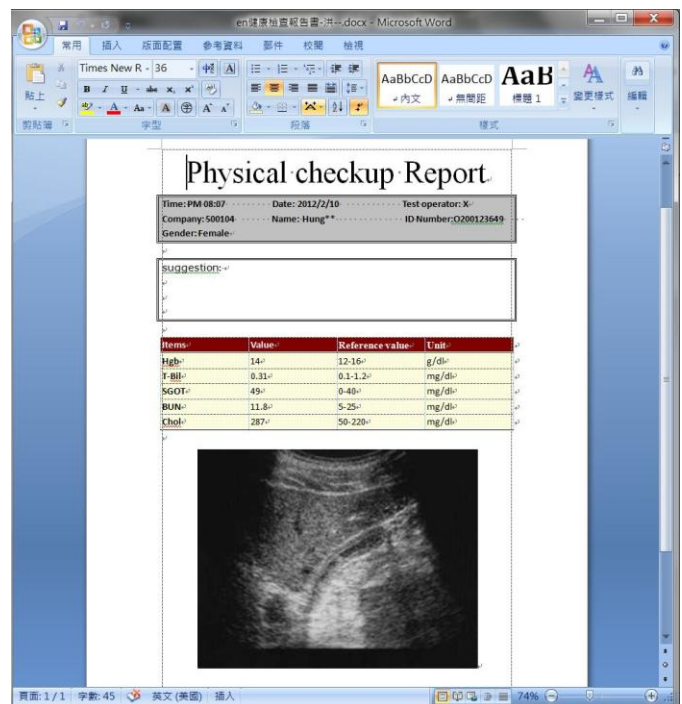


Figure 6. The GUI of report generation.

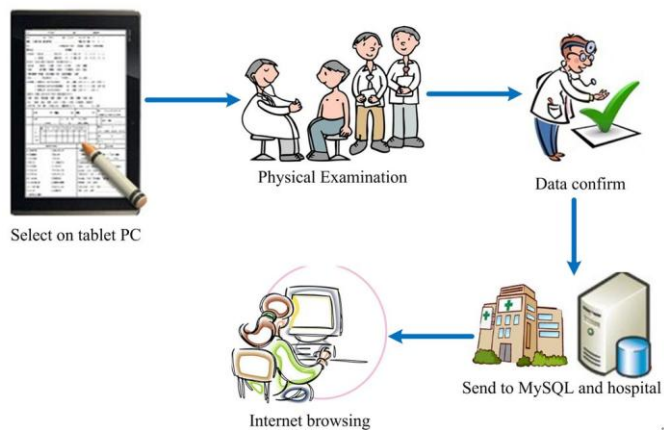


Figure 7. Developed information architecture.

5 References

- [1] S.T.C. Wong, and H.K. Huang, “A Hospital Integrated Framework for Multimodality Image Base Management”, *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 26, pp. 455—469, Jul. 1996.
- [2] H.H. Rau, C.Y. Hsu, Y.L. Lee, W. Chen, and W.S. Jian, “Developing Electronic Health Records in Taiwan”, *IT Professional*, vol.12, pp. 17—25, March-April 2010.
- [3] NI LabVIEW Software Engineering Hands-On Exercises : <http://www.ni.com>
- [4] R. Villegas, G. Montilla, H. Villegas, “DICOMDIR files reader for using in computer assisted diagnosis and surgery”, : <http://image2006.cmm.uchile.cl/papers/villegas2.doc>
- [5] CS. Yam, A. Sitek, V. Raptopoulos and M. Larson, “A simple method for extracting DICOM images from a magneto optic disk”, *AJR Am J Roentgenol*, pp 183:529—33, 2004.
- [6] R. Noumeir, “Sharing Medical Records: The XDS Architecture and Communication Infrastructure”, *IT Professional*, vol. 13, pp. 46—52, July-Aug. 2011.
- [7] P. De Meo, G. Quattrone, and D. Ursino, “Integration of the HL7 Standard in a Multiagent System to Support Personalized Access to e-Health Services”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 1244—1260, Aug. 2011.
- [8] R. Noumeir, “DICOM Structured Report Document Type Definition”, *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, pp. 318—328, Dec. 2003.
- [9] IHE: <http://www.ihe.net/>
- [10] Details about the structure of IHE Technical Frameworks and Supplements: <http://www.ihe.net/About/process.cfm> , <http://www.ihe.net/profiles/index.cfm>
- [11] Marcel Lucas Müller, Frank Ü ckert, Thomas Bürkle and Hans-Ulrich Prokosch, “Cross-institutional data exchange using the clinical document architecture (CDA)”, *International Journal of Medical Informatics*, vol.74, pp. 245—256, 2005.
- [12] F. Piero, R. Gustavo, Sá, and N.F. Fernando, “Rich Internet Application”, *IEEE Internet Computing*, vol. 14, pp. 9—12, Jul. 2010.

Simulated Docking of Laninamivir with the 1918 Pandemic Strain Influenza A/H1N1 Neuraminidase Active Site

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA
email: jhorner@cybermesa.com

Abstract

Neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Laninamivir is a neuraminidase inhibiting drug approved for general use in Japan in 2010 for the treatment of influenza, and for emergency use in the US in 2011. Here I provide a computational docking analysis of laninamivir with the active site of the neuraminidase of the 1918 strain (A/Brevig Mission/1/18 H1N1). The computed inhibitor/receptor binding energy suggests that laninamivir would not be effective against that strain.

Keywords: Influenza, H1N1, neuraminidase, laninamivir

1.0 Introduction

Neuraminidases are glycoproteins that facilitate the transmission of the influenza virus from cell to cell. Laninamivir (4S,5R,6R)-5-acetamido-4-carbamimidamido-6-[(1R,2R)-3-hydroxy-2-methoxypropyl]-5,6-dihydro-4H-pyran-2-carboxylic acid; [14]) is a neuraminidase inhibitor approved in Japan in 2010 for general use in the treatment of influenza and for emergency use in the US in 2011.

In the World Health Organization serotype-based influenza taxonomy, influenza type A has nine neuraminidase-related sero-subtypes, and these subtypes correspond at least roughly to differences in the active-site structures of the flu neuraminidases. The subtypes fall into two groups ([3]): group-1 contains the subtypes N1, N4, N5 and N8; group-2 contains the subtypes N2, N3, N6, N7 and N9. Laninamivir was designed to target the group-2 neuraminidases.

The available crystal structures of the group-1 N1, N4 and N8 neuraminidases ([1]) reveal that the active sites of these enzymes have a very different three-dimensional structure from that of group-2 enzymes. The differences lie in a loop of amino acids known as the "150-loop", which in the group-1 neuraminidases has a conformation that opens a cavity not present in the group-2 neuraminidases. The 150-loop contains an amino acid designated Asp 151; the side chain of this amino acid has a carboxylic acid that, in group-1 enzymes, points away from the active site as a result of the 'open' conformation of the 150-loop. The side chain of another active-site amino acid, Glu 119, also has a different conformation in group-1 enzymes compared with the group-2 neuraminidases (8)).

The Asp 151 and Glu 119 amino acid side chains form critical interactions with neuraminidase inhibitors. For neuraminidase subtypes with the "open conformation" 150-loop, the side chains

of these amino acids might not have the precise alignment required to bind inhibitors tightly ([8]). The active site of the 1918 strain has the 150-loop configuration.

The difference in the active-site conformations of the two groups of neuraminidases may also be caused by differences in amino acids that lie outside the active site. This means that an enzyme inhibitor for one target will not necessarily have the same activity against another with the same active-site amino acids and the same overall three-dimensional structure ([17]).

2.0 Method

The general objective of this study is straightforward: to computationally assess the binding energy of the active site of crystallized 1918 pandemic strain neuraminidase with laninamivir. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment.

Protein Data Bank (PDB) 3BEQ ([6]) is a structural description of most of the crystallized neuraminidase of Influenza A/Brevig Mission/1/18 H1N1 (the principal 1918 pandemic mutant). 3BEQ consists of

two identical chains, designated Chain A and Chain B.

3BEQ was downloaded from PDB on 31 January 2011. A PDB description of laninamivir was extracted from PDB 3TI8 ([4]) using *AutoDock Tools* v 4.2 (ADT, [9]). ADT was then used to perform the docking of laninamivir to the receptor. More specifically, in ADT, approximately following the rubric documented in [12]

-- Chain B, and the water in Chain A, of 3BEQ were deleted

-- Chain A's active-site was extracted. (3BEQ identifies the active site of Chain A as 14 amides: ARG118, GLU119, ASP151, ARG152, ARG156, TRP178, ARG224, GLU227, SER246, GLU276, GLU277, ARG292, ARG371, and TYR406.)

-- the hydrogens, charges, and torsions in the ligand and active site were adjusted using the ADT-recommended defaults

-- and finally, the ligand, assumed to be flexible wherever that assumption is physically possible, was auto-docked to the active site, assumed to be rigid, using the Lamarckian genetic algorithm implemented in ADT. The best-fit (lowest-energy) configuration from the analysis was saved.

The ADT parameters for the docking are shown in Figure 1. Most values are, or are a consequence of, ADT defaults.

```

autodock_parameter_version 4.2      # used by autodock to validate parameter set
outlev 1                            # diagnostic output level
intelec                             # calculate internal electrostatics
seed pid time                       # seeds for random generator
ligand_types C HD OA N              # atoms types in ligand
fld 3BEQ_receptor.maps.fld          # grid data file
map 3BEQ_receptor.C.map             # atom-specific affinity map
map 3BEQ_receptor.HD.map            # atom-specific affinity map
map 3BEQ_receptor.OA.map            # atom-specific affinity map
map 3BEQ_receptor.N.map             # atom-specific affinity map
elecmap 3BEQ_receptor.e.map         # electrostatics map
desolvmap 3BEQ_receptor.d.map       # desolvation map
move laninamivirA.pdbqt             # small molecule

```

```

about 22.7762 -20.7805 -52.3029 # small molecule center
tran0 random # initial coordinates/A or random
axisangle0 random # initial orientation
dihe0 random # initial dihedrals (relative) or random
tstep 2.0 # translation step/A
qstep 50.0 # quaternion step/deg
dstep 50.0 # torsion step/deg
torsdof 9 # torsional degrees of freedom
rmstol 2.0 # cluster_tolerance/A
extnrg 1000.0 # external grid energy
e0max 0.0 10000 # max initial energy; max number of retries
ga_pop_size 150 # number of individuals in population
ga_num_evals 2500000 # maximum number of energy evaluations
ga_num_generations 27000 # maximum number of generations
ga_elitism 1 # number of top individuals to survive to next
generation
ga_mutation_rate 0.02 # rate of gene mutation
ga_crossover_rate 0.8 # rate of crossover
ga_window_size 10 #
ga_cauchy_alpha 0.0 # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0 # Beta parameter Cauchy distribution
set_ga # set the above parameters for GA or LGA
sw_max_its 300 # iterations of Solis & Wets local search
sw_max_succ 4 # consecutive successes before changing rho
sw_max_fail 4 # consecutive failures before changing rho
sw_rho 1.0 # size of local search space to sample
sw_lb_rho 0.01 # lower bound on rho
ls_search_freq 0.06 # probability of performing local search on
individual
set_pswl # set the above pseudo-Solis & Wets parameters
unbound_model bound # state of unbound ligand
ga_run 10 # do this many hybrid GA-LS runs
analysis # perform a ranked cluster analysis

```

Figure 1. ADT parameters for the docking in this study

3.0 Results

The interactive problem setup, which assumes familiarity with the general neuraminidase "landscape", took about 20 minutes in ADT; the docking proper, about 28 minutes on the platform described in Section 2.0. The platform's performance monitor suggested that the calculation was more or less uniformly distributed across the four processors at ~25% of peak per

processor (with occasional bursts to 40% of peak), and required a constant 2.9 GB of memory.

Figure 2 shows the best-fit laninamivir/receptor energy and position summary produced by ADT under the setup shown in Figure 1. The estimated free energy of binding under these conditions is ~ -7 kcal/mol; the estimated inhibition constant, ~7.7 microMolar at 298 K.

```

MODEL          6
USER           Run = 6
USER           Cluster Rank = 1
USER           Number of conformations in this cluster = 1
USER
USER           RMSD from reference structure          = 61.496 A
USER
USER           Estimated Free Energy of Binding      = -6.97 kcal/mol  [(1)+(2)+(3)-(4)]

```

```

USER      Estimated Inhibition Constant, Ki   =      7.72 uM (micromolar) [Temperature =
298.15 K]
USER
USER      (1) Final Intermolecular Energy     =      -9.66 kcal/mol
USER      vdW + Hbond + desolv Energy        =      -7.32 kcal/mol
USER      Electrostatic Energy              =      -2.34 kcal/mol
USER      (2) Final Total Internal Energy     =      -1.56 kcal/mol
USER      (3) Torsional Free Energy          =      +2.68 kcal/mol
USER      (4) Unbound System's Energy   [= (2)] =      -1.56 kcal/mol
USER
USER
USER      DPF = laninamivirA_3BEQ.dpf
USER      NEWDPF move      laninamivirA.pdbqt
USER      NEWDPF about    22.776199 -20.780500 -52.302898
USER      NEWDPF tran0    7.783660 14.900163 -0.696761
USER      NEWDPF axisangle0  0.051041 0.607910 0.792364 70.549071
USER      NEWDPF quaternion0  0.029476 0.351065 0.457586 0.816394
USER      NEWDPF dihe0    121.24 76.63 152.14 144.19 61.46 -14.16 72.74 160.39 -28.29
USER
USER      x      y      z      vdW      Elec      q      RMS
ATOM      1  CAA LNV A 901      6.776  14.665 -3.341 -0.26 +0.21 +0.235 61.496
ATOM      2  CAB LNV A 901      7.292  16.072 -3.047 -0.24 -0.00 +0.103 61.496
ATOM      3  CAC LNV A 901      7.903  16.305 -1.661 -0.32 -0.01 +0.059 61.496
ATOM      4  CAD LNV A 901      7.836  15.114 -0.662 -0.16 +0.03 +0.090 61.496
ATOM      5  CAE LNV A 901      6.930  13.905 -1.032 -0.13 +0.07 +0.107 61.496
ATOM      6  OAF LNV A 901      6.325  13.844 -2.309 -0.17 -0.36 -0.334 61.496
ATOM      7  NAZ LNV A 901      9.014  17.205 -1.573 -0.15 +0.08 -0.194 61.496
ATOM      8  HAZ LNV A 901      8.940  17.997 -0.936 -0.33 -0.16 +0.184 61.496
ATOM      9  CBA LNV A 901     10.223  17.056 -2.327 +0.12 +0.00 +0.669 61.496
ATOM     10  NBC LNV A 901     10.559  15.885 -2.869 -0.17 -0.15 -0.235 61.496
ATOM     11  NBB LNV A 901     11.054  18.076 -2.330 -0.13 +0.02 -0.235 61.496
ATOM     12  1HBC LNV A 901     11.420  15.780 -3.406 -0.24 +0.15 +0.174 61.496
ATOM     13  2HBC LNV A 901      9.914  15.094 -2.867 +0.10 +0.14 +0.174 61.496
ATOM     14  2HBB LNV A 901     10.797  18.972 -1.915 -0.46 -0.14 +0.174 61.496
ATOM     15  1HBB LNV A 901     11.916  17.970 -2.867 -0.25 +0.08 +0.174 61.496
ATOM     16  NBG LNV A 901      8.170  15.300  0.721 -0.04 -0.03 -0.324 61.496
ATOM     17  HBG LNV A 901      7.439  15.653  1.340 -0.37 -0.11 +0.169 61.496
ATOM     18  CBD LNV A 901      9.445  15.029  1.297 -0.06 +0.12 +0.218 61.496
ATOM     19  OBF LNV A 901     10.320  14.445  0.665 -0.03 -0.22 -0.274 61.496
ATOM     20  CBE LNV A 901      9.637  15.445  2.722 -0.19 +0.06 +0.117 61.496
ATOM     21  CAG LNV A 901      6.172  14.373 -4.674 -0.21 +0.34 +0.204 61.496
ATOM     22  OAH LNV A 901      6.227  13.087 -5.141 -0.23 -1.38 -0.646 61.496
ATOM     23  OAI LNV A 901      6.010  15.307 -5.441 -0.40 -0.90 -0.646 61.496
ATOM     24  CAJ LNV A 901      7.195  12.570 -0.411 -0.08 +0.15 +0.210 61.496
ATOM     25  OAW LNV A 901      7.776  11.761 -1.357 +0.02 -0.05 -0.381 61.496
ATOM     26  CAX LNV A 901      7.113  10.609 -1.750 +0.01 +0.07 +0.202 61.496
ATOM     27  CAK LNV A 901      5.940  12.009  0.227 -0.15 +0.21 +0.177 61.496
ATOM     28  OAY LNV A 901      4.781  12.513 -0.337 -0.56 -0.66 -0.390 61.496
ATOM     29  HAY LNV A 901      4.366  13.173  0.207 -0.46 +0.32 +0.210 61.496
ATOM     30  CAL LNV A 901      5.970  12.083  1.737 -0.16 +0.27 +0.198 61.496
ATOM     31  OAM LNV A 901      4.910  12.785  2.340 -1.16 -0.93 -0.398 61.496
ATOM     32  HAM LNV A 901      4.515  13.471  1.815 -0.45 +0.44 +0.209 61.496
TER
ENDMDL

```

Figure 2. ADT's laninamivir energy and position predictions.

Figure 3 is a rendering of the active-site/inhibitor configuration computed in this study.

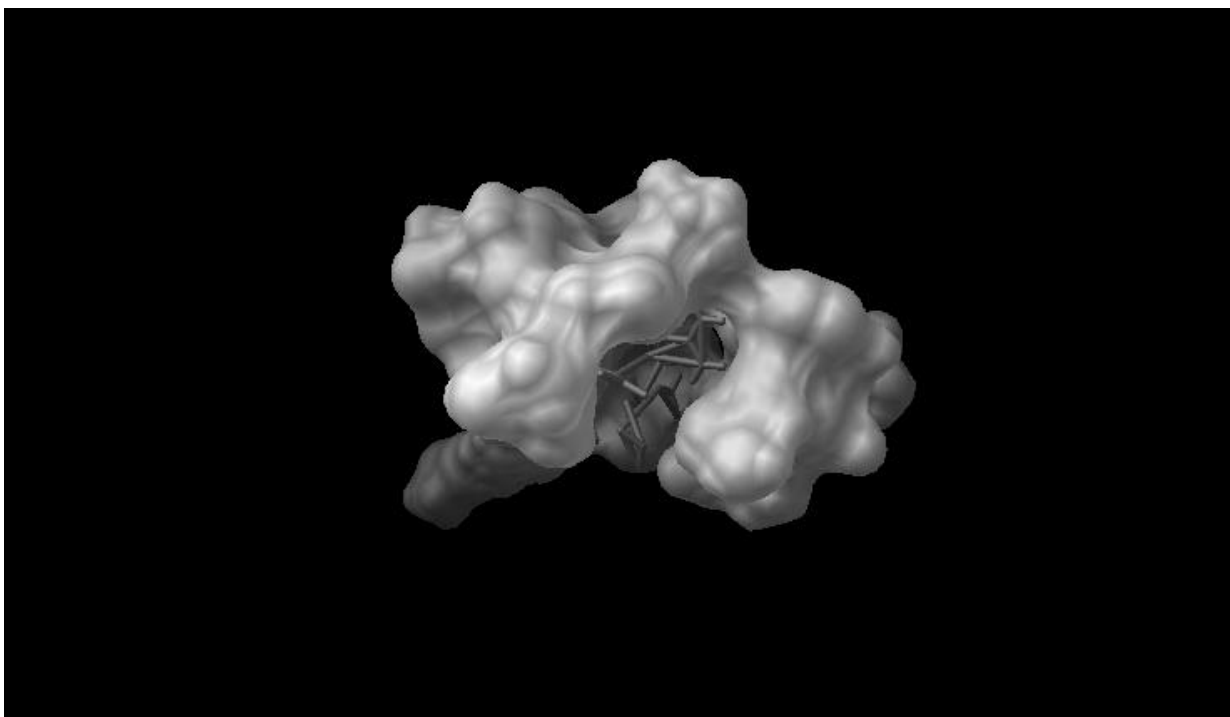


Figure 3. Rendering of laninamivir computationally docked with the active site of PDB 3BEQ. The molecular surface of the receptor is shown in white; the inhibitor, in stick form in grey. Only the interior, inhibitor-containing region of the molecular surface of the active site can be compared to *in situ* data: the surface distal to the interior is a computational artifact, generated by the assumption that active site is detached from the rest of the receptor.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The inhibition constant computed in this study (~ 7.7 microMolar at ~ 298 K) is comparable to the inhibition constant of neuraminidase inhibitors that are not clinically effective ([10], [11], [13]) against several H1N1 genotypes. That inhibition constant is less than the inhibition constant of oseltamivir (~ 11 microMolar, at 298 K; [14]), and greater than the inhibition constant of zanamivir (298 nanoMolar, at 298 K; [15]), against the 3BEQ active site. This suggests that laninamivir would not be effective against the principal 1918

pandemic mutant, A/Brevig Mission/1/18 H1N1.

2. The docking study reported here assumes that the receptor is rigid. This assumption is appropriate for the binding energy computation for PDB 3BDQ per se. However, the calculation does not reflect what receptor "flexing" could contribute to the interaction of the ligand with native unliganded receptor.

3. The analysis described in Sections 2.0 and 3.0 assumes receptor is in a crystallized form. *In situ*, at physiologically normal temperatures (~ 310 K), the receptor is not in crystallized form. The ligand/receptor conformation *in situ*, therefore, may not be identical to their conformation in the crystallized form.

4. Minimum-energy search algorithms other than the Lamarckian genetic algorithm used in this work could be applied to this docking problem. Future work will use Monte Carlo/simulated annealing algorithms.

5. A variety of torsion and charge models could be applied to this problem, and future work will do so.

6. 3BEQ has two chains, each with its own active site. The work described in this paper was performed on Chain A only. Chain B appears to have an active site highly similar to the Chain A active site. Future work will assess the ligand/receptor binding energies of Chains B.

5.0 Acknowledgements

This work benefited from discussions with Tony Pawlicki. For any problems that remain, I am solely responsible.

6.0 References.

- [1] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49.
- [2] Johnson NP and Mueller J. Updating the accounts: global mortality of the 1918-1920 "Spanish " influenza pandemic. *Bulletin of the History of Medicine* 76 (2002), 105-115.
- [3] World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization* 58 (1980), 585-591.
- [4] Vavricka CF, Li Q, Wu Y, Qi J, Wang M, Liu Y, Gao F, Liu J, Feng E, He J, Wang J, Liu H, Jiang H, and Gao GF. Structural and functional analysis of laninamivir and its octanoate prodrug reveals group specific mechanisms for Influenza NA inhibition. *PLoS Pathogens* 7 (October 2011): e1002249. doi:10.1371/journal.ppat.1002249.
- [5] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.
- [6] Xu X, Zhu X, Dwek RA, Stevens J, and Wilson IA. Structural characterization of the 1918 Influenza virus H1N1 neuraminidase. *Journal of Virology* 82 (November 2008), 10493-10501. <http://www.pdb.org/pdb/explore/explore.do?structureId=3BEQ>.
- [7] US Centers for Disease Control. *Summary: Interim Recommendations for the Use of Influenza Antiviral Medications in the Setting of Laninamivir Resistance among Circulating Influenza A (H1N1) Viruses, 2008-09 Influenza Season*. 19 December 2008. URL <http://www.cdc.gov/flu/professionals/antivirals/summary.htm>.
- [8] Luo M. Structural biology: antiviral drugs fit for a purpose. *Nature* 443 (7 September 2006), 37-38. doi:10.1038/443037a, published online 6 September 2006.
- [9] Morris GM, Goodsell DS, Huey R, Lindstrom W, Hart WE, Kurowski S, Halliday S, Belew R, and Olson AJ. *AutoDock* v4.2. <http://autodock.scripps.edu/>. 2010.
- [10] Drug Bank. *Zanamivir*. <http://www.drugbank.ca/drugs/APRD00378>.
- [11] Govorkova EA et al. Comparison of efficacies of RWJ-270201, zanamivir, and oseltamivir against H5N1, H9N2, and other avian influenza viruses. *Antimicrobial Agents and Chemotherapy* 45 (2001), 2723-2732.
- [12] Huey R and Morris GM. *Using AutoDock 4 with AutoDock Tools: A*

Tutorial. 8 January 2008.
<http://autodock.scripps.edu/>.

[13] Cheng Y and Prusoff WH. Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. *Biochemical Pharmacology* 22 (December 1973), 3099–3108. doi:10.1016/0006-2952(73)90196-2.

[14] Horner JK. Simulated docking of oseltamivir with the 1918 pandemic strain

Influenza A/H1N1 zanamivir-conformed neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. 2011. pp. 130-135.

[15] Horner JK. Simulated docking of zanamivir with the 1918 pandemic strain Influenza A/H1N1 neuraminidase active site. *Proceedings of the 2011 International Conference on Genetic and Evolutionary Methods*. CSREA Press. pp. 136-142

SESSION

SOFTWARE PACKAGES AND OTHER COMPUTATIONAL TOPICS IN MEDICIN, BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Chair(s)

TBA

ScaffoldScaffolder: An Aggressive Scaffold Finishing Algorithm

P. Bodily¹, J. Price¹, M. Clement¹, and Q. Snell¹

¹Department of Computer Science, Brigham Young University, Provo, Utah, USA

Abstract—With next generation sequencing technologies producing vast amounts of nucleotide data, it becomes imperative to streamline and automate the genome assembly process as much as possible. Contig scaffolding algorithms, ideally designed to reconstruct full chromosomes, more often tend to produce a still intractable number of disjoint sequences, requiring further manual finishing of the genome. To this end we present ScaffoldScaffolder, an aggressive automated scaffold finisher which further reduces the scaffold set using paired-end data. We evaluate the performance of ScaffoldScaffolder on Newbler scaffolds created from the *Rubus idaeus* cultivar heritage raspberry species. Further automated genome finishing methods are discussed.

Keywords: Genome Assembly; Scaffolding

1. Introduction

1.1 Motivation

Genetic variation is the root cause for numerous diseases or predispositions to life-threatening diseases such as cancer and heart disease. Genetic variation in plants is the basis for variability in crop yields, nutritional value, and flavor. The future of scientific study in these areas depends heavily on the ability to study and characterize genetic variation.

Despite the direct bearing that genetics has on each of these instances, the specific genetic variations at play are not well-characterized, their effects are not well-understood, and the ability to scientifically study them is limited. To a large extent this is due to the relatively sparse amount of data that is available. This shortage derives in large part from the cost-prohibitive and somewhat unrefined nature of the technology and software used to obtain and analyze genetic data. The first human genome was sequenced less than 10 years ago and cost upwards of 3 billion dollars. Though DNA sequencing costs have decreased significantly, the time and manual effort required to produce finished genome sequences are still very restrictive. Despite global efforts to collect and sequence any and all forms of life, only about 1,200 organisms have been sequenced, most at a primitive level, hardly enough to begin to adequately characterize the patterns responsible for genetic variations of interest¹. In order to deduce and characterize the effects of genetic variation, we need a larger number of high-quality sequenced genomes, implying the need for improved technology and software to produce them.

To this end we have undertaken to develop ScaffoldScaffolder, an automated scaffold finisher.

1.2 Background

Technology has thus far been only moderately successful at solving the problem of genomic sequencing. The most prominent methods require large-scale replication of genetic material which is then broken through sonication into an amalgam of short fragments of various sizes (called *reads*). From this mixture are extracted sequences suitable to the sequencing capacity of sequencing machines. The most cost-effective machines are capable of sequencing reads of approximately 100 bases while maintaining reasonably low error rates. Reads as long as 600 base pairs can be sequenced at a much higher price and with slightly higher error rates. In any case, the sequencers are unable to sequence anything that even begins to approximate the size of an entire chromosome which, for example in a raspberry, averages lengths of several million base pairs. The algorithmic challenge is to reassemble the full-length chromosomes from short DNA reads. The genome reconstruction is divided into two phases: the overlapping of reads to form consensus contigs and the scaffolding of contigs to form chromosomes.

In the initial phase of the *assembly* process reads are used to form long contiguous sequences of known bases. This is accomplished by combining overlapping reads to produce longer consensus sequences called *contigs* (see Figure 1a). If all read-length genomic sequences were unique, we could continue this process until we reconstructed the original chromosomal sequence in its entirety. However, due to the presence of repetitive regions throughout a genome, reads will exist which support multiple paths of reconstruction (see Figure 1b). The ambiguity of this result is often modeled as a graph where the nodes are the unambiguous consensus contig sequences produced from combining overlapped reads and the edges are possible ways in which these contigs could be sequentially combined (see Figure 1c). Often the number of contigs can outnumber the actual number of chromosomes by as much as a factor of 10^3 . The graph will often be missing nodes or edges due to insufficient coverage of certain areas of the genome or by erroneous contigs produced from errors during the read-sequencing phase.

Scaffolding is the step in the assembly process where additional information is leveraged to infer the relative distance and orientation of contigs. This is most commonly done using *paired-end data*. Paired-end data consists of pairs

¹http://www.ncbi.nlm.nih.gov/About/tools/restable_mol.html

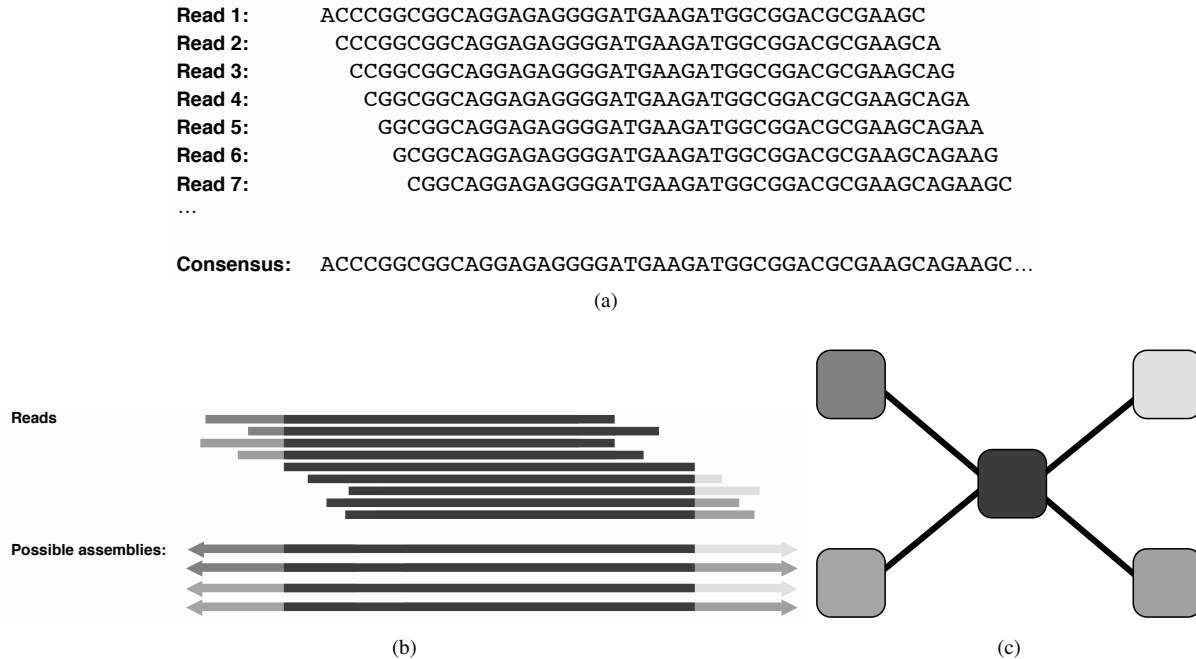


Fig. 1: (a) Short reads whose sequences overlap are overlaid such that their consensus is a reconstruction of the original sequence from which the reads are taken. (b) Repetitive regions whose length exceeds that of sequenced reads create different possible reconstructed consensus paths. (c) The different reconstructions can be modeled as a graph where unambiguous consensus sequences are collapsed into nodes and evidence for the different paths are represented as edges.

of short reads whose distance and orientation is known from the technique used to sequence them. Due to the read-size constraints mentioned above, the paired-reads are the same length or shorter than normal unpaired DNA reads. However, the paired-reads are sequenced from either end of a longer *insert* sequence of known length using one of a number of paired-end sequencing technologies (see Figure 2a). Unique mappings of the paired reads to the pre-determined contigs are supporting evidence for the inference of distance and orientation of contigs (see Figure 2b). Scaffolding thus aims to reconstruct the chromosomal sequences by orienting the contigs and fixing them at distances suggested by paired-end linkages (see Figure 2c). The *gaps* are reported using the inferred number of bases in the gap (denoted using the letter 'N'). The goal of scaffolding is to continue to properly orient and fix contigs at the correct distances until the number of scaffolds approaches the number of expected chromosomes. The quality of an assembly notably increases by using a large variety of clone sizes in the scaffolding phase [1]. However, additional measures are required to reduce the resulting scaffold number to the chromosome number.

We have developed ScaffoldScaffolder, a lightweight tool designed to automate the scaffolding of scaffolds using paired-end data. Whereas it is the purpose of a scaffolder to recover the orientation and placement of contigs inasmuch as the data will accurately allow, the purpose of the ScaffoldScaffolder is to act as a post-processing step to

aggressively reduce the number of sequences as much as possible by leveraging remaining unused linkages inferred from paired-end data. Rather than simply concatenating resulting scaffolds in random order, at random distances, and in random orientations, ScaffoldScaffolder attempts to infer the correct scaffolding, though in a somewhat less cautious manner than a scaffolder.

2. Related Work

The Newbler assembler, developed by 454 Life Sciences and distributed with 454 sequencing machines, has been used on a number of assembly projects [2], [3], [4]. Its efficiency in contigging is particularly notable given that it works natively with the .SFF data format to account for the specifics of pyrosequencing errors. Newbler requires uniquely mapping mate-pairs (i.e. paired-reads) as scaffolding evidence, disregarding reads which potentially map to multiple contigs.

Bambus [5] uses mate-pair information together with other types of linking data to infer the orientation and ordering of contigs to hierarchically construct scaffolds. The linkage data is used to create a graph where nodes are contigs and edges represent linkage evidence. Unlike many scaffolding algorithms, Bambus does not disregard ambiguous linkage evidence (for example multiply mapped pairs), and is capable of outputting pertinent data for manual finishing of ambiguous paths. However, this data is not

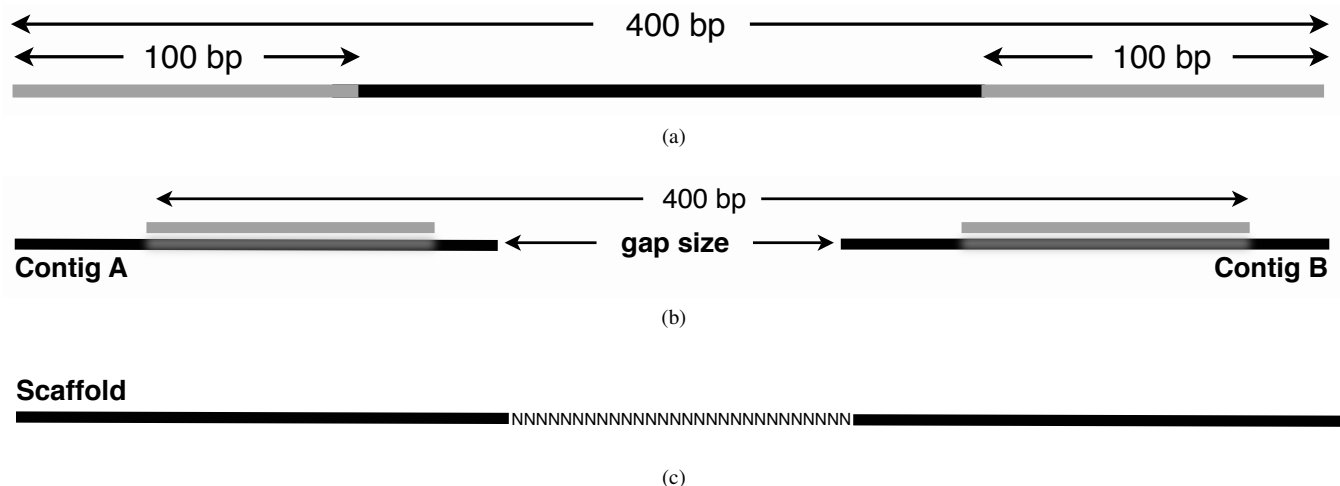


Fig. 2: (a) Paired-end reads or mate-pairs are formed by sequencing the ends of a sequence of known length. (b) Because the orientation and distance of the paired-end reads is known, they can be used to position and orient contigs relative one to another. (c) The result is a reconstructed sequence composed of known and unknown regions. Unknown bases are denoted using the letter 'N'.

used to inform the finishing algorithm in the case that automated finishing is required. Rather this simple greedy algorithm repeatedly finds the longest non-self-overlapping path without consideration of graph structures characteristic of repetitive or polymorphic sequences.

Arachne [6], [7] is a Whole Genome Shotgun (WGS) assembler which has been used to assemble heterozygous genomes [8]. In the contigging phase, Arachne uses depth of coverage and the presence of conflicting links as evidence of repetitive regions in order to avoid erroneous extension of contigs. These contigs are incorporated in filling intra-scaffold gaps.

SOAPdenovo is a short-read assembly algorithm developed by the Beijing Genomics Institute (BGI) which has been employed in a large number of genome projects [9]. The program is designed primarily to function with Illumina GA short reads in reconstruction of large genomes.

MAIA [10] integrates multiple de novo and comparative assemblies by creating a graph of the contigs from these assemblies and their alignments. Four properties for the edge weighting are implemented, namely contig length, overlap length, length of non-aligned overhang, and original assembly quality. This approach makes it possible to use specific assemblers for different next-generation data sources and enables the use of multiple known related genomes in the assembly process. The algorithm was applied on the de novo sequencing of the *Saccharomyces cerevisiae* and demonstrated improvements upon single assembly methods (Velvet, Celera, MAQ) and other hybrid methods (Velvet, Minimus). The disadvantages are that MAIA inherently relies on a very closely related genome in the assembly process and the computational expense of the algorithm

renders the approach impractical for larger genomes. The algorithm, like many, is designed for use with homozygous genomes.

3. Methods

ScaffoldScaffolder is designed to be used as an iterative algorithm where each successive iteration utilizes a paired-end library of a larger insert size than the previous iteration. Each iteration requires as input a series of sequences to be scaffolded in fasta format and any number of similarly-sized paired datasets. The high-level purpose of the algorithm is to use the paired datasets to infer scaffoldings of the input sequences and then to select and output an unambiguous subset of the scaffoldings in fasta format. It additionally outputs information detailing the specifics of the input sequences which compose the new scaffolds.

Internally the algorithm stores input sequences in the context of a graph where nodes represent sequences and edges between nodes indicate that paired data exists to suggest that two sequences should be scaffolded. We must make a slight modification on how we define nodes in the context of this problem. In classic graph theory, we say that if (u,v) is an edge in a graph $G = (V,E)$, then node v is adjacent to node u . However, in the context of our problem it is possible for two sequences to be adjacent in one of four different orientations. One possible solution to this problem relies on the biological concept of sequence orientation which defines one end of the sequence as the 5' (said *five prime*) or *upstream* end and the opposite end of the sequence as the 3' or *downstream* end inasmuch as DNA synthesis proceeds in a 5' to 3' direction. This directionality is an inherent characteristic of each sequence.

One way to uniquely distinguish between the four different orientations of two DNA sequences is to specify which two ends are adjacent (see Figure 3a). We can model this in our graph by defining our nodes as *bi-terminal*, where edges to one terminal represent adjacency to the 5' end of the represented sequence and edges to the second terminal represent adjacency to the 3' end of the same sequence (see Figure 3b). This concept of a graph can be reduced to the standard definition of a graph by making each terminal its own node and creating an edge between them.

In the ScaffoldScaffolder, this *scaffold graph* is initialized only with the sequence nodes; edges are later progressively added as each input dataset is processed for linking evidence. It is assumed that contigs represent unique sequences and thus there is a one-to-one relationship between nodes and sequences.

The algorithm uses an external mapping algorithm, Bowtie [11], to map reads in the input paired-end datasets to the sequences to be scaffolded. While the algorithm is heavily modularized to support other mapping algorithms (including GNUMAP [12] and BLAST [13]), testing has been limited

to Bowtie. Experimentation to date has required mappings to be unique (meaning no more than a single alignment location exists for the mapped read) in order to maximize confidence in the resulting scaffolds. ScaffoldScaffolder currently gives the user the option of adjusting this parameter as well as parameters dictating read-trimming options, alignment-mismatch options, and options for skipping the first n reads in a dataset. A parameter allows the user to specify the paired-end orientation either as *-fr* (Illumina paired-end protocol), *-rf* (Illumina mate-pair protocol), or *-ff* (454 mate-pair protocol).

From the Bowtie results ScaffoldScaffolder then identifies pairs for which both ends are uniquely mapped. In cases where both ends map within the same sequence, the distance between the mappings is cataloged in order to infer an insert size for the library. In cases where ends map to distinct sequences, the algorithm infers the orientation and gap size between the two base sequences. Assuming that the gap size is viable (i.e. nonnegative), the weight of the corresponding edge in the scaffold graph is linearly incremented and the inferred gap size for the scaffolding of the two oriented base sequences is cataloged. The final gap size is the mean of the inferred gap sizes.

The process of mapping paired-reads and then loading the scaffold graph according to the mapping results is repeated for each provided paired-end source in the respective iteration of the algorithm. At the conclusion of this phase, the scaffold graph contains a number of ambiguous linkages where a given base sequence may have multiple possible scaffoldings in the upstream and/or the downstream direction. ScaffoldScaffolder assumes that a given base sequence will be scaffolded with only one sequence in either direction. In order to reduce the graph to include an unambiguous subset of scaffold edges, the edges are sorted by weight, following which edges are greedily considered for inclusion in the final graph. If adding an edge creates an ambiguous scaffolding, the edge is skipped. A minimum support parameter determines the minimum number of unique pairs required as support for an edge to be included.

Scaffold sequences are constructed from the disambiguated scaffold graph and these sequences, together with any unscaffolded sequences, are output in fasta format by decreasing order of length.

4. Results

We tested the ScaffoldScaffolder algorithm on Newbler scaffolds created for the heterozygous *Rubus idaeus cultivar heritage* raspberry genome.

Contigs were first assembled from the reads using the Newbler assembler. Due to memory and time constraints, the 5k dataset was not incorporated into the Newbler assembly. Aside from this exception, the same data used in the Newbler assembly was used as input to ScaffoldScaffolder.

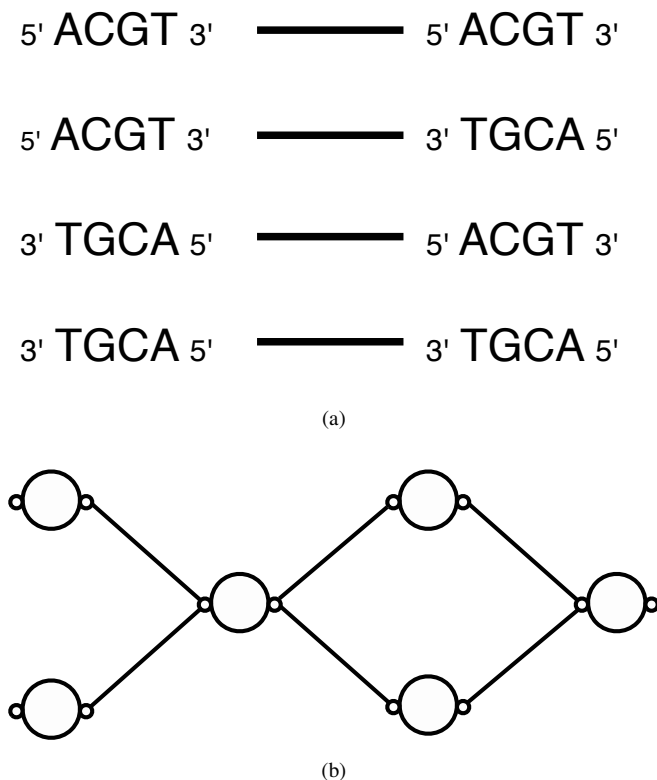


Fig. 3: (a) There are four possible ways for two sequences to be adjacent. The correct orientation can be uniquely defined by specifying which ends are adjacent. (b) Orientation can be preserved in a graph model using bi-terminal nodes where each terminal represents a sequence end.

Table 1: Reduction of Newbler Scaffolds via Multiple Iterations of ScaffoldScaffolder

Iteration (insert size)	Reads Uniquely Aligned	Scaffold Count	Max Scaffold Size	Avg Scaffold Size
<i>Initial</i>		13,037	4,456,429	19,313
400b	171,490,959	11,620	4,456,429	21,905
3kb	397,429	11,271	4,456,429	22,643
5kb	99,333,568	8,695	4,456,429	30,976
20kb	893,745	7,638	4,678,214	37,961

The assembler parameters were set to require a minimum length of 30 bases, a minimum overlap length of 70 bases, and a minimum overlap identity of 98 bases. The large genome assembly, heterozygotic, and scaffold flags were enabled. Using these parameters, Newbler produced 123,121 contigs and 13,037 scaffolds.

ScaffoldScaffolder was parameterized to use Bowtie for the mapping of paired reads, with a maximum of 3 mismatches, and only allowing uniquely mapping reads. The minimum support required for valid links was 1.

ScaffoldScaffolder was able to reduce the scaffold count by 5399, representing a reduction of over 40% of the scaffolds produced using Newbler's scaffolding algorithm alone (see Table 1).

The 3kb and 20kb datasets (those produced using the 454 mate pair protocol) had noticeably lower rates of alignment. We suspect this derives from the inability of the Bowtie aligner to consider insertions or deletions when aligning reads. This proves troublesome for reads sequenced using the 454 protocol which often have insertions and deletions in homopolymorphic sequences. Consequently, selection of other short-read mapping algorithms capable of handling indels could further improve the performance of the ScaffoldScaffolder algorithm.

5. Discussion

ScaffoldScaffolder attempts to provide an algorithmic solution to automated finishing using paired-end data. Although it may be argued that the aggressive nature of the algorithm will lead to inaccuracies in the resultant assembly, similar inaccuracies are common to other prevalent finishing methods of which we will briefly discuss two.

5.1 Genetic Linkage Map

Biological assays are capable of inferring the relative distance along chromosomes of a number of specific genetic sequences based on what is called the *recombination rate* of protein-coding sequences (i.e. genes). Recombination refers to the rearrangement and exchange of genetic material that occurs when chromosomes cross over one another. The likelihood of such a rearrangement occurring between two genes, known as the recombination rate, increases as a function of the distance between the two genes. Thus the

relative distance and ordering of certain observable genes can be inferred biologically in order to create a *genetic linkage map*. These genes, whose sequences are known, can be used to guide the finishing of the assembly. Assuming that such a genetic linkage map is available, this process is quite accurate, but may still fail to place a number of scaffolds.

5.2 Related Genomes

A second approach to genome finishing attempts to infer the distance and orientation of contigs by using the known sequence of a closely related genome. We refer to the degree of genetic similarity in gene-order between different species as *synteny*. To the extent that the genomes of two species are syntenic, the ordering and orientation of similar sequences on the related genome can be used to guide the assembly of the target scaffolds. The challenge with this approach is proper identification and treatment of genomic differences.

6. Conclusion

In this research, we present ScaffoldScaffolder, an aggressive automated scaffold finisher. We have illustrated its effectiveness in significantly reducing a set of Newbler scaffolds created for the *Rubus idaeus cultivar heritage* raspberry genome. Future development aims to address the complexities of scaffolding heterozygous genomes with inclusion of structural/sequence-based heuristics to identify and assemble distinct haplotypes. Improved input data analysis will aim to infer parameters so as to reduce the information required from the user for execution.

References

- [1] A. Zharkikh, M. Troggio, D. Pruss, A. Cestaro, G. Eldridge, M. Pindo, J. T. Mitchell, S. Vezzulli, S. Bhatnagar, P. Fontana, R. Viola, A. Gutin, F. Salamini, M. Skolnick, and R. Velasco, "Sequencing and assembly of highly heterozygous genome of vitis vinifera L. cv pinot noir: Problems and solutions," *Journal of Biotechnology*, vol. 136, no. 1-2, pp. 38 - 43, 2008, genome Research in the Light of Ultrafast Sequencing Technologies. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0168165608001880>
- [2] J. Miller, A. Delcher, S. Koren, E. Venter, B. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton, "Aggressive assembly of pyrosequencing reads with mates," *Bioinformatics*, vol. 24, no. 24, p. 2818, 2008.
- [3] M. Pop, "Genome assembly reborn: recent computational challenges," *Briefings in bioinformatics*, vol. 10, no. 4, pp. 354-366, 2009.

- [4] J. Reinhardt, D. Baltrus, M. Nishimura, W. Jeck, C. Jones, and J. Dangel, "De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*," *Genome research*, vol. 19, no. 2, pp. 294–305, 2009.
- [5] M. Pop, D. Kosack, and S. Salzberg, "Hierarchical scaffolding with bambus," *Genome Research*, vol. 14, no. 1, pp. 149–159, 2004.
- [6] S. Batzoglou, D. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. Mesirov, and E. Lander, "Arachne: a whole-genome shotgun assembler," *Genome research*, vol. 12, no. 1, pp. 177–189, 2002.
- [7] D. Jaffe, J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J. Mesirov, M. Zody, and E. Lander, "Whole-genome sequence assembly for mammalian genomes: Arachne 2," *Genome research*, vol. 13, no. 1, pp. 91–96, 2003.
- [8] J. Vinson, D. Jaffe, K. O'Neill, E. Karlsson, N. Stange-Thomann, S. Anderson, J. Mesirov, N. Satoh, Y. Satou, C. Nusbaum, *et al.*, "Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*," *Genome research*, vol. 15, no. 8, pp. 1127–1135, 2005.
- [9] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, *et al.*, "De novo assembly of human genomes with massively parallel short read sequencing," *Genome research*, vol. 20, no. 2, pp. 265–272, 2010.
- [10] J. Nijkamp, W. Winterbach, M. van den Broek, J.-M. Daran, M. Reinders, and D. de Ridder, "Integrating genome assemblies with maia," *Bioinformatics*, vol. 26, no. 18, pp. i433–i439, 2010. [Online]. Available: <http://bioinformatics.oxfordjournals.org/content/26/18/i433.abstract>
- [11] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.
- [12] N. Clement, Q. Snell, M. Clement, P. Hollenhorst, J. Purwar, B. Graves, B. Cairns, and W. Johnson, "The gnumap algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing," *Bioinformatics*, vol. 26, no. 1, pp. 38–45, 2010.
- [13] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

A Bayesian Network Calculator of the Combined Effects of Climate Change, Chytridiomycosis, and Land-Use Change on Global Amphibian Diversity

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

Abstract

Amphibians are threatened worldwide by climate change, chytridiomycosis, and land-use change. Recently, data sufficient to estimate the combined effects of these threats, at least in the near term, has become available. The combined effect can be modeled as a linear combination of the individual effects. Here I describe EGAD, a Bayesian network implementation of such a model. The tool is especially useful in assessing the sensitivity of the estimate of the combined effect to uncertainty in the components. In addition, it can automatically recalibrate itself as new data becomes available.

1.0 Introduction

Amphibians are threatened worldwide by climate change, chytridiomycosis, and land-use change ([1]). The loss of a large fraction of the world's amphibians could profoundly disrupt the control of insect vectors of a variety of human diseases, including malaria, sleeping sickness, and dengue fever.

EGAD is a Bayesian network ([4]) calculator of the globally averaged "threat" of

- (F) -- fractional climate change
- fractional chytridiomycosis change
- fractional land use change

to species diversity of each of

- (O) -- frogs
- salamanders
- caecilians

based on [1].

All "changes" in (F) are measured as the fraction of the area of Earth's surface inhabited by at least one species in the amphibian orders in (O) in the reference state of (F) (identified in [2]). By fiat, "change", as used in the context of (F), means "change that reduces amphibian diversity". Any increase in the area affected by climate change, chytridiomycosis, or human land use, can reduce amphibian diversity.

EGAD assumes that the net global average threat ("Threat to O", reduction in the species diversity) of each of (O) is linear in (i.e., is a weighted sum of) each of (F).

2.0 Method

The weighting coefficients for each of (F) in the threat formulas for each of (O) are derived from the data is reported in [2], Table S1. [2], Table S1 reports effects by pairs of threats

1. Climate change (CL) and chytridiomycosis (CH)
2. Climate change and land use (LU)
3. Chytridiomycosis and land use

for each member of O.

Analysis reveals that this pair-wise data is modelable as a system of three independent linear equations, one for each of O. This system was solved for each of CL, CH, and LU, for each of O. For each of O, these values were then summed.

The weighting coefficient for a threat factor (F) in the formula for Threat_to_O was computed by dividing the observed value (in Table S1) for each of these factors by that sum. Table 1 shows these weighting factors, in the form N (M), where N is the non-normalized value of the weighting coefficient from Table S1, and (M) is the sum-normalized value of that coefficient.

Table 1. Weighting coefficients of threat terms in formulas for threats to amphibian orders

Order	CL	CH	LU
Frogs	132 (0.377)	109 (0.311)	109 (0.311)
Salamanders	58 (0.249)	152 (0.652)	23 (0.099)
Caecilians	17 (0.202)	2 (0.024)	65 (0.774)

More specifically, the formulas for each Threat_to_O are:

$$\begin{aligned} \text{Threat_to_Frogs} &= (0.377) * \text{Fractional_Climate_Change} + \\ & (0.311) * \text{Fractional_Chytridiomycosis_Change} + \\ & (0.311) * \text{Fractional_Land_Use_Change} \\ \text{Threat_to_Salamanders} &= (0.249) * \text{Fractional_Climate_Change} + \\ & (0.652) * \text{Fractional_Chytridiomycosis_Change} + \\ & (0.099) * \text{Fractional_Land_Use_Change} \\ \text{Threat_to_Caecilians} &= (0.202) * \text{Fractional_Climate_Change} + \\ & (0.024) * \text{Fractional_Chytridiomycosis_Change} + \\ & (0.774) * \text{Fractional_Land_Use_Change} \end{aligned}$$

Bayesian prior probabilities were set at 0.1 for each member of (F). (Other prior probability distribution are of course possible.)

Given these formulas and prior probabilities, for each user-selected combination of Fractional_Climate_Change, Fractional_Chytridiomycosis_Change, and Fractional_Land_Use_Change, EGAD computes the posterior probability of Threat_to_Frogs, Threat_to_Salamanders, and

Threat_to_Caecilians in accordance with Bayes' Theorem ([4], pp. 17, 124).

3.0 Results

EGAD is implemented as a Windows *Netica* ([3]) application. A nominal user view of EGAD is shown in Figure 1.

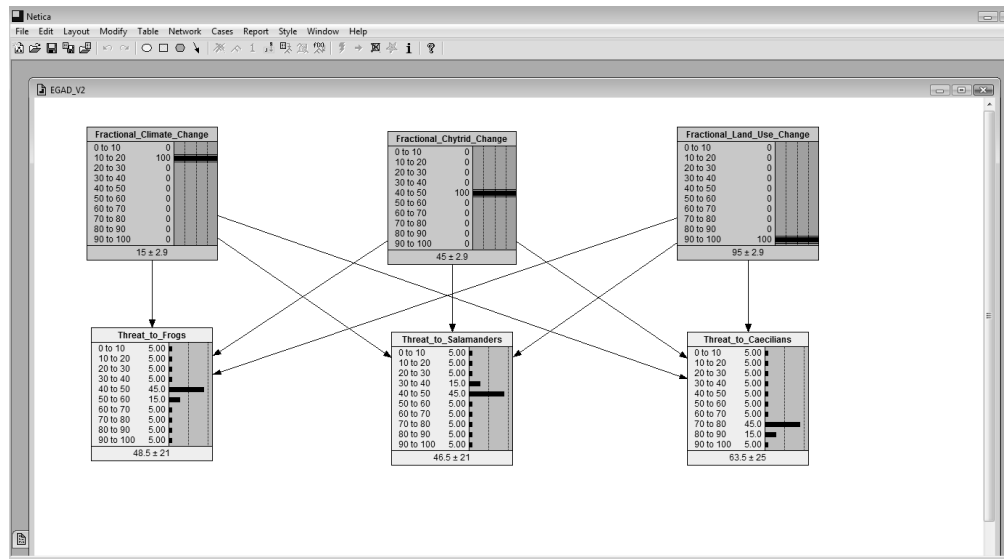


Figure 1. A nominal EGAD user view.

On the screen, there is one box each for each of (F) and one box for each of (O). Arrows depict the dependence of each of (O) on each of (F). The top row of boxes represent (F), expressed as discrete percentage ranges of change from reference values for the members of (F) defined in [2]. The bottom row of boxes represent the fractional decrease in species diversity, expressed as a percentage, based on the user selections for each of (F), one box for each of (O).

Each box in Figure 1 has three regions, delimited by horizontal borders.

The top region of a box contains the name of a (random) variable of interest, e.g., "Fractional_Climate_Change".

The middle region of a box consists of three elements (read horizontally):

- i. a textual 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 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 example, in Figure 1 the box in the lower left is associated with the variable Threat_to_Frogs, conditional on each Fractional_XX_Change. For example, the probability that Threat_to_Frogs has a value lying in "70 to 80 (percent reduction in diversity)" is 0.50.

The bottom region of a box reports the "mean \pm one_standard_deviation" of the distribution shown in (ii). For example, in Figure 2, the mean for the distribution for Threat_to_Frogs is 57.4 (percent) and the standard deviation is 22 (percent).

For typical operation, the top row of boxes have a grey background; the bottom row of boxes, a pink background. 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 1, Fractional_Chytroid_Change is an "input" with an asserted value of "40 to 50" percent. A box with a pink background means the variable corresponding to that box is intended as an "output" (also called a "calculated") variable. For example, in Figure 1, Threat_to_Salamanders is an "output"/"calculated" variable that has a probability distribution, with most of the probability in the "50 to 60" percent range.

The basic operation of *EGAD* is simple. The user places the mouse pointer over a percentage

label in a threat-factor ((F)) box and clicks once. The resulting threat values (fractional reduction in species diversity) in each of (O) will appear in the "Threat_to_O", where O = [frogs | salamanders | caecilians].

EGAD can also analyze problems in which we don't have asserted values for all nominal input variables. Suppose, for example, we don't know what the fractional climate change is, and assume as a starting configuration the one shown in Figure 1. *EGAD* can, given these "inputs", recalculate all the probabilities in the list (see Figure 2). Note how the probability distributions for Threat_to_O change in this setup: they tend to "smear", which is what we would expect, given that this setup contains less information than the setup for Figure 1. Note also that Fractional_Climate_Change, under these settings, defaults to its prior probability distribution.

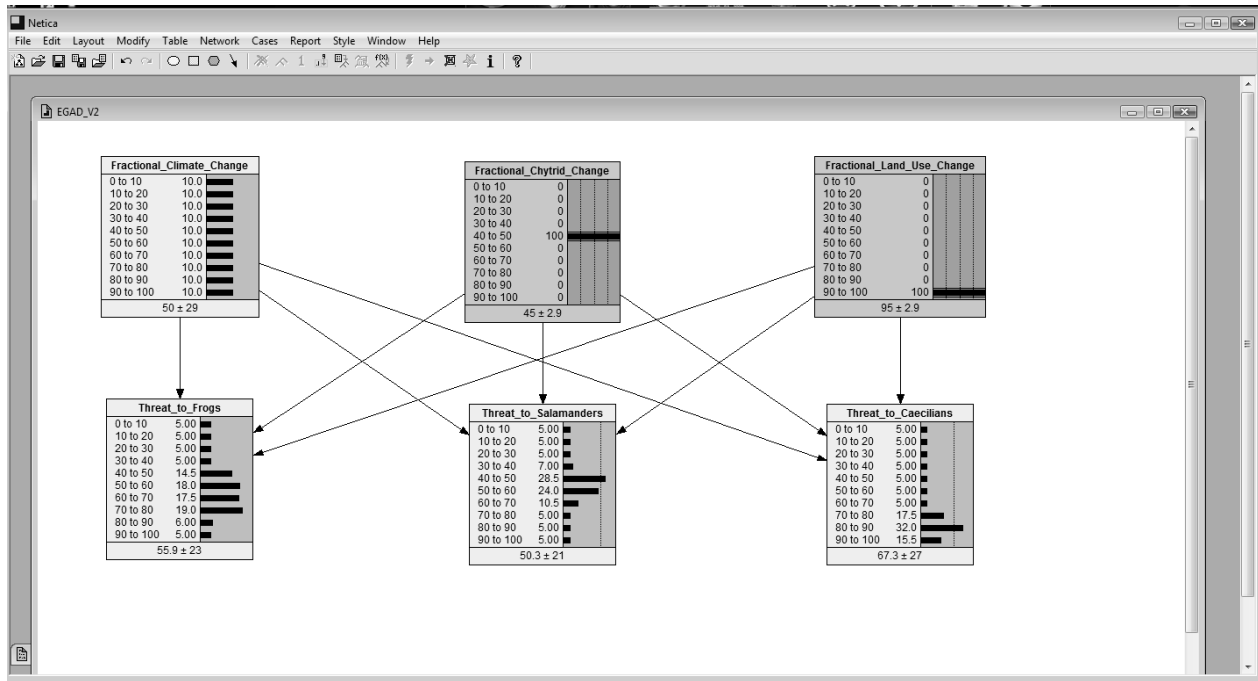


Figure 2. The result of changing Fractional_Climate_Change to an output variable, starting with the configuration of Figure 1.

Because any variable in *EGAD* can be toggled between asserted, and calculated, status, we can use amphibians as "probes" of the nominal input variables. Suppose, for example, we wanted to use changes in salamander diversity to estimate the global change in chytridiomycosis. We could start with the configuration in Figure 1, toggle Fractional_Chytrid_Change to be an output variable (pink background), and toggle Threat_to_Salamaders to be an asserted-value variable (grey background). To toggle the mode of these variables, we place the mouse cursor between a value-range label and the numeric literal to its right in the middle region of the box and click the left mouse button once or twice until the background color has the desired value.

We then place the mouse cursor over "20 to 30" in Threat_to_Salamaders and click the left mouse button once.

The result is shown in Figure 3. *EGAD* determines that the probability that the Fractional_Chytrid_Change has a value lying in the 0 to 10 percent range is 0.4 (40%). In addition, *EGAD* computes new probabilities for Threat_to_Frogs and Threat_to_Caecilians.

This is just one example of the $\sim 10^6$ predictions *EGAD* can make based on different input conditions.

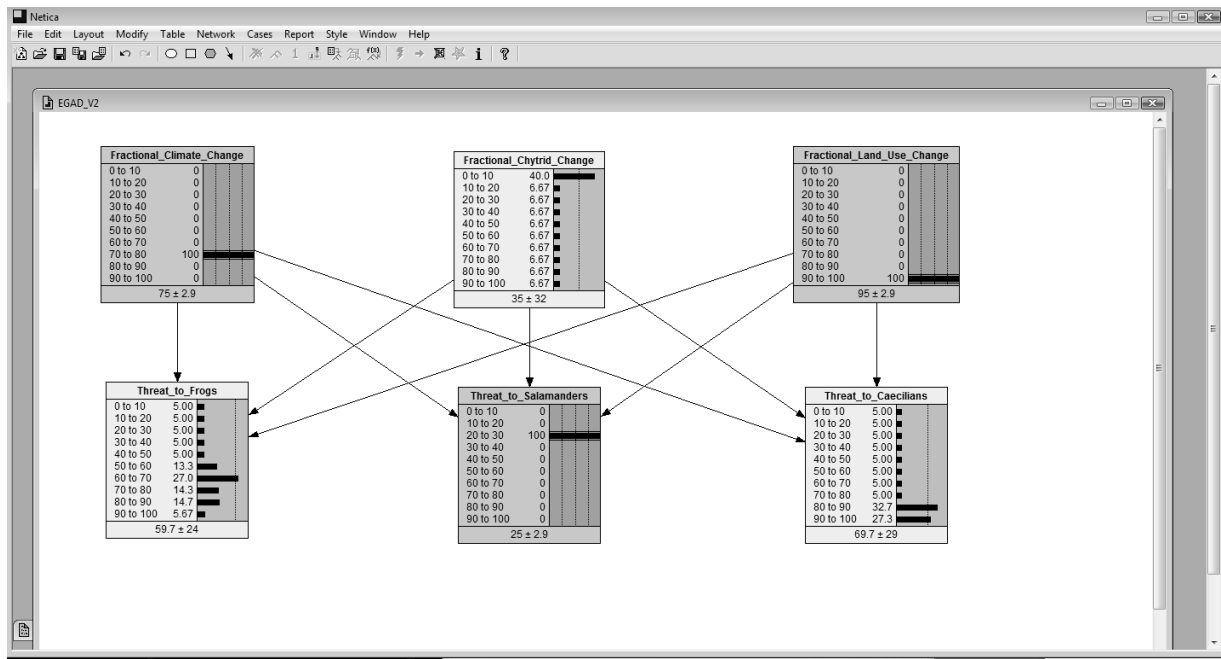


Figure 3. The result of changing Fractional_Chytrid_Change to be an output variable, and Threat_to_Salamaders to be an input variable, then asserting that the Threat_to_Salamaders is "20 to 30" (percent decrease in salamander species diversity), starting with the configuration in Figure 1.

This latter behavior -- the propagation of probability changes across the net -- is one of the most powerful features of Bayesian network modeling. Not only does this kind of model show localized probability changes, but it also shows how such changes constrain probability distributions elsewhere in the net. When a network contains at least one variable that is connected to more than one other variable, such network-wide dependencies can help to support more nuanced and sensitive testing than would be possible in two-variable models.

For any application running under it (such as *EGAD*), *Netica* provides an impressive spectrum of mouse-selectable analysis, network editing, simulation, automated learning-from-data, graphics, and reporting functions. *Netica* also provides a C-like, richly featured programming language and support library. APIs to C, C++, C#, Visual Basic, Matlab, and CLisp are available.

4.0 Discussion and conclusions

1. Does *EGAD* produce results we expect? Providing an exhaustive answer to the question is not tractable, but the tool produces results we expect in several "intuitive" cases. Let $YY = [\text{Climate} \mid \text{Chytrid} \mid \text{Land_Use}]$. Then:

a. If each Fractional_YY_Change is set to "0 to 10", starting with the configuration shown in Figure 2, most of the probability in the distributions shown in Threat_to_O becomes "0 to 10". In other words, *EGAD* predicts small changes in the amphibians' environment produces small changes in their species diversity.

b. If each Fractional_YY_Change is set to "90 to 100", starting with the configuration shown in Figure 2, most of the probability in the distributions shown in Threat_to_O becomes "90 to 100". In other words, *EGAD* predicts large changes in the amphibians' environment produces large changes in their species diversity.

c. If each Fractional_YY_Change is set to "40-50" ["50-60"], starting with the configuration shown in Figure 2, most of the probability in the distributions shown in Threat_to_O becomes "40 to 50" ["50 to 60"].

d. Based on the coefficients shown in Section 3.0, we expect changes in chytridiomycosis to affect salamanders most, frogs somewhat less so, and caecilians least. If we start with the configuration shown in Figure 2, and set each of climate change and land use to "0 to 10", then change the asserted value of Fractional_Chlytrid_Change one value-interval at a time, the probability distributions in Threat_to_O exhibit exactly the expected numerical order of effects.

e. Based on the coefficients shown in Section 2.0, we expect changes in land use to affect caecilians most, frogs somewhat less so, and salamanders least. If we start with the configuration shown in Figure 1, and set each of climate change and chytrid change to "0 to 10", then change the asserted value of Fractional_Land_Use_Change one value-interval at a time, the probability distributions in Threat_to_O exhibit exactly the expected numerical order of effects.

2. *EGAD* depends heavily on the data in [1] and [2], on the assumption that the combined effect is linear in its components, and to a lesser degree, on the uniform prior probability posits noted above.

3. The tool is especially useful for providing probability-constrained predictions when we have only partial information about threats or effects.

5.0 References

[1] Hof C, Araújo MB, Jetz W, and Rahbek C. Additive threats from pathogens, climate, and land-use change for global amphibian diversity. *Nature* online publication. doi:10.1038/nature10650. 2011.

[2] Supplementary Information for [1].
<http://www.nature.com>.

[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.

A Web-based multi-Genome Synteny Viewer for Customized Data

Kashi V. Revanna¹, Chi-Chen Chiu², Daniel Munro¹, Alvin Gao³, and Qunfeng Dong^{1,2}

¹Department of Biological Sciences,

²Department of Computer Science and Engineering, and

³The Texas Academy of Mathematics and Science, University of North Texas, Denton, Texas, USA

Abstract - *Web-based synteny visualization tools are important for sharing data and revealing patterns of complicated genome conservation and rearrangements. Such tools should allow biologists to upload genomic data for their own analysis. Recently, we published a web-based synteny viewer, GSV, which was designed to satisfy the above requirement [1]. However, extending the functionality of GSV to visualize multiple genomes is important to meet the increasing demand of the research community.*

We have developed a multi-Genome Synteny Viewer (mGSV). Similar to GSV, mGSV is a web-based tool that allows users to upload their own genomic data files for visualization. Multiple genomes can be presented in a single integrated view with an enhanced user interface. Users can navigate through all selected genomes to examine conserved genomic regions as well as the accompanying genome annotations. A web server hosting mGSV is provided at <http://cas-bioinfo.cas.unt.edu/mgsv>.

Keywords: synteny, genome browser, visualization, bioinformatics

1 Background

Since patterns of genome conservation and rearrangements can be complicated, visualization tools are critical to reveal those patterns. A variety of web-based synteny visualization tools exist for this purpose (e.g., SynBrowse [2] and CoGe [3]). Compared to standalone bioinformatics software, those web-based analysis tools are more convenient for users since no local software installation or maintenance is necessary. However, some of these tools only allow users to analyze a small number of pre-selected genome sequences available at those web resources. This limitation is becoming a serious issue since biologists often need to examine synteny for their own sequences of interest that are typically not available at those web resources.

2 Design and Implementation

To use the mGSV web tool, users submit one or two input files, as described below, and are then presented with

first a synteny overview page, and then the main synteny browser.

2.1 mGSV input files

The synteny data file allows users to specify the genomic location of each conserved region in each pair of genomic sequences. Users can provide additional information such as alignment score or percentage of similarity or identity to characterize each of the conserved regions, which can then be used to filter regions shown in the synteny display. An optional genome annotation file can also be submitted to list the accompanying genomic features (e.g., genes) to be displayed as annotation tracks along with the reference genomes.

2.2 Synteny overview page

After the data upload, users are first presented with an overview display, in which all the input genomes are arranged in a circle showing the overall conserved regions among each other. An “Associations Provided” table is also shown in the overview page listing all pairs of genomes specified in the user-uploaded input data and the number of conserved regions for each pair. When the genome order has been chosen, the user is brought to the main synteny browser.

2.3 Main synteny browser

At the top of the main synteny browser, multiple pull-down menus are available that allow users to select specific genomes to display in the order of their choice. Additional pull-down menus can be added and removed, so that each genome can be displayed more than once if necessary. Buttons at the top left corner allow users to control all the genomes displayed by zooming in/out, moving left/right or viewing entire genomes on all genomes. mGSV is then divided into two main display windows with control panels (for zoom and filtering functions) on the left and synteny displays on the right.

The conserved regions between any pairs of selected genomes are displayed as colored translucent blocks. When users click on a conserved region, a pop-up menu appears showing its numerical start and end positions. Users can zoom in/out, move left/right or select specific regions on

individual genomes for display by using the embedded control panels on the left of the view. Users can also filter the conserved regions based on their associated characteristics listed in the synteny files such as length of the conserved regions, similarity score, and so on.

If an annotation file is also provided, a selected annotation track (e.g., gene) will be displayed inside each selected genome. Users can easily switch among the tracks or change the colors and shapes of the selected tracks on the fly.

3 Discussion

Although embedding sequence comparison software may facilitate users, we have chosen not to do so in mGSV mainly for three reasons: (1) Sequence comparison among large genomes is not often practical at a web server due to heavy computational demands. (2) It is unrealistic for a centralized web server to decide which software or methods users should use for their data set. (3) Sequence comparison is not the only means for synteny identification. Other types of data (e.g., genetic mapping) may also provide synteny information.

4 Conclusions

mGSV is a web-based synteny visualization tool that enhances the original functionalities of GSV by allowing biologists to upload their own data sets and visualize the synteny among multiple genomes simultaneously in a single integrated view. The novel design and the implementation of mGSV provide the research community with an important alternative to currently available tools.

5 References

- [1] Revanna KV, Chiu CC, Bierschank E, Dong Q: **GSV: a web-based genome synteny viewer for customized data.** *BMC Bioinformatics* 2011, **12**:316.
- [2] Pan X, Stein L, Brendel V: **SynBrowse: a synteny browser for comparative sequence analysis.** *Bioinformatics* 2005, **21**(17):3461-3468.
- [3] Lisch D *et al*: **Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids.** *Plant Physiol* 2008, **148**(4):1772-1781.

An S-System Analysis of the Light Response of the Microalga *Chlamydomonas reinhardtii* during Biohydrogen Production

Jack K. Horner
PO Box 266
Los Alamos NM 87544 USA
email: jhorner@cybermesa.com

Abstract

Producing biohydrogen on a commercial scale will likely require the genetic re-engineering of natural hydrogen-producing organisms. Kinetic modeling of hydrogen-producing metabolic pathways can cost-effectively help to characterize systemic (e.g., mass/energy/charge conservation) constraints in these organisms. *In vitro* kinetic studies suggest that the activity of the hydrogenases in several photolytic biohydrogen producers (PBPs) could be increased to as much as four times their nominal *in vivo* rate. It is much less clear, however, whether the *in vitro* activity maximum could be realized *in vivo*. Here I use an S-system photosynthesis-based PBP (PS-PBP) simulator to analyze the light-saturation response of *C. reinhardtii*. The analysis strongly suggests that the H_2 production of the alga cannot be increased at incident light intensities greater than ~ 10 hv.

Keywords: biohydrogen, S-system, metabolic modeling

1.0 Introduction

Kinetic modeling of hydrogen-producing metabolic pathways can cost-effectively help to characterize systemic (e.g., mass/energy conservation) sensitivities in photolytic biohydrogen producers, even if all the details of hydrogen-gas producing metabolic pathways are not known. Among the more promising candidates for hydrogen-production optimization are photolytic biohydrogen producers (PBPs) such as the microalga *Chlamydomonas reinhardtii* ([7], [8]). It is generally held that the hydrogen-producing pathways in many PBPs incorporate segments of the PS-I and PS-II photosynthetic pathways ([6],[13]), and electrons from the anaerobic

degradation of starch, to help accumulate the electron free energy required to allow a hydrogenase to convert protons to H_2 ([14]). *In vitro* kinetic studies suggest that the activity of hydrogenases isolated from several PBPs could be increased to as much as four times their nominal *in vivo* rate ([1]). Among other constraints *C. reinhardtii* exhibits a light-saturation response, hypothesized to arise from "shading" of the light-receptor structures by each other, and by electron-throughput limitations of the organism's light receptors ([17]). Here I use *bioh2gen* ([15]), an S-system ([2], [11]) PS-PBP kinetics simulator, to argue that within the context of the model, the H_2 production of *C. reinhardtii* cannot be increased with incident light intensities greater than ~ 10 hv.

2.0 S-systems

An S-system ([11],[12]) is a power-law-oriented, differential, difference-equation

system of ordinary differential equations (SODE) each of whose dependent variables X_i is described by a kinetic equation of the form

$$dX_i/dt = \alpha_i \prod_j X_j^{g_{i,j}} - \beta_i \prod_j X_j^{h_{i,j}}$$

Eq. 2.1

where

- the left-hand side of Eq. 2.1 is the first derivative of X_i with respect to time
- $i, j = 1, 2, 3, \dots, N$
- $\{X_i\}$ is the set of real-valued dependent variables of the system
- for any given X_i , only those independent and dependent variables X_j that have an action on X_i are included as factors in the products on the right-hand-side (RHS) of Eq. 2.1. The factors in the first term on the RHS of Eq. 2.1 correspond to just those entities that increase or inhibit the production of X_i ; the factors in the second term of the RHS of Eq. 2.1 correspond to just those entities that contribute to, or inhibit, the consumption of X_i .
- $\alpha_i, \beta_i > 0$
- $g_{i,j}, h_{i,j}$ are real-valued

There is a natural mapping from a biochemical map, K , to equations that have the form of Eq. 2.1. In particular, let $K = \langle \{X_k\}, E \rangle$, $E \in \{X_k\} \otimes \{X_k\}$, $k = 1, 2, \dots, N$, be a directed graph in which each distinct $X_i \in \{X_k\}$ corresponds to a distinct variable (e.g., the concentration of a distinct chemical species in the map), and $w \in E$ if and only if $w = (X_m, X_n)$ is a directed edge in K , $m \neq n = 1, 2, \dots, N$.

α_i and β_i are called *generalized rate constants* (or just rate constants) for X_i , and $g_{i,j}$ and $h_{i,j}$ are called the *generalized kinetic orders* (or just kinetic orders) for X_i , on analogy with standard chemical kinetic

theory. The subexpression i_j indicates the action of X_j on X_i .

An S-system has several desirable features, including the fact that it is fully characterized by its rate constants and kinetic orders. Any SODE can be *recast* ([10],[11]) as an S-system without loss of accuracy or precision; the recasting, however, is not in general unique. In addition to biochemical systems, S-systems have been successfully used to model epidemics, forest diversification, and world dynamics.

3.0 A network model of hydrogen production in PS-PBPs

I will call bioH₂ producers that exploit portions of the PSII or PSI pathways “photosynthetic” PBPs (PS-PBPs). The schematized PS-PBP model used in the

present study is shown in Figure 1 and is similar to [3], [4], [5], [9] and [14]. It represents a consensus working hypothesis held by the biohydrogen research community about the high-level metabolics of hydrogen production in PS-PBPs ([7]).

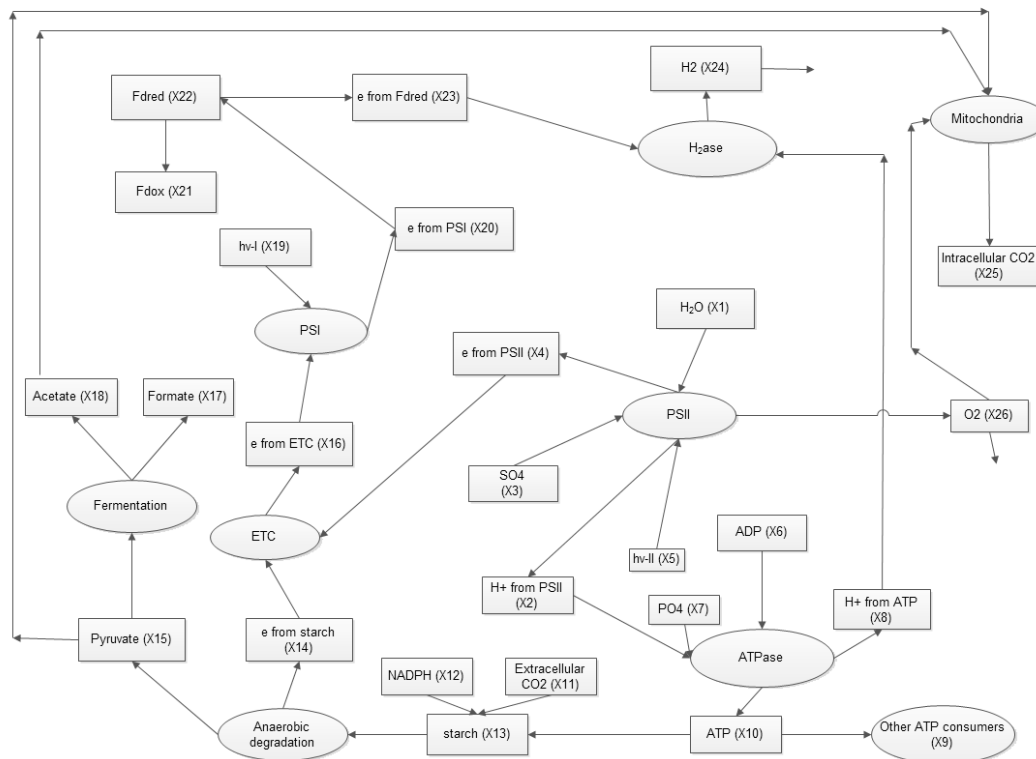


Figure 1. Schematized hydrogen producing metabolic network for PS-PBPs. Rectangles represent sources or sinks of physical quantities of interest (such as mass, concentration, or photon count) named in those rectangles, ellipses represent transforms (which may be complexes of reactions not individually modeled here), and an arrow from an ellipse to a rectangle means that the transform named in the ellipse affects the quantity/concentration of the chemical species named in the rectangle. Legend: PSI = photosynthesis stage I; PSII = photosynthesis stage II; SO₄ = sulfate; hv-I = photons incident to PSI; hv-II = photons incident to photosynthesis PSII; ADP = adenosine diphosphate; ATP = adenosine triphosphate; PO₄ = inorganic phosphate; O₂ = oxygen gas; ATPase = adenosine triphosphatase; e from starch = electrons from anaerobic starch degradation; H₂ase = hydrogenase; ETC = electron transport chain; e from PSII = electrons from PSII; e from PSI = electrons from PSI; Fdred = ferredoxin, reduced; Fdox = ferredoxin, oxidized; H₂ = hydrogen gas; H⁺ from PSII = protons from PSII; H⁺ from ATP = protons from ATPase. Not all interactions exist in all PS-PBP species.

In sulfur-deprived *C. reinhardtii*, oxygen gas production under the experimental conditions of [7] (1-L, 6×10^6 cell/mL preparation) is about 1 mmol/h after beginning of sulfur deprivation, and spontaneously ceases ~10 h thereafter. 30 - 50 h after beginning of sulfur deprivation, the algae begins releasing hydrogen at a rate of ~0.17 millimole H_2 /h (1-L, 6×10^6 cell/mL preparation) after beginning of

sulfur deprivation. ~100 h after beginning of sulfur deprivation, hydrogen production ceases. These trajectories provide strong constraints on any model of bio H_2 production by *C. reinhardtii*.

The S-system equations used in this study are shown in Figure 2.

```
// protons from PSII
X2' = a2 X1^g2_1 X3^g2_3 X5^g2_5 - b2 X10^h2_8 X2^h2_2 X5^h2_5

// e from PSII
X4' = a4 X1^g4_1 X3^g4_3 X5^g4_5 - b4 X16^h4_16 X4^h4_4

// protons from ATPase
X8' = a8 X6^g8_6 X7^g8_7 X2^g8_2 - b8 X8^h8_8 X24^h8_24

// other ATP consumers
X9' = a9 X10^g9_10 - b9 X9^h9_9

// ATP
X10' = a10 X2^g10_2 X7^g10_7 X6^g10_6 - b10 X13^h10_13 X9^h10_9 X10^h10_10

// starch
X13' = a13 X12^g13_12 X11^g13_11 X10^g13_10 - b13 X14^h13_14 X15^h13_15 X13^h13_13

// e from starch
X14' = a14 X13^g14_13 - b14 X16^h14_16 X14^h14_14

// pyruvate
X15' = a15 X13^g15_13 - b15 X25^h15_25 X18^h15_18 X17^h15_17 X15^h15_15

// e from ETC
X16' = a16 X14^g16_14 X4^g16_4 - b16 X20^h16_20 X16^h16_16

// formate
X17' = a17 X15^g17_15 - b17 X17^h17_17

// acetate
X18' = a18 X15^g18_15 - b18 X15^h18_25 X18^h18_18

// e from PSI
X20' = a20 X16^g20_16 - b20 X22^h20_22 X20^h20_20

// Fdox
X21' = a21 X22^g21_22 - b21 X21^h21_21

// Fdred
// X22' = a22 X20^g22_20 - b22 X21^g22_21 X23^g22_23 X22^h22_22

// e from Fdred
X23' = a23 X22^g23_22 - b23 X24^h23_24 X23^h23_23

// H2 gas
X24' = a24 X23^g24_23 X8^g24_8 - b24 X24^h24_24

// Intracellular CO2
X25' = a25 X15^g25_15 X18^g25_18 X26^g25_26 - b25 X25^h25_25

// oxygen
X26' = a26 X1^g26_1 X3^g26_3 X5^g26_5 - b26 X26^h26_26 X25^h26_25 X5^h26_5
```

Figure 2. S-system equations for the dependent variables used in this study. “^” is exponentiation. “>>” means “expression continuation”. “’” means “first derivative with respect to time”. Note that the equation for X2' has light as a *consumption* factor because activity *decreases* as light intensity increases above an optimal value.

Table 1 shows the values of the independent variables of the system.

Table 1. Values of the independent variables of the system.

Independent variable	Value (relative units)
X1 (water)	1
X3 (SO ₄)	0.3
X5 (hv-II)	2.363
X6 (ADP)	100
X7 (PO ₄)	100
X11 (Extracellular CO ₂)	3e-3
X12 (NADPH)	1e-6
X19 (hv-I)	2.363

Much of the system in Figure 1 is based on PSII and PSI kinetics. The model was calibrated (to produce the "nominal" configuration) on PSII/PSI kinetic data in [16], setting all generalized rate constants to 0.1, except a₂ (= 3e-4), b₂ (= 1e-4), a₄ (=0.01), a₂₄ (=1e-4), b₂₄ (= 0.001), a₂₆ (=10), and b₂₆ (=1000); these exceptions were based on *in vitro* experimental values obtained in [7]. All generalized kinetic orders were set to 1.

bioh2gen and the model used in [14] differ in a few ways. First, following the conventions in [11] for modeling metabolic systems in the absence of gene-circuit dynamics, no enzyme is an explicit variable of *bioh2gen*; several enzymes are variables in [14]. Second, *bioh2gen* employs more rate constants derived from experiment than does the model used in [14]. Third, all the

kinetic orders in *bioh2gen* were set to 1; two kinetic orders were set to 2 in [14]. Fourth, *bioh2gen* study models the photon inputs to each of PSII and PSI individually; the model in [14] represents only the photon inputs to PSII.

The nominal H₂ and O₂ production rates of *bioh2gen* were compared to [7], and the response of the organism to light intensities ranging from 0.01 - 20.0 hv were computed.

4.0 Results and discussion

Figure 3 show the nominal (hv-I and hv-II = 2.363) hydrogen and oxygen output predicted by the model described in Section 3.0. The H₂ and O₂ outputs agree well with [7].

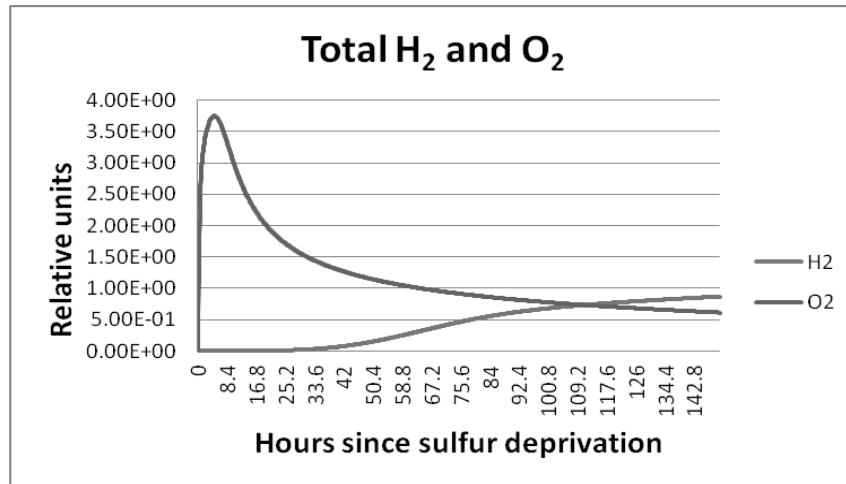


Figure 3. Nominal (for $h\nu = 2.363$) total hydrogen and oxygen gas production as a function of time (units on the horizontal axis are hours after t_0). The values predicted by the model agree well with the results shown in [7].

Figure 4 shows the H₂ gas production in the model as a function of incident light intensity at PSII and PSI.

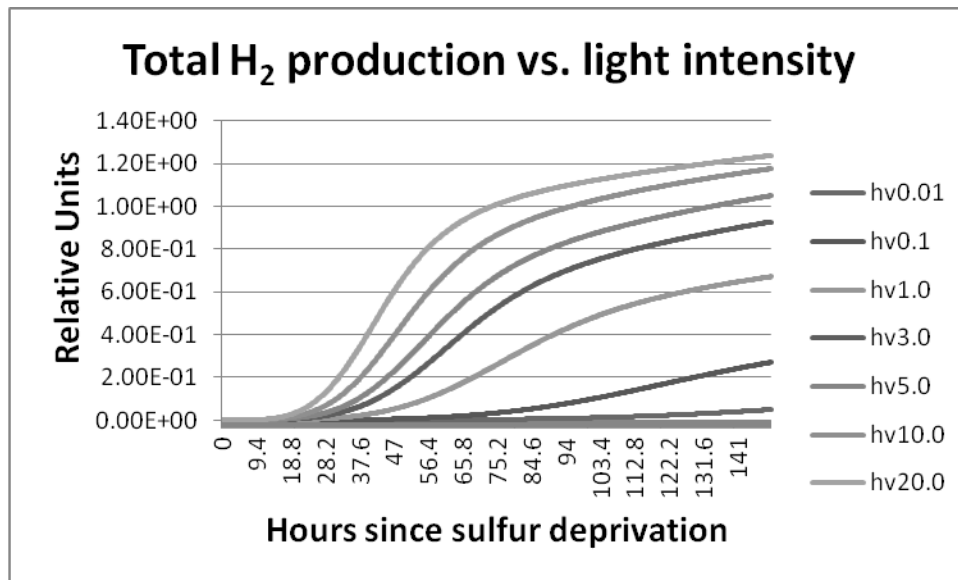


Figure 4. H₂ production as a function of incident light intensity at PSII and PSI. Note the saturation effect as the intensity exceeds ~ 10.0 hv.

Figure 4 strongly suggests that, within the model described in Section 3.0, the H₂ production of *C. reinhardtii* cannot be increased with light intensities $> \sim 10$ hv. These results are generally consistent with the results reported in [17].

5.0 Acknowledgements

This work benefited from discussions with Maria Ghirardi and Michael Seibert of the National Renewable Energy Laboratory, Anastasios Melis of the University of California/Berkeley, Anatoly Tsygankov of the Institute of Basic Biological Problems (Pushchino, Russia), Orlando Jorquera of the Federal University of Bahia, Murray Wolinsky of Los Alamos National Laboratory, and Jorge Soberón of the University of Kansas Biodiversity Institute. For any errors that remain, I am solely responsible.

6.0 References

- [1] Cammack R. Hydrogenases and their activities. In Cammack R, Frey M, and Robson R, eds. *Hydrogen as a Fuel: Learning from Nature*. Taylor and Francis. 2001.
- [2] Ferreira AEN. *Power Law Analysis and Simulation (PLAS)*. Version 1.2 beta, Build 0.120. URL <http://correio.cc.fc.ul.pt/~aenf/plas.html>. March 2011. Note: the link to the PLAS software appears is broken as of 1 January 2012. A copy of the software is available on request from the author of the present paper.
- [3] Horner JK. An S-system model of hydrogen production in microalgae. *International Society for Computational Biology 2002, Special Interest Group for Biological Simulation Satellite Meeting (SIGSIM2002), Computer Modeling of Cellular Processes*. Edmonton, Alberta, Canada.
- [4] Horner JK. Leveraging biohydrogen research: a kinetic modeling approach. *Hydrogen and Fuel Cells Conference 2003*. Vancouver, British Columbia, Canada.
- [5] Horner JK and Wolinsky MA. A power-law sensitivity analysis of the hydrogen-producing metabolic pathway in *Chlamydomonas reinhardtii*. *International Journal of Hydrogen Energy* 27 (2002), 1251-1255.
- [6] Lawlor DW. *Photosynthesis*. Third Edition. Springer. 2001.
- [7] Melis A et al. Sustained photobiological hydrogen gas production upon reversible inactivation of oxygen evolution in the green algae *Chlamydomonas reinhardtii*. *Plant Physiology* 122 (2000), 127-135.
- [8] Melis A. Green alga hydrogen production: progress, problems, and prospects. *International Journal of Hydrogen Energy* 27 (2002), 1217-1228.
- [9] Horner JK. *bioh2gen, Version 1*. Available on request from the author. 2004.
- [10] Savageau MA. Growth of complex systems can be related to the properties of their underlying determinants. *Proceedings of the National Academy of Sciences* 76 (1979), 5413-5417.
- [11] Voit EO. *Computational Analysis of Biochemical Systems*. Cambridge. 2000.
- [12] Drazin PG. *Nonlinear Dynamics*. Cambridge. 1992.
- [13] Markvart T and Landsberg PT. Solar cell model for electron transport in photosynthesis. *Proceedings of the 29th IEEE Photovoltaic Specialists Conference (2002)*, 1348-1351.
- [14] Jorquera O, Kiperstok A, Sales EA, Embirucu M, and Ghiardi ML. S-systems sensitivity analysis of the factors that may influence hydrogen production by sulfur-deprived *Chlamydomonas reinhardtii*. *International Journal of Hydrogen Energy* 33 (2008), 2167-2177.
- [15] Horner JK. *bioh2gen, Version 5*, a PLAS simulator for biohydrogen production by photosynthetic biohydrogen producers. Source code is available on request from the author.
- [16] NPO Bioinformatics Japan. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. <http://www.genome.jp/kegg/>. 2012.
- [17] Polle JEW, Kanakagiri S-D, and Melis A. tla1, a DNA insertional transformant of the green alga *Chlamydomonas reinhardtii* with a truncated light-harvesting chlorophyll antenna size. *Planta* 217 (2003), 49-59.

Taming the Chatroom Bob: The role of brain-computer interfaces that manipulate prefrontal cortex optimization for increasing participation of victims of traumatic sex and other abuse online

J. Bishop

Mathematics, Engineering, Intelligent Systems and Future Technologies Group
Centre for Research into Online Communities and E-Learning Systems
Institute of Life Sciences, Swansea University, Singleton Park, Swansea, SA2 8PP

Abstract – *Chatroom Bobs, which derived from the concept of 'Uncle Bob' being a name for a less than responsible family man, are characterised by being online community users driven by seeking out satisfaction for their 'urgeances' (or biological drives). Some of these are akin to the 'office loser' who tries to impress others but is despised, others have more ulterior motives for sexual satisfaction. This paper presents an intervention – called MEDIAT – which uses TAGTeach to retrain people who are sexually damaged by society and demonstrate impairment in how they interact with others. The paper presents an equation for measuring such 'social orientation impairment' as a reflection of its relationship to serotonergic and dopaminergic activity in the prefrontal cortex as a result of differences in 'Neuro-response plasticity'. The paper concludes that by using MEDIAT to reverse dopaminergic-serotonergic asynchronicity caused by traumatic experience can lead to increased constructive participation in online and other environments.*

Keywords: Personality disorders, social orientation impairments, evolution, human-computer interaction

1 Introduction

The chatroom bob is a prolific character in online communities, characterised by constant references to sex and other desires [1]. They range from the one extreme of dangerous people who want to seduce others to get their way with them in whatever form they want, to more harmless ones who simply post 'rude jokes' or make double entendres. The term 'chatroom bob' was first described in NetLingo as "A nickname girls give to the kind of guy who uses the Internet primarily to hang out in chat rooms and search for photos of naked women. If he finds a pretty girl's Web site, he will send flirty e-mail messages ad nauseum, even though he would 'never in a million years' approach her face-to-face". Chatroom bobs may come from a number of well understood backgrounds, from victims of sex abuse to people who otherwise lack maturity in the way they have psychologically

developed to understand others and have relationships with them. Many chatroom bobs will experience specific social behavior traits (SBTs) which this paper argues differ based on someone's neurological make-up, specifically the degree to which their prefrontal cortex is optimal or not.

A social behavioral trait can restrict an actor's optimal performance in an environment, and this can be seen to be a 'social orientation impairment' (SOI) and also an SBT that enables optimal performance, which can be seen as a 'social orientation advantage' (SOA). It is clear that the extent of these vary between different types of chatroom bob. Those who are able to use their social orientation to seduce others and take advantage of them could be considered to have an SOA. On the other hand, those who are not able to convince others of their worldview, such as through not reading their theory of mind can be seen to have SOIs.

The SOIs that result from those SBTs which have their basis in medical conditions, are either from a physical basis, such as due to traumatic brain injury or genetic mutations [2], or from a mental basis, such as due to childhood sex abuse. For instance it is known that the genes TPH1 and TPH2 associated with the prefrontal cortex are associated with known social impairments, like autism and schizophrenia.

Those SBTs that are derived from medical conditions can be seen to differ from those caused by genetic differences, such as those which affect the sex or race or a person, which may result in differences in gender or cultural identification. The acceptance of a person's social behavioral traits by the environment affect whether these become an SOI or an SOA. To explain this throughout the paper the concept of "Darwin's birds" will be used as well as "Norman's doors". Essentially, Darwin's birds, refers to the concept that a difference in a person's make-up can have huge effects on their chances of survival in an environment. In *The Origin of the Species*, Darwin showed that birds with one type of beak would survive over the others where the food sources were best suited to consumption by those with that type of beak.

Norman's doors on the other hand refers to a concept that humans, when designing their own environments can do so that is disadvantageous to those without a high-level knowledge of that system, as it operates differently to what they would expect. In other words, a Norman door could make a Darwin bird disabled if it did not have the right beak to open that door. A disability in this context is where an actor's SBT becomes an SOI due to the environment imposed on them by other actors who may not share that SBT.

1.1 The role of the pre-frontal cortex in social orientation construction

The pre-frontal cortex is composed of several anatomical regions that are responsible for numerous functions including planning, language production, working memory, artistic expression, some aspects of emotional behavior and attention among others[3]. Neuroimaging studies have provided some of the most consistent evidence that dysfunction of the prefrontal cortex is a characteristic of schizophrenia[4]. The Diagnostic and Statistical Manual for Mental Disorders (DSM) also makes it clear that one should not give a diagnosis for an autism spectrum condition where there are grounds for schizophrenia, which suggests an overlap in symptoms or causation. This leads one to suggest that the difficulties people with schizophrenia have with regards to constructing an accurate interpretation of the situation they are in, may be reflected in people with autism who have difficulty constructing an interpretation at all. Damage to the prefrontal cortex is associated with impaired emotional and social interactions such as angry outbursts, increased lability, interpersonal skills deficits, insensitivity, and sexual disinhibition[5]. Some of these have been found in people with autism and social phobia, and attempts have been made to develop technological interventions that help people with these disabilities overcome them [6, 7].

The pre-frontal cortex is involved in the behavioral inhibitory mechanism and not just participating in the behavioral excitatory mechanism [8] and this is also something known to play a big role in bipolar disorder[9], which may also explain why some persons get less effect at work when they have developed a thinking pattern of discarding any opportunities in the environment. Because the pre frontal cortex is involved in the organization of behavior, abstraction, and consciousness, its disruption could also facilitate violent behavior indirectly by interfering with the individual's perception of the situation [10], which may be why persons with schizophrenia misinterpret others and therefore express inappropriate actions.

1.2 Social behavioral traits

An SBT that manifests itself as an SOI and which is propagated as a disability has serious consequences of an actor's psychopathy, specifically their ability to appreciate and navigate the social and emotional world compared to those not disabled by the environment in which they are. The ability of an actor to form reactions to a particular situation they are in is in part related to the plasticity of their 'neuro-response' functioning [11], which is aided by what is called a 'seduction mechanism'. The seduction mechanism is the main change stimulus an actor responds to that transforms them from one set of mental or physical states to another through influencing their dopaminergic and serotonergic activity. Recent research finds that neurochemical and neurophysiological hyper-reactivity of the dopaminergic reward system may comprise a neural substrate for impulsive-antisocial behavior and substance abuse in psychopathy [12].

Dopaminergic hyper-reactivity has been shown to be a factor in compulsive use of Internet environments [13, 14]. This is known to significantly affect the psychopathy of users, with particular regard to those chatroom bobs with sex addiction [15], as one can see from Table 1.

Table 1. The psychopathy of the Chatroom Bob

Groups	RELATION Distorted Attachment Chatroom Bob	TRANSACTION Adaptable Chatroom Bob	VIOLATION Hyper-Sexualized Chatroom Bob	Hypothesised Dopamine/Serotonin Link
Dimensions				
Previous convictions	No	No	Yes	Reduced serotonergic activity (Yes), increased (No)
Use of identity	Own	Other	Other	Increased dopaminergic activity
Indecent image use	No	No	Yes	Low dopaminergic flow, High serotonergic involvement
Contact other offenders	No	No	Yes	High dopaminergic flow, low serotonergic involvement
Offence-supportive belief	Friendship and love	Exchange compliance	Dehumanised as object	Increased dopaminergic flow
Speed of contact	Long before meeting	Tailored escalation	Fast sex talk and action	Increased dopaminergic flow
Contact method	Personalized contact by phone	Contingent contact approach	Non-personal contact approach	Reduced serotonergic involvement
Contact maintenance	Persistence of caring and love	Offers of help and service	Threats of punishment	High dopaminergic flow, low serotonergic involvement
Offence outcome	All want to meet offline	Some want to meet offline	Some want to meet offline	N/A

This paper will show the role of dopaminergic and serotonergic activity in understanding these people. It had been previously thought in virtual reality disciplines that the terms ‘involvement’ and ‘flow’ were synonymous, but the separation of the two offers huge insights into the role the computer plays as a dopaminergic and serotonergic antagonist and agonist [16]. Flow, in this context, is a term that refers to a dynamic state of arousal that characterizes consciousness when experience is attended to for its own sake [17] and the higher the flow the higher the loss of consciousness. It would therefore be appropriate to link this to increased dopaminergic activity.

Decision-making in such a state becomes more fluid and actors respond almost without thought for the consequences of their actions. This has advantages, particularly in chat rooms, where constructive conversations can flourish and people can have a sense of self-worth and feel their contributions are welcomed. Equally, in a state of flow, those users who have an anti-social disposition, known as Snerts [1], will not see the consequences of them posting offensive messages, known as flames, and in particular the effect this will have on deterring lurkers from becoming posters. Involvement on the other hand, in this new context, refers to the amount of effort one has to put into a task for it to have the desired effect. It is proposed in this paper that involvement is linked to serotonergic activity and flow is linked to dopaminergic activity. This could create a view of SOIs, as persons with inopportune neuro-response plasticity variation (INRPV), which results in serotonergic-dopaminergic asynchronicity (SDA) in the worst of occasions.

1.3 The role of serotonergic-dopaminergic asynchronicity in influencing social orientation

Serotonergic-dopaminergic asynchronicity (SDA) can be seen to manifest in situations where the neuro-response plasticity is high at the same time as serotonergic activity and dopaminergic activity being high. This can be caused in a particular situation an actor has constructed, such as having obsessive thinking about inviting a person one is attracted to online to meet in the real world resulting in increased dopamine levels. Coupled with the anxiety due to apprehension this increases serotonin levels while at the same time the increase attention focus driving up neuro-response plasticity for rapid thinking and responding. A sudden rejection by the now unrequited love, might throw the person into turmoil due to their increased serotonergic responses to the resulting anxiety. In order to deal with the ‘rejection’ the actor tries to get that person out of their mind, which while driving down dopaminergic activity does not at the same time drive down neuro-response plasticity and therefore results in serotonergic-dopaminergic asynchronicity. One can see in this case, that the actor thought they had an optimal Darwin beak, being a suitable companion for their ideal mate, but when that person revealed themselves to be a Norman door, that was not suited to their Darwin beak, then that led them to believe their beak was not suitable for pecking at their preferred doors,

which was what they thought was the be-all-and-end-all of life, even if mistakenly.

It could be argued that such a traumatic series of events creates a mental block, lodged in their prefrontal cortex, which is referred to as a phantasy. This seriously affects their pre-disposition to specific social behavioral traits, which if ego-dystonic will result in a self-constructed social orientation impairment. The actor would normally seek to avoid an action yet they have a drive to do it, and in other situations where they would normally seek to get involved for gratification, meaning they would experience severe dissonance when attempting to engage. This in turn causes severe discomfort. Both of these can involve phantasies that restrict the optimal synaptic flow of connections to the prefrontal cortex, and can thus result in the impairments common to SOIs. These conditions include an impaired ability to form appropriate responses in social and emotional activities, due to lack of utilization of the blocked neural pathways.

Table 2 Rules of calculating Neuro-response plasticity Productivity (knol)

Rule	Example
If dopaminergic flow increases and serotonergic involvement remains unchanged, then it leads to higher knol and humanpower.	A person with an SOI is highly involved in applying emotion recognition training. They become so involved in the task this acts as a seduction mechanism to lose track of time.
If dopaminergic flow decreases and serotonergic involvement remains unchanged, then it leads to lower knol and humanpower.	A person with an SOI is highly involved in applying emotion recognition training. The difficulty, acts as a seduction mechanism to reduce the amount of time they can spend on the task.
If serotonergic involvement increases and dopaminergic flow remains unchanged, then it leads to lower pression and higher humanpower.	An SOI is really interested in an activity to the exclusion of others and then something disrupts that concentration, acting as a seduction mechanism to increase the time they spend on the task, decreasing efficiency.

The following equation shows how to identify the value of a particular phantasy (on a scale of -5 to 5) affecting neuro-response plasticity using the values of the cognitions specific to the particular individual whose social orientation is being constructed.

$$p_i = \left(\frac{((x + x_1) * (y + y_1) - \bar{z}))}{c} \right)$$

Equation 1 Calculating a Phantasy

In this context x refers to the object cognition and y the subject cognition. The element x1 refers to a number that is added to the object variable to aid conversion into a phantasy. Equally the element y1 refers to the addition to the subject to aid conversion, and the z̄ refers to a value that is required to shift the new variable, which once divided by the centrepoint

Table 3 Phantasy construction from interaction between pre-frontal cortex detachments/interests and other cognitions

x cognition	x ₁	y cognition	y ₁	\bar{z}	C	Pre-frontal cortex function
Detachment	3	Goal	3	36	-9.6	Problem-solving
Detachment	0	Plan	0	0	-2.4	Self-control
Detachment	0	Value	0	12	-2.4	Conscience
Detachment	0	Belief	0	0	-3.6	Working Memory
Detachment	0	Interest	0	30	-6	Empathy
Detachment	0	Detachment	0	18	-3.6	Deception
Interest	0	Goal	0	45.5	8.9	Problem-solving
Interest	0	Plan	3	1.5	4.3	Self-control
Interest	0	Value	0	20.5	3.9	Conscience
Interest	0	Belief	4	2	6.4	Working Memory
Interest	0	Interest	0	50.5	9.9	Empathy
Interest	0	Detachment	0	30	-6	Deception

(c) then fits onto a -5 to +5 scale as a phantasy. In the case of the previous example one could use the cognitions of 'Interest' (i.e. 1 to 10) and 'Detachment' (i.e. 0 to 6). The actor that suffered from unrequited love, had an interest (x) in the other person of 10, because they were very much attracted to them. They also had a detachment of 0 (y), because they were essential to their future in their minds. With the appropriate x₁ being 0 and y₁ being 0, \bar{z} being 30 and c being -6, this results in the phantasy being 5, which is the strongest. This is the equivalent of an optimal Darwin beak being able to peck at a sub-optimal Norman door without any dissonance or other disruption. Following the 'rejection' which kept the optimal Darwin beak whilst forcing an undesirable optimal Norman door, the actor went into an unwanted dilemma. This resulted in a sudden shift in the once desirable phantasy to a state that was undesirable. The actor's interest in the other person remained at 10, yet their detachment decreased to 0, resulting in the phantasy becoming -5.

1.4 The impact of phantasies on neuro-response plasticity

In this study, neuro-response plasticity is measured through what is called 'Pression' (P). This come from the French word for pressure, and is a common word in French personal injury law, particularly with regard to psychiatric injury, so it is suitable to be used here. Equation 2 shows how to calculate a

Pression by factoring a number of phantasies (p_i) where 'i' is the identifier for the phantasy and n is the number of phantasies. Force (F) reflects the maximum number of hours of working time someone should produce in order to maintain a healthy amount of productivity, which is 48, based on the European Working Time Directive.

$$P = \left(\left(\sum_{i=1}^n p_i \right) / 5 \right) + F$$

Equation 2 Calculating a Pression (Pressure)

Equation 2 shows how to calculate a knol, which is the unit used to represent the serotonergic-dopaminergic synchronicity (usually between 0 and 1, but minus infinity and plus infinity are possible). The symbol H stands for humanpower, which is the individual's weekly Neuro-response plasticity potential, calculated by squaring the potential force (F) of 48 hours and subtracting the baseline (B) of actual working time (i.e. 37 hours for most full-time workers in Great Britain) from it. Also, n is the number of phantasies and i is the identifier of the phantasy in question.

2 Measuring Neuro-response plasticity in the prefrontal cortex

It has long been known that past memories can affect the ways in which one interacts with others and the environment, though these need not always be traumatic [18]. These 'phantasies' in the prefrontal cortex, including the cognitions are called, 'detachments'. It is proposed that this detachment cognition place an important role in understanding the emotions that create unwanted behavior and thoughts in those with social orientation impairments. Someone with significant INRPV will have significant more detachments in terms of quantity and strength, which impairs access from the rest of the brain to the functions in the prefrontal cortex where they are located.

Table 2 shows a number of phantasies, based on the emotional category 'bothered'[19], which reflect the hypothesised interaction between the pre-frontal cortex functions identified in the literature and the other cognitions identified in the ecological cognition research [11, 20].

$$k = \frac{P}{H}$$

Equation 3 Calculating a knol

Figure 1 shows a representation of Equation 3 in relation to two persons. Once a person with an SOI represented in blue who can only work 16 hours per week (H1), and the other a person with a SOA who can work 37 hours per week (H2),

represented in green. The purple lines represent serotonergic involvement (I), which is the amount of task-focussed anxiety that someone has to go through in order for them personally to achieve a particular goal. The red and yellow lines represent the amount of dopaminergic flow (F1) an actor has when performing that task, which reflects the ease at which it can be performed. The startpoint (y) for a dopaminergic flow measurement is a Pression of 96, which is 48 hours force (F) squared minus the baseline (B) and then multiplied by a knol of 1. The endpoint (x) for measuring knol is where that line meets the cross between an actor's Pression value and their Humanpower. In the case of dopaminergic flow, the startpoint (y) for calculating dopaminergic flow is 0 and the endpoint (x) is the same as their Humanpower (H).

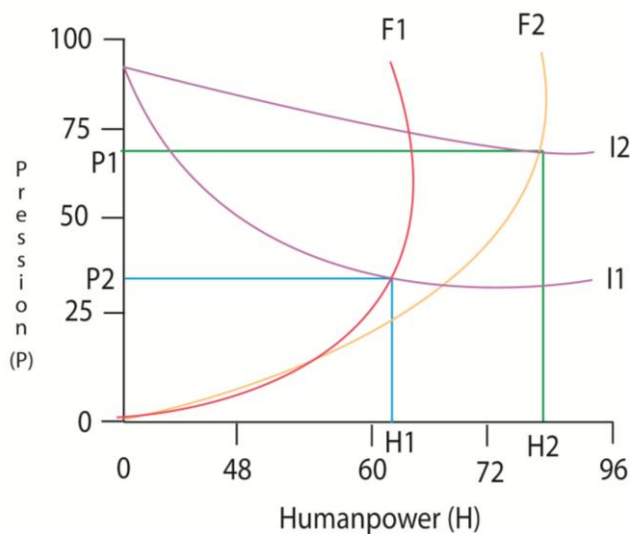


Figure 1 Calculating Neuro-response plasticity Productivity (knol)

In the case of the example in Figure 1, the role of calculating a knol in understanding neuro-response plasticity (NRP) shall be explained. The SOA actor has a Humanpower (H1) of 64 (16+48) and the SOI actor a Humanpower (H2) of 88 (37+48). An example of an actor with an SOI, who will have at least one phantasy will be used. This phantasy is valued at -5, representing a sub-optimal Darwin beak pecking at an optimal Norman door and an actor with an SOA who has at least one phantasy, valued at 5, representing a sub-optimal Darwin beak, pecking at an optimal Norman door is the example to be used. This would give the SOI a Pression of 47 and the SOA actor a Pression of 49. When the SOA actor's Pression (P1) is divided by their Horsepower (H1), this gives them a knol of 0.73.

2.1 An intervention and cause

The concept of measuring traumatic phantasies in the prefrontal cortex could lead to huge leaps in humanitarian well-being and social justice. For instance, the concept of 'psychiatric shock', such as that experienced by the author during the sex abuse in childhood, could take on a new

dimension beyond that discussed in cases like *Alcock v Chief Constable of South Yorkshire Police* [1992] 1 AC 310. With the advancement of Internet abuse through misuse of 'trolling' [21] to harm others from the safety of their personal computers, then there is going to need to be a way of measuring psychological trauma to determine the extent of psychiatric injury, in the same way one can with physical abuse. This could mean those families denied justice in the aforementioned case because they observed the indescribable injury to a loved one on television, should have strong evidence of the psychiatric shock they sustained so the judgement can be overturned.

Many of the social relation theoretic principles that apply to online dating between adults [21], also apply in relation to attraction to minors by online sex offenders seeking gratification, generically known as paedophilic chatroom bobs [1]. Using the findings above in relation to the dopaminergic and serotonergic activity involved in coming to terms with a formerly dystonic sexual identity can offer insights into how this can facilitate undesirable activity in paedophilic chatroom bobs which reduces the positive participation of minors in online environments. Table 5 presents links between the three types of grooming-orientated paedophilic chatroom bob identified by Gottschal [22] with the findings above.

3 Towards the Mediated Emotion Demonstration Intervention using Avatars and TAGTeach

Using a number of patents [23-25] to create an iterative prototypical intervention called the 'The Mediated Emotion Demonstration Intervention using Avatars and TAGTeach' (MEDIAT) is proposed. The intervention, which developed throughout the study, can provide a simulated environment with visual feedback, very similar to that used elsewhere [26], in order to display an output onto a computer screen which shows neuro-response plasticity and dopaminergic and serotonergic synchronicity. Beyond the scope of the original patent, it was possible for users to track their thought processes and then use a metronome-like tool, such as a clicking pen, to indicate when they have achieved the desired mental state, so it is easier to gain that mental state when absent from the brain-scanner, just by clicking the pen for instance.

3.1 Role of MEDIAT for re-training online sex predators

The links between the dimensions and paedophilic chatroom bob activity and dopaminergic and serotonergic activity can be generally seen to affect the conscience of these groomers. A higher dopaminergic flow and lower serotonergic involvement will make the paedophilic chatroom bob feel less guilty and more motivated to engage in their activities, and thus make them have a stronger advantageous 'Darwin beak' and make their victims less like a Norman door. A reduced dopaminergic flow and increased serotonergic involvement

will manifest itself when a sex offender is committing an inappropriate sexual act, and this will normally be connected with a high level of neuro-response plasticity (i.e. attention focus). This suggests that any form of therapy for turning around paedophilic chatroom bobs to make them see minors as Norman doors may need to look at interventions that reduce dopaminergic-serotonergic asynchronicity so that when the individual has a phantasy that consists low dopaminergic flow and high serotonergic involvement that their associated neuro-response plasticity is low. Equally it may be appropriate to develop appropriate phantasies so that the paedophilic chatroom bob's Darwin's beak changes to one more suitable for not identifying with Norman doors. Such phantasies would involve the individual having high dopaminergic flow when thinking about an 'appropriate' sexual partner, as well as low serotonergic involvement. When this is paired at the same time with high neuro-response plasticity then this could redirect the drive of paedophilic chatroom bobs that are interested in minors, to being interested in age-appropriate users of online environments.

4 Discussion

This paper has provided strong preliminary evidence of the role of the prefrontal cortex in the manifestation of social orientation impairments, such as autism. Further research, such as through more extensive brain imaging is needed to validate the assumptions made in the study, such as with regards to the links between pre-frontal cortex functions and memories carrying the attributed emotions.

The link between abnormal sex-related thoughts or traumas and increased Neuro-Response Plasticity, increased serotonergic involvement, reduced dopaminergic flow and the dopaminergic-serotonergic asynchronicity (DSA) opens up opportunities for exploring treatment opportunities to people with sexual identity disorders as suggested in the previous section. Studies could look at whether interventions that reduce DSA by reducing attention focus (i.e. NRP) on undesirable thoughts at the same time as reducing the dopaminergic while maintaining the high serotonergic involvement can reverse someone's propensity to criminal manifestations of abnormal sexual thinking, such as paedophilia, rape, among others.

In terms of advancing diagnosis and measurement of phantasies associated with these abnormal thought processes, social psychological studies can identify whether the use of subjective quantitative instruments such a Q-methodology can be used to identify common phantasies among different social groups, particularly as the scale for phantasies proposed above is -5 to 5, as it typical of these studies.

5 References

- [1] J. Bishop. "Increasing Capital Revenue in Social Networking Communities: Building Social and Economic Relationships through Avatars and Characters"; *Social Computing: Concepts, Methodologies, Tools, and Applications* (IGI Global) S. Dasgupta (Ed.), 1987-2004:2009.
- [2] C. S. Leblond, J. Heinrich, R. Delorme, C. Proepper, C. Betancur, G. Huguet, M. Konyukh, P. Chaste, E. Ey, M. Rastam, H. Anckarsäter, G. Nygren, C. Gillberg, J. Melke, R. Toro, B. Regnault, F. Fauchereau, O. Mercati, N. Lemièrè, D. Skuse, M. Poot, R. Holt, A. P. Monaco, I. Järvelä, K. Kantojärvi, R. Vanhala, S. Curran, A. Collier, P. Bolton, A. Chiocchetti, S. M. Klauck, F. Poustka, C. M. Freitag, R. Waltes, M. Kopp, E. Duketis, E. Bacchelli, F. Minopoli, L. Ruta, A. Battaglia, L. Mazzone, E. Maestrini, A. F. Sequeira, B. Oliveira, A. Vicente, G. Oliveira2, A. Pinto, S. W. Scherer, D. Zelenika, M. Delepine, M. Lathrop, D. Bonneau, V. Guinchat, F. Devillard, B. Assouline, M. C. Mouren, M. Leboyer, C. Gillberg, T. M. Boeckers & T. Bourgeron. "Genetic and Functional Analyses of SHANK2 Mutations Suggest a Multiple Hit Model of Autism Spectrum Disorders"; *PLoS Genetics*, 8., 2, 2012.
- [3] E. Miller & L. Buys. "Is Generation X the new Civic Generation? An exploratory analysis of social capital, environmental attitudes and behaviours in an Australian community". Paper presented to the Social Change in the 21st Century Conference, Centre for Social Change Research, Queensland University of Technology. Queensland University of Technology, 2004. .
- [4] K. B. Yancey. "2008 NCTE Presidential Address: The Impulse to Compose and the Age of Composition"; *Research in the Teaching of English*, 43., 3, 2008, 2009.
- [5] D. Pekarsky. "Excellence in Teaching—Here Too, it Takes a Village"; *Journal of Jewish Education*, 75., 3, 203-215, 2009.
- [6] J. Bishop. "The Role of Augmented E-Learning Systems for Enhancing Pro-social Behaviour in Socially Impaired Individuals"; *Assistive and Augmentive Communication for the Disabled: Intelligent Technologies for Communication, Learning and Teaching* (IGI Global) B-T Lau (Ed.), 2011.
- [7] J. Bishop. "The Internet for educating individuals with social impairments"; *Journal of Computer Assisted Learning*, 19., 4, 546-556, 2003.
- [8] M. Watanabe. "Prefrontal unit activity during delayed conditional Go/No-Go discrimination in the monkey. II. Relation to Go and No-Go responses"; *Brain research*, 382., 1, 15-27, 1986.
- [9] D. R. Hirshfeld-Becker, J. Biederman, S. Calltharp, E. D. Rosenbaum, S. V. Faraone & J. F. Rosenbaum. "Behavioral inhibition and disinhibition as hypothesized precursors to psychopathology"; *Biological psychiatry*, 53., 11, 985-999, 2003.

- [10] N. D. Volkow, L. R. Tancredib, C. Grant, H. Gillespie, A. Valentine, N. Mullani, G. J. Wang & L. Hollister. "Brain glucose metabolism in violent psychiatric patients: a preliminary study"; *Psychiatry Research: Neuroimaging*, 61., 4, 243-253, 1995.
- [11] J. Bishop. "Ecological Cognition: A New Dynamic for Human-Computer Interaction"; *The Mind, the Body and the World: Psychology after Cognitivism* (Imprint Academic) B. Wallace, A. Ross, J. Davies & T. Anderson (Eds.), 327-3452007.
- [12] J. W. Buckholtz, M. T. Treadway, R. L. Cowan, N. D. Woodward, S. D. Benning, R. Li, M. S. Ansari, R. M. Baldwin, A. N. Schwartzman & E. S. Shelby. "Mesolimbic dopamine reward system hypersensitivity in individuals with psychopathic traits"; *Nature neuroscience*, 13., 4, 419-421, 2010.
- [13] D. H. Han, Y. S. Lee, K. C. Yang, E. Y. Kim, I. K. Lyoo & P. F. Renshaw. "Dopamine genes and reward dependence in adolescents with excessive internet video game play"; *Journal of Addiction Medicine*, 1., 3, 133, 2007.
- [14] B. D. Ng & P. Wiemer-Hastings. "Addiction to the internet and online gaming"; *CyberPsychology & Behavior*, 8., 2, 110-113, 2005.
- [15] L. E. Marshall, M. D. O'Brien, A. R. Beech, L. A. Craig AND K. D. Browne. "Assessment of sexual addiction"; *Assessment and treatment of sex offenders: A handbook* (Wiley) A. R. Beech, L. A. Craig & K. D. Browne (Eds.), 1632009.
- [16] K. Gillan & J. Pickerill. "Transnational anti-war activism: Solidarity, diversity and the Internet in Australia, Britain and the United states after 9/11"; *Australian Journal of Political Science*, 43., 1, 59-78, 2008.
- [17] M. Csikszentmihalyi. "Flow: the psychology of optimal experience". New York: Harper & Row, 1990.
- [18] S. M. Valente. "Evaluating and Managing Adult PTSD in Primary Care"; *The Nurse practitioner*, 35., 11, 41, 2010.
- [19] O. Golan & S. Baron-Cohen. "Systemizing Emotions: Using Interactive Multimedia as a Teaching Tool"; *Learners on the Autism Spectrum: Preparing Highly Qualified Educators* (Autism Asperger Publishing Company) K. D. Buron (Ed.), 235-2542008.
- [20] J. Bishop. "Increasing participation in online communities: A framework for human-computer interaction"; *Computers in Human Behavior*, 23., 4, 1881-1893, 2007.
- [21] J. Bishop. "Increasing Capital Revenue in Social Networking Communities: Building Social and Economic Relationships through Avatars and Characters"; *Social Networking Communities and eDating Services: Concepts and Implications* (IGI Global) C. Romm-Livermore & K. Setzekorn (Eds.), 2008.
- [22] P. Gottschalk. "A Dark Side of Computing and Information Sciences: Characteristics of Online Groomers"; *Journal of Emerging Trends in Computing and Information Sciences*, 2., 9, 2011.
- [23] A. Junker AND C. R. Berg. "Brain-body actuated system". 09/857,660, 2003. .
- [24] A. Junker. "Brain-body actuated system". 1997. .
- [25] J. Bishop. "Assisting Human Interaction". PCT/GB2011/050814, 2011. .
- [26] E. Hazelkorn. "International Comparisons: The Good, the Bad and the Ugly"; *Other resources*, 24, 2010.

Research on BRATUMSS System of Detecting Transmission Model and Breast Tissues Target Spectrum Distribution

Zhifu Tao¹, Zhonglin Han², Meng Yao^{2,3*}, Yizhou Yao²
Blair Fleet³, Erik D. Goodman³, Huiyan Wang⁴ and John R. Deller⁴

¹Department of Electronic Information Engineering, Suzhou Vocational University, Suzhou China

²Institute of information science and technology East China Normal University, Shanghai China

³BEACON Michigan State University, East Lansing, MI

⁴ECE Michigan State University, East Lansing, MI

*Corresponding Author, e-mail: myao@ee.ecnu.edu.cn

Abstract - BRATUMASS system uses the difference on dielectric constant between breast cancer tissues and normal breast tissues from target tissues microwave response back scatter echo to screen the various tissues. The characteristics of detection object can be determined through of analysis of characteristic of echo. The paper also forwards the transmission loss model depending on the characteristics of the system detection data, and gives fitting analysis results of in vivo real data.

Keywords: BRATUMASS, Transmission Loss, Echo Coefficient, Dielectric Constant

1 Introduction

The environment of microwave propagation determines the transmission loss. For the environment is too complex, people usually summarize empirical model in different environments based on testing data when establishing microwave propagation prediction models. BRATUMASS system makes use of the difference on dielectric constant between breast cancer tissues and normal breast tissues which obtain object tissues back scatter echo through of microwave irradiation. By analyzing echo characteristics, and thus determine the system of characteristics of detecting targets. Since microwave signal power launched by the BRATUMASS system is about 6mW which is received by receiving antenna after transmission in sounding target space. In a relatively longer distance, attenuation becomes the main factor of sampling data quality. This paper is to solve the estimation problem of signal transmission loss of BRATUMASS system and to identify region of tissues echo in power spectrum and provide useful reference for the separation of echo. BRATUMASS system transmission loss model is proposed in analysis of in vivo real data and gives the corresponding valuation formula.

2 BRATUMASS System Transmission Loss Estimation Model

Figure 1 is BRATUMASS system experimental model and antenna structure at present. Zheng, S[1] and others have simulated the electric field distribution within the breast in the electromagnetic field of 1.5GHz and prove that there is a huge difference of the electric field distribution between the breast model of uniform tissue distribution and malignant breast tissues. The transmission of electromagnetic wave within the breast is attributed to near-field problems. Considering the vicinity region of transmission antenna, all possible transmission patterns have been encouraged and there are many high-order modes each of which has its own particular transmission direction and the transmission path. In theory, mode number is infinite, so fast fading phenomenon is very obvious. After a propagation distance high-order mode is almost faded out, then enter the transfer mode mainly based on the base band transmission in which attenuation has become slow down obviously.

BRATUMASS system adopts FM microwave, and the center frequency is 1.5GHz. The propagation velocity of microwave in medium of the detection region (breast tissues)

$$v = 0.766 \times 10^8 \text{ m/s} \quad [1] \quad \text{and} \quad \text{wavelength is}$$

$$\lambda = \frac{v}{f} = \frac{0.766 \times 10^8 \text{ m/s}}{1.5 \times 10^9 \text{ Hz}} = 0.0510 \text{ m} = 5.1 \text{ cm}$$

In the BRATUMASS detection environment, detection target is located in near-field where guided propagation has not been established and propagation of electromagnetic wave here is the multimode, which is similar to propagation mode in free space, so electromagnetic wave mode in free space can be used for prediction. In free space we have

$$10 \lg \frac{P_r}{P_t} = 10 \lg \left[\frac{G_t G_r \lambda^2}{(4\pi)^2 d^2} \right] \quad (1)$$

Where, P_t is the transmission power, P_r is the power of receiver, G_t, G_r is seperately gain of transmitting antenna and receiving antenna, d is line-of-sight distance between transmitting antenna and receiving antenna, unit is m, λ is electromagnetic wavelength, f is electromagnetic frequency, here adopt 1.5GHz.

The microwave propagation speed of breast tissues detected by BRATUMASS system is $v = 0.766 \times 10^8 \text{ m/s}$ Suppose gain of transmitting antenna and receiving antenna are both 1, and then formula (1) can be given as:

$$10 \lg \frac{P_r}{P_t} = 10 \lg \left[\frac{G_t G_r \lambda^2}{(4\pi)^2 d^2} \right] = 20 \lg (0.766 \times 10^8) - 20 \lg 4\pi - 20 \lg fd \quad (2)$$

Based on the above data, we can obtain the following formula

$$10 \lg \frac{P_r}{P_t} = -20 \lg d - 47.8214 \quad (3)$$

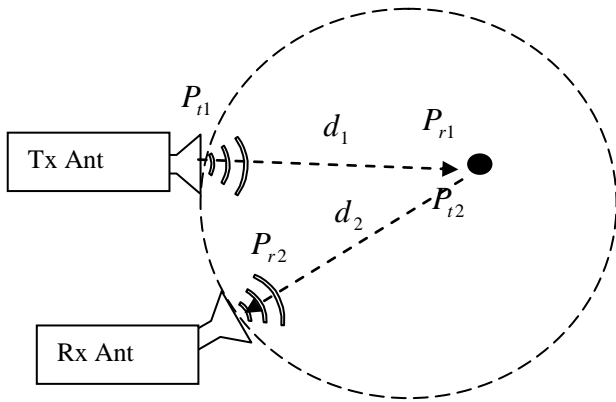


Figure1 BRATUMASS System Sketch Map

The following is the calculation of compensating relationship.

As seen in figure 1, suppose transmission power at transmitting antenna is P_{t1} , the power of electromagnetic wave reaching on target surface is P_{r1} , the distance between transmitting antenna and target is d_1 , 2-order (echo power) transmission power on target surface aroused by electromagnetic wave is P_{t2} , the power at receiving antenna from echo is P_{r2} , the distance between target and receiving antenna is d_2 . Considered separately from transmitting antenna and receiving antenna, we can obtain.

$$\frac{P_{r2}}{P_{t1}} = \frac{P_{r1}}{P_{t1}} \times \frac{P_{r2}}{P_{t2}} \times \frac{P_{t2}}{P_{r1}} \quad (4)$$

And $\frac{P_{t2}}{P_{r1}}$ is echo power ratio of target, and that is $\frac{P_{t2}}{P_{r1}} = \eta_\epsilon$, the logarithm of (4) is:

$$10 \lg \frac{P_{r2}}{P_{t1}} = 10 \lg \frac{P_{r1}}{P_{t1}} + 10 \lg \frac{P_{r2}}{P_{t2}} + 10 \lg \frac{P_{t2}}{P_{r1}} \quad (5)$$

Substitute (3) into (5)

$$10 \lg \frac{P_{r2}}{P_{t1}} = 10 \lg \frac{P_{r1}}{P_{t1}} + 10 \lg \frac{P_{r2}}{P_{t2}} + 10 \lg \eta_\epsilon = (-20 \lg d_1 - 47.8214) + (-20 \lg d_2 - 47.8214) + 10 \lg \eta_\epsilon$$

If $d_1 \approx d_2 = d$ then formula above can be simplified as follows:

$$10 \lg \frac{P_{r2}}{P_{t1}} = 2 \times (-20 \lg d - 47.8214) + 10 \lg \eta_\epsilon \quad (6)$$

Considering (6) is the estimation value of object in homogeneous medium, secondary emission signal power of target aroused by microwave is weaker than that of microwave antenna ignoring 2-order dispersion and other factors, relationship between echo signal intensity in detection space of BRATUMASS system and distance is:

$$10 \lg \frac{P_r}{P_t} = (-20 \lg d - 47.8214) + 10 \lg \eta_\epsilon \quad (7)$$

3 Simulation Results of BRATUMASS System

In order to validate the above estimation model, this paper adjusts the sampling of system as follows: Place BRATUMASS system combined antenna (shown in figure2d) at the triangle mark position shown in figure2a; Place sheet metal of which diameter is 1cm at the ellipse mark position

(and the metal $\eta_\epsilon \approx 1$). The size of combined antenna and the position of object space are shown in figure2b and Figure2c. The base circle size of breast is different depending on different objects. Corresponding data of 14 objects are got. [2]

Figure3 and Figure 4 show two examples of the curve relationship between responding echo of sheet metal and loss estimation curve. And the distance between the sheet metal and detection antenna are separately 175mm and 122.5mm the (3) curve is attenuation value of estimation by formula (3) for metal material. The (6) curve is the attenuation value of estimation by formula (6).

As shown in Figure 3 and Figure 4, signals of (distance) 100mm-200mm received by BRATUMASS system receiving antenna basically located in the estimation region of (3) and (6). And echo of sheet metal is fairly close to the (3) curve of (3). So, it is basically reasonable to use transmission model in free space to estimate attenuation relationship.

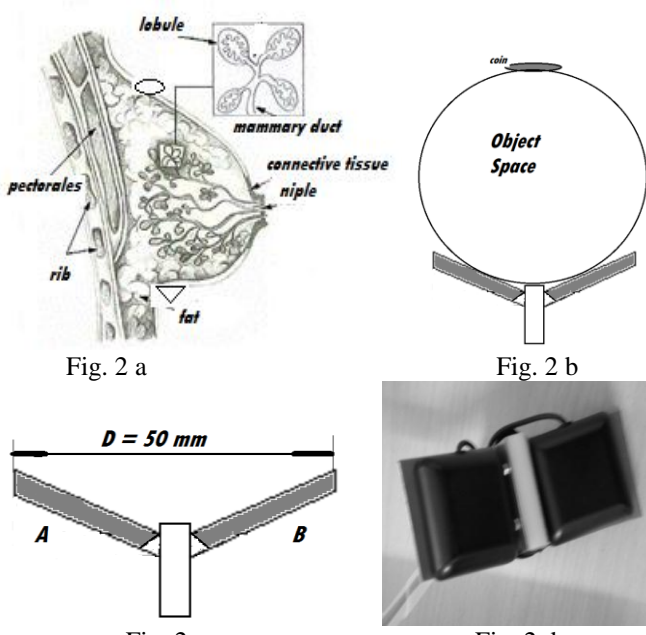


Fig. 2 a

Fig. 2 b

Fig. 2 c

Fig. 2 d

Figure 2. Experimental antenna and position of sampling sketch map in target space. 2a structure of breast, triangle mark point is the position of real data sampling and ellipse mark point is position of sheet metal. 2b. BRATUMASS system testing position sketch map of which red point is the position of real data sampling and green point is position of sheet metal. 2C. Sketch map of parts of combined antenna, A is transmitting antenna. B is receiving antenna. 2d. Combined antenna

Figure 5 shows sheet metal echo spectrum of 14 objects detected by formula (5) accords with the estimation of formula (3). Of which the mark ∇ is the peak value distribution of metal piece echo spectrum.

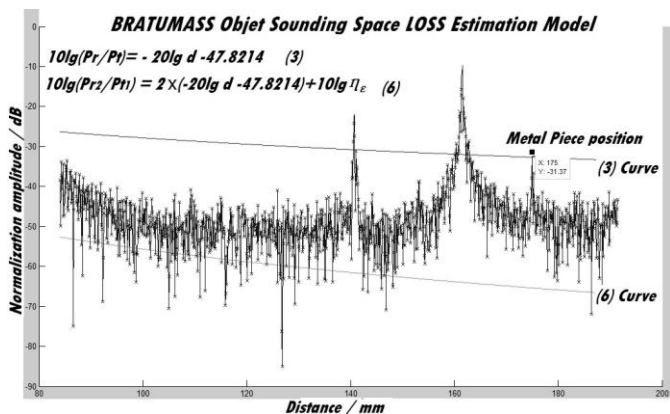


Figure 3. Echo Power Spectrum when distance between metal piece and detection antenna is 175mm

As seen from the figure 5, the estimation value is close to the distribution of real echo. The echo peak value distribution of sheet metal is close to the estimation of formula (3), which accords with the distribution of $\eta_{\epsilon} \approx 1$.

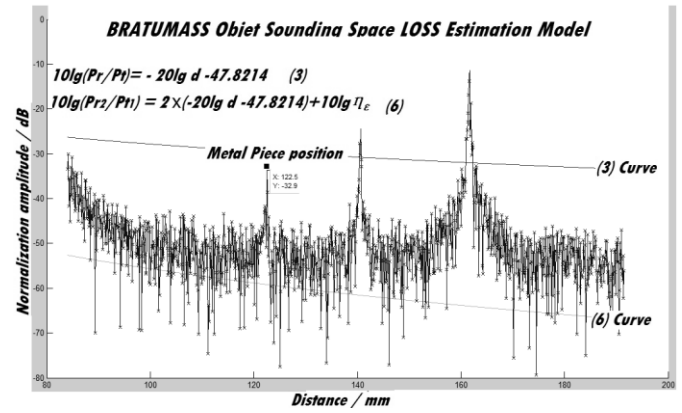


Figure 4. Echo Power Spectrum when distance between metal piece and detection antenna is 122.5mm

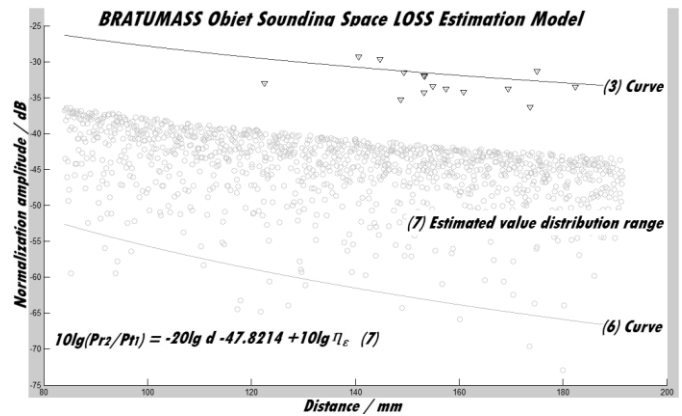


Figure 5. Sheet metal echo spectrum of 14 objects detected by (5) accords with Estimation of (3), of which the mark ∇ is the peak value distribution of metal piece echo spectrum

4 Conclusions

According to the theory of near-field transmission, this paper gives the estimation of microwave transmission loss in target space of BRATUMASS system. The echo of sheet metal is close to the ideal value distribution of $\eta_{\epsilon} = 1$ based on the real testing data. Echo of real breast tissues is close to simulation distribution when dielectric constant is between 10.04-14.93. The dielectric constant of normal breast tissue is between 10 and 15 under frequency of 1.5GHz [3], and sheet metal can't be inserted into living breast for detection because of current conditions. The attenuation can only be estimated by measuring microwave intensity outside. Mode data accords well at distance of beyond 2λ .

The destination of BRATUMASS system is to confirm the distribution information within tissue of breast object. This paper only involves the region location of useful information of echo in the power spectrum, and details of isolation technical of useful information will continue to be discussed in the following articles.

5 Acknowledgement

This work has been performed while Prof. M. Yao was a visiting scholar in Michigan State University, thanks to a visiting research program from Prof. Erik D. Goodman. M. Yao would also like to acknowledge the support of Shanghai Science and Technology Development Foundation under the project grant numbers 03JC14026 and 08JC1409200, as well as the support of TI Co. Ltd through TI (China) Innovation Foundation. And this work is in part supported by National Science Foundation of China grant number 61002003.

6 References

- [1] Zheng, S. "Breast Tumor Imaging Method Investigation in UWB near-field microwave environment". Master Thesis, East China Normal University, 1-35 May 2006.
- [2] Zhongling Han et al, "Application of quarter Iteration of FRFT in BRATUMASS for Weak Signal Extraction". Proceedings of the 2010 International conference on Bioinformatics and Computational Biology July 2010
- [3] Zhifu Tao, Qifeng Pan, Meng Yao, and Ming Li. "Reconstructing Microwave Near-Field Image Based on the Discrepancy of Radial Distribution of Dielectric Constant". ICCSA 2009, 717-728 Oct 2009

Open-source Dental Laboratories Web Content Management Systems vs. Commercial Web Content Management Systems

Reham Alabduljabbar and Samir El-Masri

Information Systems Department College of Computer and Information Sciences
King Saud University Riyadh, Kingdom of Saudi Arabia

Abstract - Desktop applications for managing dental laboratories operation are available; however, these applications are not sufficient. Many information need to be shared between dentists, assistant, laboratory technicians and others. With the development of the Internet technology, there is a need for an effective web application to manage such operation and to control accessing this shared information. The Dental Laboratory WCMS would be used as a channel for dentists to technician, dental clinic-to-dental laboratory, to provide the long-distance service. With dental laboratory WCMS, the medical data, pictures and patient records all can be accessible online. Dental laboratories can track and manage lab cases and payments online. Besides, dental clinics and dentists can access to track the lab cases, and they can be notified by email about the status of their lab cases. Both have an account and can view their lab case history, their current balance, and pay bills online. It aids in reducing paperwork and automating the approach to process lab cases. Any dental laboratory can have its own instance of the system to manage its content and to ease communications with dental clinics its working with. The system adopts three layers technical architecture to design the system as in [1]. The main contribution of this paper is to design a simple Web Content Management System (WCMS) for Dental Laboratories and to discuss why there is a need to develop a standalone WCMS for Dental Laboratories whilst other open source WCMSs can be utilized such as Joomla, Drupal and WordPress.

Keywords: Web content management system, Dental laboratory system, Three-tier architecture, Commercial content management system, Open source content management system

1. Introduction

Not all dental clinics have their own dental laboratories. Small to medium clinics send their patient lab cases to local, national or sometimes international dental laboratories. Communication is done through mailing handwritten forms or sending files and bills by email. Usually each dental clinic works with one dental laboratory. However, dental laboratories receive lab cases from different dental clinics. Lots of time is wasted in both sides on trying to track down missing lab case information over the phone or email, working out unreadable handwritten prescriptions, or following up on billing and payments.

Maintaining a long-term relationship between dental laboratories and their customers (dental clinics and dentists) urges active communication process between two sides. According to a Marketing Director at the Continental Dental Laboratories: “*communication—or a lack of it—will make or break the relationship between a laboratory and the dentist*”. Until now, this communication process is done through a handwritten prescription and an impression that may or may not have been able to completely give the required information to the technician to meet the dentists’ expectations. Vice President, Sales & Marketing at Trident Dental Laboratories agrees that the fewer laboratories have to depend on verbal or written instructions, the better. The laboratory work depends on the prescription form received from the clinic, any simple mistake or incomplete information will result in loss in money and customers [2].

It is not only about lab case management, the dental laboratory, the dental technicians, and the laboratory owner have an obligation towards dentists to share their knowledge with them and to educate them regarding new products. Whether for product education or case management, communication between the dentist and

the laboratory is the most important factor that has a great influence on the success of the relationship [2].

The market is crowded with desktop applications for managing dental laboratories operation; however, we are looking for a system that utilizes both case management and a relationship between dental laboratory and its customer. Many information need to be shared between dentists, dentist assistant, laboratory technicians, laboratory owner and others. With the development of the Internet technology, there is a need for an effective web application to manage such operation and to control accessing this shared information. Thus, the motivation of this paper is to design a simple Web Content Management System (WCMS) for creating dental laboratories websites.

In daily services, a dentist in a certain clinic fills a form as in Figure 1 to order a lab case from a certain dental laboratory. Then, the dental assistant and other staff arrange with the laboratory for pickup, payment and delivery. This paper-based recording imposes several major drawbacks namely miscommunication between the laboratory and the clinic and lack of visual interactivity.

Figure 1- Dental Laboratory Form

With dental laboratory WCMS, the medical data, pictures and patient records all can be accessible online. Dental laboratories can track and manage lab cases and

payments online. Besides, dental clinics and dentists can access to track the lab cases, and they can be notified by email about the status of their lab cases. Both have an account and can view their lab case history, their current balance, and pay bills online. It aids in reducing paperwork and automating the approach to process lab cases.

Any dental laboratory can have its own instance of the system to manage its content and to ease communications with dental clinics its working with.

The basic idea of web content management systems is to get organized and find a logical, consistent and easy way to place content on the web [3]. It allows non-technical users to create, edit, manage and control a large, dynamic collection of web material (HTML documents, images and video). WCMS involves a lifecycle starting from creation to destruction of content. The lifecycle includes reviewing the content before publishing it and it may include archiving before destroying. WCMS helps in keeping the site more consistent, ease the navigation, and most important it aids in controlling and tracking the content [4].

Figure 2 shows the framework for the dental laboratory WCMS. The core of this dental laboratory WCMS is the content, which is the patient lab case that is being sent from a certain dental clinic to the dental laboratory to be produced. The full content lifecycle starts from a dentist in a clinic submitting new patient lab case to the system and then, a laboratory technician is assigned to process this lab case. The content will be archived and later destroyed after delivering the patient lab case and receiving the payment.

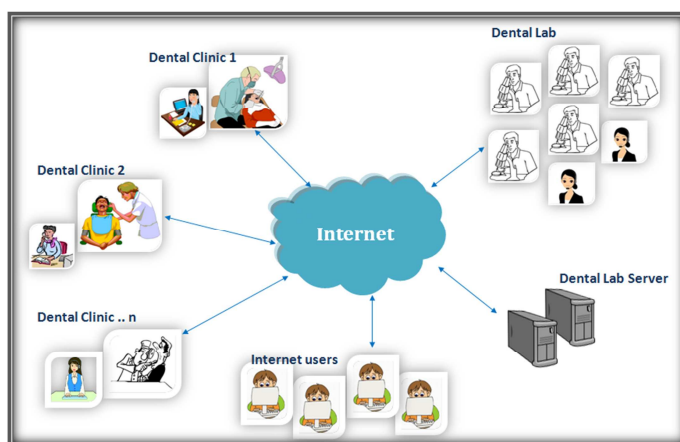


Figure 2-System Framework of Dental Laboratory WCMS

2. Related Background

2.1 Commercial Dental Laboratories WCMS




Similar dental laboratory management systems envisaged is available on the international market in various formats, however the products available have several disadvantages. The greatest and most obvious disadvantages are:

- They are costly and usually unaffordable for small to medium dental clinics.
- They do not combine web management and content management.

- They normally do not have user friendly interface.
- They do not offer any kind of personalization or customization to their customers.
- They only offer dentist account; there is no dental clinic account.

We conduct a research and create a short list of systems in order to be further examined and narrowed or widened to fit the small to medium dental laboratories need. Table 1 shows a comparison between some of the systems.

Table 1: Comparison between a few system

	 ¹	 ²	 ³
Web-based	√	Only via DDX ⁴	X
On line scheduling	√	√	√
Rescheduling	√	Only via DDX	X
Patient lab case on line tracking	√	Internal only	√
Billing	√	√	√
Dentists Profile	√	√	√
Clinic Profile	X	X	X
Attaching files	√	Only via DDX	X
Reports	√	√	X
Customer Service	√	√	X
Chatting system	X	X	X
Personalization	X	X	X
Customization	X	X	X

¹ <http://evidentlabs.com/>

² <http://www.labnet.net/>

³ <http://www.sarals.com/Precise.htm>

⁴ DDX is a web based system that turns Labnet into a Web-enabled application.

2.2 Open-source WCMS Utilized by Dental Laboratories

In addition to these commercial WCMS solutions, many open source solutions are utilized using *Joomla*, *Drupal* and *WordPress*. We have studied some websites which are utilized by different open-source solutions:

- Websites powered by **Joomla**:
 - Quest Dental Laboratory⁵
 - A+ Dental Laboratory⁶
- Websites powered by **Drupal**:
 - Vision Dental Laboratory⁷
 - Mascol Dental Laboratory⁸
- Websites powered by **WordPress**:
 - ArrowHead Dental Laboratory⁹
 - Keller Laboratory¹⁰
 - Nik Dental Laboratory¹¹

Following are the common features of these websites:

- **Sending a Lab Case**: By filling handwritten form and schedule for pickup.
- **Online Account**: For scheduling a pickup.
- **Lab Case Tracking**: Not available.
- **Shipment Tracking**: Via carrier.
- **Billing and Statements**: Sent by mail.
- **Product Education**: promotes education for dentists through all the typical means, such as direct mail, journal ads, articles that they place in journals, and their Web site content.

Among all of the envisaged websites, none of them offer clinic and dentist account. Besides, they do not have a user-friendly interface. They are more likely static

⁵ <http://questdental.us/>

⁶ <http://www.a-plusdentallab.com/>

⁷ <http://visiondentallaboratory.co.uk/>

⁸ <http://mascoladentallab.com/>

⁹ <http://www.arrowheaddental.com/>

¹⁰ <http://www.kellerlab.com/>

¹¹ <http://www.nikdentallab.com/>

WebPages of mostly informational content with simple designs. On the next section, a discussion is presented whether it is better to build a custom WCMS or to utilize an open-source solution.

2.3 Commercial vs. Open-source WCMS

As mentioned previously, there are many open-source solutions. Some of them are being created for many years, empowered by developers with technical background. A question may arise why we need standalone WCMSs for dental laboratories? Why dental laboratories do not utilize open source WCMSs solution? Although this is not the main goal of this research, it is worth it to bring up this discussion.

On the one hand, having a commercial WCMS specifically for dental laboratories will allow for full flexibility in developing [5]. Once it is available, many dental laboratories can utilize it instead of utilizing open-source solutions, this is because, the entire application is setup so it works exactly how needed by the dental laboratory. It will be faster to implement and associates a certain degree of safety as opposed to open-source. Moreover, it offers more support and stronger training documentation than open-source [6]. Probably the most important concern regarding the commercial WCMS especially for small to medium dental laboratories is the cost.

On the other hand, for small to medium dental laboratories, open source WCMSs offer a low cost alternative to commercial solutions. Besides, Troubleshooting is made easier because of the technical support and online community. However, potential concern regarding the open-source solutions is the security. As the source code is available for public, attackers can use the source code to identify vulnerabilities. Thus, these systems raise significant security issues [7].

According to [8] in 2010 and [9] in 2011, the best WCMSs today are *Joomla* and *Drupal*. Having

studied them, *Drupal* is the best to utilize when developing large websites with hundreds of pages but smaller websites with lesser number of pages are better developed by *Joomla* [10]. *Drupal* is not very user-friendly, terms are confusing and the admin interface is relatively poor, whilst *Joomla* is more user-friendly with a more active developer and designer community. With *Drupal*, unlimited user permissions levels can be created, but *Joomla* offers only three user levels (Public, Registered and Special). *Drupal* does not support multimedia, photo galleries by default but *Joomla* supports multimedia by its default editor [11]. Of course both of their developers have overcome these issues by developing modules to extend their usability. However, modules not always free or easy to install.

To sum up the discussion, both approaches the commercial and the open-source have their advantages and disadvantages. It is depending on the requirements of the system, so there is no absolute answer to which is 'best'. However, open-source WCMSs does not fit the requirements of the system presented in this paper. We need many levels of permissions with user-friendly interface. In addition to secure websites to handle payments.

As we will see in the next section, the system and user requirements are not supported by the available dental laboratories commercial WCMS, thus, we are designing a new commercial WCMS for Dental Laboratories.

3. Overall System Analysis and Design

3.1 System Analysis

3.1.1 System Components and System Users for Dental Laboratory WCMS

The system has three major types of system engines (components) similar to the system proposed on [12, 14] in addition to the data warehouse repository:

- **A Content Editorial Engine** provides content and repository maintenance and approval functions for different levels of administrators in the dental laboratory.
- **A Content Reception Engine** collects content from external sources, and then delivers it to different parts of the system for approval and publication.
- **A Content Publishing Engine** stores approved content and send them to different parties via different channels (such as email, fax, and text messaging). It also serves as the Web storefront of the dental laboratory for user enquiries.

Figure 3 depicts an overview of the Dental Laboratory WCMS highlighting the main system components and system users. A Dental Laboratory WCMS must be designed specifically to match the need and interest of each system user within and related to the dental laboratory. Besides the management, there are four main types of system users involved, namely, *Content Creators*, *Content Providers*, *Content Distributors*, and *Content Users*:

1. ***Content Creators*** collectively refer to internal users who are involved in the content creation processes of the dental laboratory. The Dental Laboratory WCMS should be able to accommodate the different operational and administrative requirements of these different roles of internal users and to maintain appropriate security control. They interact mainly with Content Editorial Engines of the Dental Laboratory WCMS.
2. ***Content Providers*** are external sources (such as PayPal) providing content (such as payments) to the dental laboratory through a Content Reception Engine. To ensure timeliness, content from trusted sources are usually forwarded automatically to the Content Publishing Engine for immediate delivery.
3. ***Content Distributors*** are external service providers that render the content and deliver them to clients via different (traditional or electronic) channels, such as mass fax, mail, email, hardcopy delivery, and so on.

4. **Content Users**, who can be internal or external to the dental laboratory, are classified into three types in our case. Content Users obtain content access through a Content Publishing Engine. Based on their subscription data, the Content Publishing Engines also actively send appropriate content to the subscribed users. The three types are:

- *Public Visitors* – Anonymous users are often allowed to access some limited amount of public content through a portal. This helps attract them to visit the dental laboratory’s Web site.
- *Clients (dental clinics and dentists)* – Customers who do basic business with the dental laboratory are allowed access and subscription to all unrestricted content. They have their own gateway where they can track their lab cases, pay and check their bills.
- *Internal Users* – Internal staff can access “internal only” content related to them, as well as all the content for external users. They are also automatically subscribed to relevant content, according to their job functions, secretary, technician, driver, and so on.

3.1.2 Description of the Main System’s Functions and Workflow

As mentioned earlier, the core of this dental laboratory WCMS is the content, which is the patient lab case that is being sent from dental clinic or a dentists to the dental laboratory to be processed. The full content lifecycle content creation, content editing, content approval and content publishing, which together consist of the core part of the website content management system. Following is description of each cycle:

1. **Lab case creation:** ability to create new lab case and submit it. In addition, content creation should provide basic editing methods and editing tools and be able to upload and download images in the content. Dentists can login using their accounts into the system and submit their cases online using an electronic form. In addition, the system will allow dentists to choose the technician if they wish. If the technician is available and able to accept the case and finish it by required time then the case will be assigned to that technician. Otherwise, the system will notify the dentist that the technician is not available and will give him/her the choice to choose someone else or just leave it for the laboratory staff to assign the lab case to a technician. Approximate pricing will be calculated after filling the form and indicating the due date for delivery. In addition, the system will offer the ability to attach files to lab cases. Dentists can attach images and any other file that is needed for better understanding of the lab case.

2. **Lab case editing:** editing should satisfy the requirement of submitted lab cases in terms of querying, previewing, modifying, deleting, submitting, etc. such basic operation and management, edit content items without affecting the published work and distribute the authority to approval submitted contents. In addition, lab case editing should track the process of lab case submission and approval status which has been rejected or in the process of approval. Thus, laboratory staff can view the lab cases as soon as they are submitted so they can arrange a pickup.

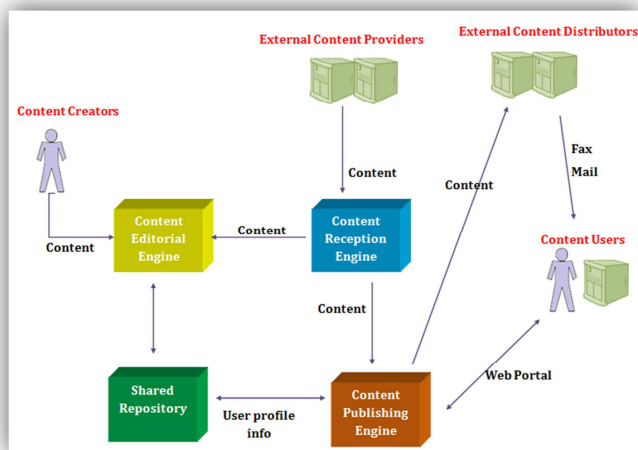


Figure 3. System users and components

When the lab case arrived, its status is updated as received.

3. **Approval process:** Approval process can add, modify, delete and manage the authority of the allocation roles and individuals. Provide a workflow that is configurable for users to allow different approval processes with varying case item status during the authoring, establish a variety of roles within a workflow process, and assign workflow to classes of content items as well as roles and individuals. The workflow in the dental laboratory will be managed by the ability to check the status of every lab case. In addition, the dental clinic will be notified by email about the status of the lab cases whether they have been received, processed or out for delivery, etc. Of course dentists can login to system and check the status as well.
4. **Case approval:** ability to approve the submitted lab case which is in the edit process, query and view the submitted lab cases for approval. There are two kinds of approval states: passed and not passed. The function that approval process should implement is to ensure each approval step can authorize only one approver.
5. **Case publishing:** The editorial content can be published after passing the approval process, in the process of content publishing, cancellation and republished function should be provided. Published content should apply static html pages as contents storage form, with the publication of the contents, the unpublished content and page module can be updated as well, which also makes publishing Web pages convenient to manage and update. Thereby increase the speed of page browse and access. Laboratory staff will be alerted when lab cases are ready to deliver.
6. **Website configuration and management:** Ability to classify and manage the website columns of the publishing content, including the basic operation of the columns, such as columns add, modify,

delete, as well as the template option of page modules and website columns. Besides, Website resource management can achieve the basic functions, such as upload and download files management, configuration and management of the published website parameters.

7. **Rights authority:** include allocation of operating authority in various sectors of system function module.

3.2 System Design

In the dental laboratory context, the initial structure that would be used in the WCMS had the following characteristics:

1. One central site for the dental laboratory (each dental laboratory would contain a unique domain).
2. The entire application is installed on a hosting web server for that dental laboratory.
3. A separate instance would be created for every dental clinic or dentists.

The proposed WCMS adopts Three-Tier Architecture design as shown in Figure 4 and is composed of three layers [1]. Since there are many computers with different kinds of operating system will work in coordination in the system, platform independent system architecture is needed to adapt the change in future use. Following is brief description for each tier [15]:

- The first (bottom) layer of the system is the database layer. It saves the system's content such as lab cases' data, dentists' data and images, etc. Content is frequently stored as XML since variety of data sources are used and each has its own characteristics [16]. XML is used to solve the incompatibility of different structures. XML facilitates reuse and enable flexible presentation options [13]. This layer mainly complete the local query, extract and transform distributed information from heterogeneous data sources. It uses Wrapper technology including the queries translate function and result translate function. It can

translate the query result which gets from middle layer to local process. Extract the query result and makes a XML document. Finally return the document to middle layer [4].

- The second (middle) layer is the system transaction layer, which is consisted of the system function modules, such as the lab case tracking. This layer mainly contains the query splitter and results Integrator two functions. In order to implement centre process, the system must use a common model which comes from different sources of information from a variety of data XML's characteristics determine that it is can describe a variety of data. It is means that it is a common data model. Heterogeneous data integration can solve this problem. It also enables the dynamic data release. Therefore, this system uses the XML model as a common model. It provides a unified query view by middleware layer on the client. It accepts the client's query command, split into various sub-queries and assigns to various data sources; and then integrates the results of its inquiries, sends to the client's browser displays for the user [4].
- The third (top) layer is the user interface layer, which included the client side of the system and displays the content to the users.

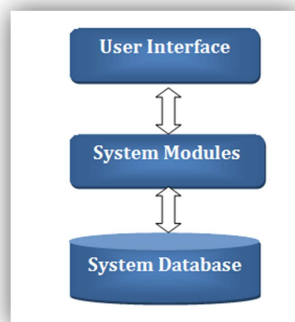


Figure 4-The system architecture

4. Conclusion and Future Work

This paper proposed Three-Tier architecture design to develop a WCMS for Dental Laboratories based on a study of the requirements of dental laboratories. The system would enhance the clinical management level and would be used as a channel for dentists to technician, dental clinic-to-dental laboratory, to provide the long-term relationship and information sharing. The paper also discussed the need to have a standalone WCMS for dental laboratories other than the open-source WCMS. The system will significantly improve performance of dental laboratories and will assure long-relationship term between dental laboratories, dental clinics and dentists. It is expected that in a couple of years, the proposed system will be developed and implemented in several dental labs in Riyadh, Saudi Arabia.

5. References

- [1] Jian Yu; , "Distributed Data Processing Framework for Oral Health Care Information Management Based on CSCWD Technology," Information Science and Engineering (ICISE), 2009 1st International Conference on , vol., no., pp.2312-2315, 26-28 Dec. 2009.
- [2] Article on "Dental Labs: A Vital Key to Your Success", Inside Dentistry, September 2009, Volume 5, Issue 8, Published by AEGIS Communications, available at: <http://www.dentalaegis.com/id/2009/09/dental-labs-a-vital-key-to-your-success> (accessed in May 2011).
- [3] McNay, H.E.; , "Enterprise content management: an overview," Professional Communication Conference, 2002. IPCC 2002. Proceedings. IEEE International , vol., no., pp. 396- 402, 2002.
- [4] Richard Vidgen, Steve Goodwin, Stuart Barnes; "Web Content Management, e-Everything: e-Commerce, e-Government, e-Household, e-Democracy", 2001, 14th Bled Electronic Commerce Conference, Bled, Slovenia.
- [5] Nakwaski M. and Zabierowski W., "Content Management System for Web

- Portal”, TCSET'2010, February 23-27, 2010, Lviv-Slavske, Ukraine.
- [6] Article on “*Custom CMS vs. Open Source CMS*”, December 2009, Sleepless Media - Website Design in Santa Cruz, California, USA, available at: <http://www.sleeplessmedia.com/blog/2009/12/custom-cms-vs-open-source-cms/> (accessed in May 2011).
- [7] Meike, M.; Sametinger, J.; Wiesauer, A.; , "Security in Open Source Web Content Management Systems," Security & Privacy, IEEE , vol.7, no.4, pp.44-51, July-Aug. 2009.
- [8] Article on “*The Best Three CMS today*” April 2010, available at: <http://tech.comfiles.com/the-best-three-cms-today> (accessed in May 2011).
- [9] Arah T., article on “*The Best CMS: Joomla 1.6 vs Drupal 7.0*” February 2011, PC Pro blog, available at: <http://www.pcpro.co.uk/blogs/2011/02/02/Joomla-1-6-vs-Drupal-7-0/> (accessed in June 2011).
- [10] Bose S., article on “*Drupal vs. Joomla: Advantages and Disadvantages*”, February 2011, Evon Technologies, available at: <http://technology.ezinemark.com/Drupal-vs-Joomla-advantages-and-disadvantages-7d2d2b2f178c.html> (accessed in May 2011).
- [11] Burg S., article on “*Joomla and Drupal - Which One is Right for You?*”, December 2009, available at: <http://www.alledia.com/blog/general-cms-issues/Joomla-and-Drupal-version-2/> (accessed in May 2011).
- [12] Gu, Y.; Warren, J.; Stanek, J.; Suthers, G.; , "A System Architecture Design for Knowledge Management (KM) in Medical Genetic Testing (MGT) Laboratories," Computer Supported Cooperative Work in Design, 2006. CSCWD '06. 10th International Conference on , vol., no., pp.1-6, 3-5 May 2006.
- [13] JSR Subrahmanyam; ‘ “*Future Trends Of Content Management Systems (CMS) for e-Learning: A Tool Based Database Oriented Approach* “ Technical, Quantum Softech Limited, Hyderabad, India.
- [14] Kwok, K.H.S.; Chiu, D.K.W.; "A Web services implementation framework for financial enterprise content management," System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on , vol., no., pp. 10 pp., 5-8 Jan. 2004.
- [15] Preuner, G. and Schrefl, M. (2000). "A three-level schema architecture for the conceptual design of web-based information systems: from web-data management to integrated web-data and web-process management", World Wide Web 3, 2 (Mar. 2000), 125-138.
- [16] Yuelan Liu; Lu Yang; Yanqing Zheng; , "Research and design of open teaching and learning platform based on XML middleware," E-Health Networking, Digital Ecosystems and Technologies (EDT), 2010 International Conference on , vol.2, no., pp.356-359, 17-18 April 20.

An S-System Analysis of the Sulfur-Deprivation Response of the Microalga *Chlamydomonas reinhardtii* during Biohydrogen Production

Jack K. Horner
PO Box 266
Los Alamos NM 87544 USA
email: jhorner@cybermesa.com

Abstract

Producing biohydrogen on a commercial scale will likely require the genetic re-engineering of natural hydrogen-producing organisms. Kinetic modeling of hydrogen-producing metabolic pathways can cost-effectively help to characterize systemic (e.g., mass/energy/charge conservation) constraints in these organisms. *In vitro* kinetic studies suggest that the activity of the hydrogenases in several photolytic biohydrogen producers (PBPs) could be increased to as much as four times their nominal *in vivo* rate. It is much less clear, however, whether the *in vitro* activity maximum could be realized *in vivo*. Here I use an S-system photosynthesis-based PBP (PS-PBP) simulator to analyze the H₂ production of sulfur-deprived *C. reinhardtii*, a water-photolyzing PS-PBP microalga. The analysis strongly suggests that maximum H₂ production by the alga requires some sulfate to be present in order to enable an initial purge (in the form of O₂) of the oxygen arising from the photolysis of water, accompanied by a corresponding rise in proton (also from the photolysis of water) pressure which helps to drive the H₂ formation. After the initial O₂ burst, residual oxygen from the photolysis of water is consumed by CO₂ formation in the mitochondria.

Keywords: biohydrogen, S-system, metabolic modeling

1.0 Introduction

Kinetic modeling of hydrogen-producing metabolic pathways can cost-effectively help to characterize systemic (e.g., mass/energy conservation) sensitivities in photolytic biohydrogen producers, even if all the details of hydrogen-gas producing metabolic pathways are not known. Among the more promising candidates for hydrogen-production optimization are photolytic biohydrogen producers (PBPs) such as the microalga *Chlamydomonas reinhardtii* ([7], [8]). It is generally held that the hydrogen-producing pathways in many PBPs incorporate segments of the PS-I and PS-II photosynthetic pathways ([6],[13]), and electrons from the anaerobic degradation of starch, to help accumulate the

electron free energy required to allow a hydrogenase to convert protons to H₂ ([17]). *In vitro* kinetic studies suggest that the activity of hydrogenases isolated from several PBPs could be increased to as much as four times their nominal *in vivo* rate ([1]). *C. reinhardtii* produces H₂ only if deprived of sulfur. Here I use *bioh2gen* ([15]), an S-system ([2], [11]) PS-PBP kinetics simulator, to argue that maximum H₂ production by the alga requires some sulfate to be present in order to enable an initial purge (in the form of O₂) of the oxygen arising from the photolysis of water, accompanied by a corresponding rise in proton pressure which helps to drive H₂ formation. After the initial O₂ burst, residual oxygen from the photolysis of water

is consumed by CO₂ formation in the mitochondria.

2.0 S-systems

An S-system ([11],[12]) is a power-law-oriented, finite-difference system of ordinary differential equations (SODE) each of whose dependent variables X_i is described by a kinetic equation of the form

$$dX_i/dt = \alpha_i \prod_j X_j^{g_{i,j}} - \beta_i \prod_j X_j^{h_{i,j}}$$

Eq. 2.1

where

- the left-hand side of Eq. 2.1 is the first derivative of X_i with respect to time
- $i, j = 1, 2, 3, \dots, N$
- $\{X_i\}$ is the set of real-valued dependent variables of the system
- for any given X_i , only those independent and dependent variables X_j that have an action on X_i are included as factors in the products on the right-hand-side (RHS) of Eq. 2.1. The factors in the first term on the RHS of Eq. 2.1 correspond to just those entities that increase or inhibit the production of X_i ; the factors in the second term of the RHS of Eq. 2.1 correspond to just those entities that contribute to, or inhibit, the consumption of X_i .
- $\alpha_i, \beta_i > 0$
- $g_{i,j}, h_{i,j}$ are real-valued

There is a natural mapping from a biochemical map, K , to equations that have the form of Eq. 2.1. In particular, let $K = \langle \{X_k\}, E \rangle$, $E \in \{X_k\} \otimes \{X_k\}$, $k = 1, 2, \dots, N$, be a directed graph in which each distinct $X_i \in \{X_k\}$ corresponds to a distinct variable (e.g., the concentration of a distinct chemical species in the map), and $w \in E$ if and only if $w = (X_m, X_n)$ is a directed edge in K , $m \neq n = 1, 2, \dots, N$.

α_i and β_i are called *generalized rate constants* (or just rate constants) for X_i , and $g_{i,j}$ and $h_{i,j}$ are called the *generalized kinetic orders* (or just kinetic orders) for X_i , on analogy with standard chemical kinetic theory. The subexpression i_j indicates the action of X_j on X_i .

An S-system has several desirable features, including the fact that it is fully characterized by its rate constants and kinetic orders. Any SODE can be *recast* ([10],[11]) as an S-system without loss of accuracy or precision; the recasting, however, is not in general unique. In addition to biochemical systems, S-systems have been successfully used to model epidemics, forest diversification, and world dynamics.

3.0 A network model of hydrogen production in PS-PBPs

I will call bioH₂ producers that exploit portions of the PSII or PSI pathways “photosynthetic” PBPs (PS-PBPs). The schematized PS-PBP model used in the

present study is shown in Figure 1 and is similar to [3], [4], [5], [9] and [14]. It represents a consensus working hypothesis held by the biohydrogen research community about the high-level metabolics of hydrogen production in PS-PBPs ([7]).

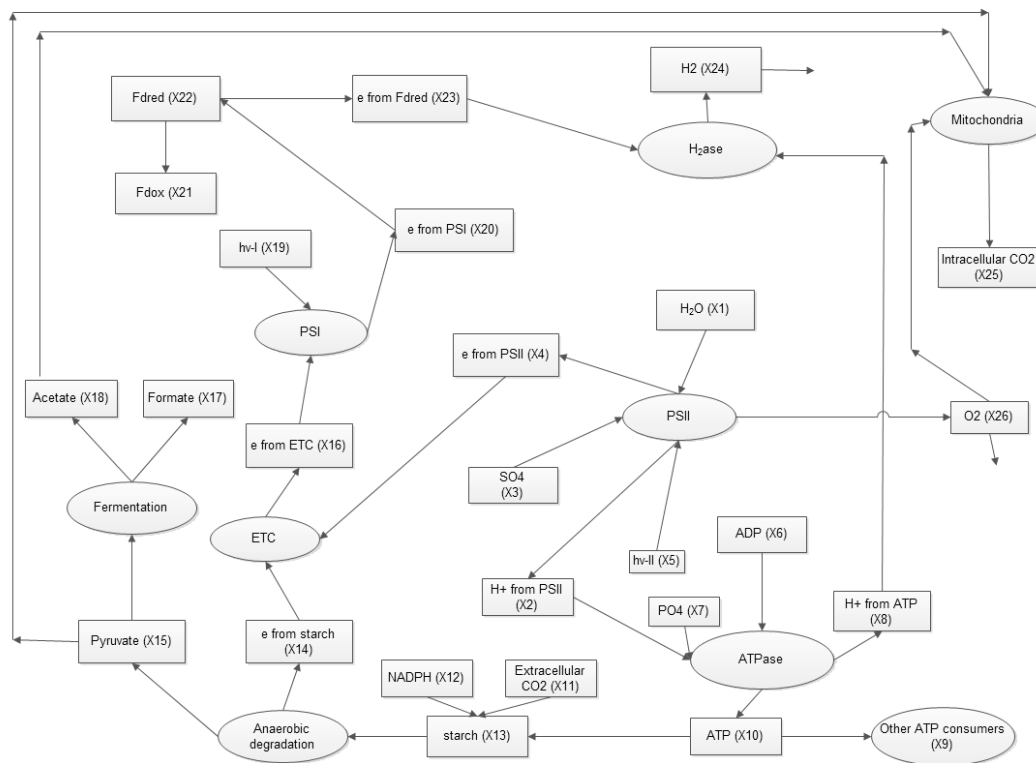


Figure 1. Schematized hydrogen producing metabolic network for PS-PBPs. Rectangles represent sources or sinks of physical quantities of interest (such as mass, concentration, or photon count) named in those rectangles, ellipses represent transforms (which may be complexes of reactions not individually modeled here), and an arrow from an ellipse to a rectangle means that the transform named in the ellipse affects the quantity/concentration of the chemical species named in the rectangle. Legend: PSI = photosynthesis stage I; PSII = photosynthesis stage II; SO₄ = sulfate; hv-I = photons incident to PSI; hv-II = photons incident to photosynthesis PSII; ADP = adenosine diphosphate; ATP = adenosine triphosphate; PO₄ = inorganic phosphate; O₂ = oxygen gas; ATPase = adenosine triphosphatase; e from starch = electrons from anaerobic starch degradation; H₂ase = hydrogenase; ETC = electron transport chain; e from PSII = electrons from PSII; e from PSI = electrons from PSI; Fdred = ferredoxin, reduced; Fdox = ferredoxin, oxidized; H₂ = hydrogen gas; H⁺ from PSII = protons from PSII; H⁺ from ATP = protons from ATPase. Not all interactions exist in all PS-PBP species.

In sulfur-deprived *C. reinhardtii*, oxygen gas production under the experimental conditions of [7] (1-L, 6×10^6 cell/mL preparation) is about 1 mmol/h after beginning of sulfur deprivation, and spontaneously ceases ~10 h thereafter. 30 - 50 h after beginning of sulfur deprivation, the algae begins releasing hydrogen at a rate of ~0.17 millimole H_2 /h (1-L, 6×10^6

cell/mL preparation). ~100 h after beginning of sulfur deprivation, under the experimental conditions of [7], hydrogen production ceases. These trajectories provide strong constraints on any model of bio H_2 production by *C. reinhardtii*.

The S-system equations used in this study are shown in Figure 2.

```

// protons from PSII
X2' = a2 X1^g2_1 X3^g2_3 X5^g2_5 - b2 X10^h2_8 X2^h2_2 X5^h2_5

// e from PSII
X4' = a4 X1^g4_1 X3^g4_3 X5^g4_5 - b4 X16^h4_16 X4^h4_4

// protons from ATPase
X8' = a8 X6^g8_6 X7^g8_7 X2^g8_2 - b8 X8^h8_8 X24^h8_24

// other ATP consumers
X9' = a9 X10^g9_10 - b9 X9^h9_9

// ATP
X10' = a10 X2^g10_2 X7^g10_7 X6^g10_6 - b10 X13^h10_13 X9^h10_9 X10^h10_10

// starch
X13' = a13 X12^g13_12 X11^g13_11 X10^g13_10 - b13 X14^h13_14 X15^h13_15 X13^h13_13

// e from starch
X14' = a14 X13^g14_13 - b14 X16^h14_16 X14^h14_14

// pyruvate
X15' = a15 X13^g15_13 - b15 X25^h15_25 X18^h15_18 X17^h15_17 X15^h15_15

// e from ETC
X16' = a16 X14^g16_14 X4^g16_4 - b16 X20^h16_20 X16^h16_16

// formate
X17' = a17 X15^g17_15 - b17 X17^h17_17

// acetate
X18' = a18 X15^g18_15 - b18 X15^h18_25 X18^h18_18

// e from PSI
X20' = a20 X16^g20_16 - b20 X22^h20_22 X20^h20_20

// Fdox
X21' = a21 X22^g21_22 - b21 X21^h21_21

// Fdred
// X22' = a22 X20^g22_20 - b22 X21^g22_21 X23^g22_23 X22^h22_22

// e from Fdred
X23' = a23 X22^g23_22 - b23 X24^h23_24 X23^h23_23

// H2 gas
X24' = a24 X23^g24_23 X8^g24_8 - b24 X24^h24_24

// Intracellular CO2
X25' = a25 X15^g25_15 X18^g25_18 X26^g25_26 - b25 X25^h25_25

// oxygen
X26' = a26 X1^g26_1 X3^g26_3 X5^g26_5 - b26 X26^h26_26 X25^h26_25 X5^h26_5

```

Figure 2. S-system equations for the dependent variables used in this study. “^” is exponentiation. “>>” means “expression continuation”. “’” means “first derivative with respect to time”. Note that the equation for X2' has light as a *consumption* factor because activity *decreases* as light intensity increases above an optimal value.

Table 1 shows the values of the independent variables of the system.

Table 1. Values of the independent variables of the system.

Independent variable	Value (relative units)
X1 (water)	1
X3 (SO ₄)	0.3
X5 (hv-II)	2.363
X6 (ADP)	100
X7 (PO ₄)	100
X11 (Extracellular CO ₂)	3e-3
X12 (NADPH)	1e-6
X19 (hv-I)	2.363

Much of the system in Figure 1 is based on PSII and PSI kinetics. The model was calibrated (to produce the "nominal" configuration) on PSII/PSI kinetic data in [16], setting all generalized rate constants to 0.1, except a₂ (= 3e-4), b₂ (= 1e-4), a₄ (=0.01), a₂₄ (=1e-4), b₂₄ (= 0.001), a₂₆ (=10), and b₂₆ (=1000); these exceptions were based on *in vitro* experimental values obtained in [7]. All generalized kinetic orders were set to 1.

bioh2gen and the model used in [14] differ in a few ways. First, following the conventions in [11] for modeling metabolic systems in the absence of gene-circuit dynamics, no enzyme is an explicit variable of *bioh2gen*; several enzymes are variables in [14]. Second, *bioh2gen* employs more rate constants derived from experiment than does the model used in [14]. Third, all the kinetic orders in

bioh2gen were set to 1; two kinetic orders were set to 2 in [14]. Fourth, *bioh2gen* study models the photon inputs to each of PSII and PSI individually; the model in [14] represents only the photon inputs to PSII.

The nominal H₂ and O₂ production rates of *bioh2gen* were compared to [7], and the responses of the model to sulfate concentrations ranging from 0.01 to 5.0 (relative units) were computed.

4.0 Results and discussion

Figure 3 show the nominal (hv-I and hv-II = 2.363) hydrogen and oxygen output predicted by the model described in Section 3.0. The H₂ and O₂ outputs agree well with [7].

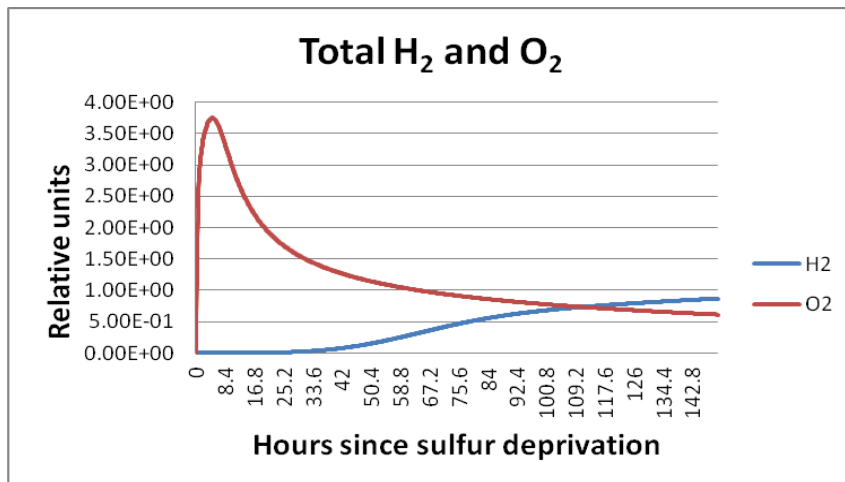


Figure 3. Nominal (for SO₄ = 0.3) total hydrogen and oxygen gas production as a function of time (units on the horizontal axis are hours after t₀). Note the initial oxygen burst. The values predicted by the model agree well with the results shown in [7].

Figure 4 shows the H₂ gas production in the model as a function of sulfate concentration.

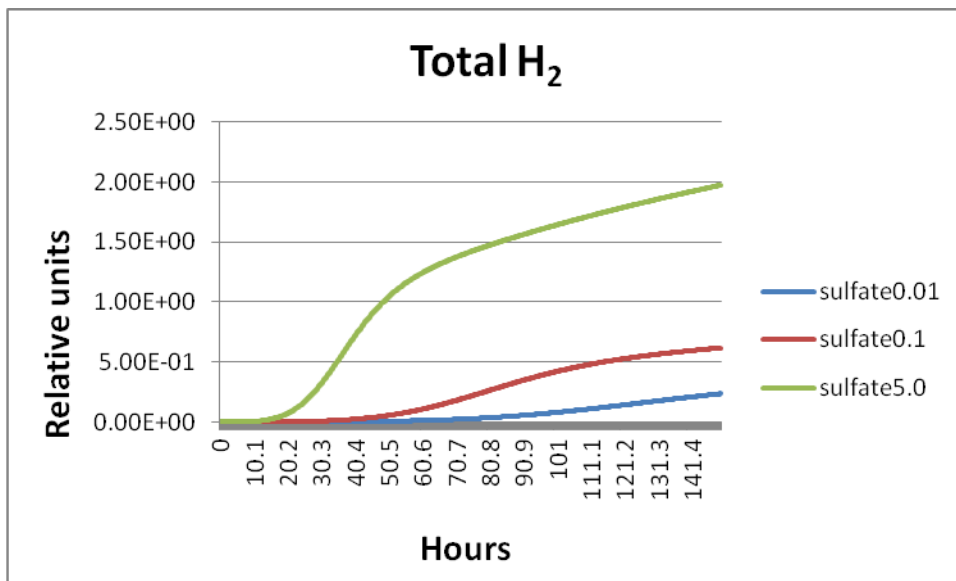


Figure 4. H₂ production (in relative units) as a function of sulfate concentration. (The trajectory for a sulfate concentration of 5.0 relative units may not be biologically realistic.)

Figure 4 strongly suggests that, within the model described in Section 3.0,

maximum H₂ production by the alga requires some sulfate to be present in order

to facilitate an initial burst (in the form of O_2 (See Figure 3)) of the oxygen resulting from the photolysis of water, accompanied by a corresponding rise in proton (also from the photolysis of water) pressure which helps to drive the H_2 formation. (Even if the algal growth medium contains no sulfate, some sulfate is still available within the reactions supporting PSII.) After the initial O_2 burst, residual oxygen from the photolysis of water is accommodated by CO_2 formation in the mitochondria (X25 in the model).

5.0 Acknowledgements

This work benefited from discussions with Maria Ghirardi and Michael Seibert of the National Renewable Energy Laboratory, Anastasios Melis of the University of California/Berkeley, Anatoly Tsygankov of the Institute of Basic Biological Problems (Pushchino, Russia), Orlando Jorquera of the Federal University of Bahia, Murray Wolinsky of Los Alamos National Laboratory, and Jorge Soberón of the University of Kansas Biodiversity Institute. For any errors that remain, I am solely responsible.

6.0 References

- [1] Cammack R. Hydrogenases and their activities. In Cammack R, Frey M, and Robson R, eds. *Hydrogen as a Fuel: Learning from Nature*. Taylor and Francis. 2001.
- [2] Ferreira AEN. *Power Law Analysis and Simulation (PLAS)*. Version 1.2 beta, Build 0.120. URL <http://correio.cc.fc.ul.pt/~aenf/plas.html>. March 2011. Note: the link to the PLAS software appears is broken as of 1 January 2012. A copy of the software is available on request from the author of the present paper.
- [3] Horner JK. An S-system model of hydrogen production in microalgae. *International Society for Computational Biology 2002, Special Interest Group for Biological Simulation Satellite Meeting (SIGSIM2002), Computer Modeling of Cellular Processes*. Edmonton, Alberta, Canada.
- [4] Horner JK. Leveraging biohydrogen research: a kinetic modeling approach. *Hydrogen and Fuel Cells Conference 2003*. Vancouver, British Columbia, Canada.
- [5] Horner JK and Wolinsky MA. A power-law sensitivity analysis of the hydrogen-producing metabolic pathway in *Chlamydomonas reinhardtii*. *International Journal of Hydrogen Energy* 27 (2002), 1251-1255.
- [6] Lawlor DW. *Photosynthesis*. Third Edition. Springer. 2001.
- [7] Melis A et al. Sustained photobiological hydrogen gas production upon reversible inactivation of oxygen evolution in the green algae *Chlamydomonas reinhardtii*. *Plant Physiology* 122 (2000), 127-135.
- [8] Melis A. Green alga hydrogen production: progress, problems, and prospects. *International Journal of Hydrogen Energy* 27 (2002), 1217-1228.
- [9] Horner JK. *bioh2gen, Version 1*. Available on request from the author. 2004.
- [10] Savageau MA. Growth of complex systems can be related to the properties of their underlying determinants. *Proceedings of the National Academy of Sciences* 76 (1979), 5413-5417.
- [11] Voit EO. *Computational Analysis of Biochemical Systems*. Cambridge. 2000.
- [12] Drazin PG. *Nonlinear Dynamics*. Cambridge. 1992.
- [13] Markvart T and Landsberg PT. Solar cell model for electron transport in photosynthesis. *Proceedings of the 29th IEEE Photovoltaic Specialists Conference (2002)*, 1348-1351.
- [14] Jorquera O, Kiperstok A, Sales EA, Embiruçu M, and Ghiardi ML. S-systems sensitivity analysis of the factors that may influence hydrogen production by sulfur-deprived *Chlamydomonas reinhardtii*. *International Journal of Hydrogen Energy* 33 (2008), 2167-2177.
- [15] Horner JK. *bioh2gen, Version 5*, a PLAS simulator for biohydrogen production by photosynthetic biohydrogen producers. Source code is available on request from the author.
- [16] NPO Bioinformatics Japan. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. <http://www.genome.jp/kegg/>. 2012.
- [17] Ghirardi ML and Amos W. Hydrogen photoproduction by sulfur-deprived green algae - status of the research and potential of the system. *Biocycle* 2004, 45-59.

Pseudoneglect and its Many Faces: Laterality of Motor Control Underpins Asymmetries in Line Bisection, Initial Visual Exploration, Optimal Viewing Position, Point of Subjective Equality and Visual Span

I. Derakhshan, MD, Neurologist

Formerly, Associate Professor of Neurology, Case Western Reserve (Cleveland) and Cincinnati Universities, Ohio, USA. Currently, private practice of neurology: 415 Morris Street, Suite 401, Charleston, WV 25301. Tel 304 343 4098, Fax 304 343 4598, Email idneuro@hotmail.com

Abstract: *Contrary to the accepted belief, incontrovertible evidence indicates that all commands are issued in one hemisphere (major, command center, speech hemisphere), with majority of people (~80%) left hemispheric in laterality of their command center. Thus, those commands intended for moving the nondominant side of the body, including the eyes, are first transmitted to the opposite hemisphere via the corpus callosum, before they are implemented by the minor hemisphere (which is devoid of any conscious awareness). This article reviews the many faces of the behavioural consequences of this one-way callosal traffic circuitry underpinning lateralities of motor and sensory control; including our inability to divide a line in exactly two halves without using a ruler (i.e. pseudoneglect). Review of the relevant time resolved data in the literature indicates that the well-known laterality indexed asymmetries in visual span (e.g. the right visual field advantage in lexical decisions), optimal viewing position in reading (OVP), point of subjective equality (PSE) and the tau effect are among the many faces of pseudoneglect, all based on the laterality of motor and sensory control which is unidirectional: from the major to the minor hemisphere for motor and from the minor to the major hemisphere for sensory signals (arising from the nondominant side of the body).*

Humanity possesses twin disabilities, for which the reason has been discovered recently. Firstly, we cannot divide a line precisely in halves and resort to a straight-edge or a tape measure for securing a valid result. This is not because we do not know that the middle is between the two halves; rather, we cannot determine exactly where the half mark must lie without actually measuring the line

and deducing the middle mathematically. Short of this, depending on our neural handedness (i.e. see below for the distinction between neural hand behavioral handedness), we either deviate to the left or to the right of the veridical center by a percentage point of the length of the line without ever realizing that we have erred in the process [1]. Similarly, when viewing a target in the middle of our view, vast majority of right handed people will initiate a search to the left of the midline, focusing slightly to the left of the middle of a word target, or display a leftward point of subjective equality (PSE) when acknowledging the arrival of the stimuli just to the left of the midline using the right hand [2, 3]. It has been shown repeatedly that interference with this automatic process will result in diminished efficiency in the reading a text [4-6]. Whereas the first of these twin failings appears sensory in nature the second is more clearly motoric. Thus, starting from the same distance of a musical keyboard, humans are incapable of striking two keys at the very same time. Musicologists have known about this disability for a long time and since the melody of a tune is traditionally written for the right and the harmony for the left hand have formally named the phenomenon the "melody-lead of the right hand" in piano players [7]. Importantly, however, it has been known for over forty years that severing the anterior aspect of the corpus callosum (the neural bridge between hemispheres) intensifies the above described interlimb asynchrony by further prolonging the performance of left hand [8, 9].

The purpose of this article is to explore evidence that the twin phenomena mentioned above are based on the fact that the interhemispheric traffic, underpinning laterality of motor control, is one-way (from the major to the minor hemisphere, as here defined) and that the

nature of signals employed for the purpose of activating the nondominant side is purely excitatory. It will be shown that the line bisection deviation mentioned above represents an automatic trading of "time" and "space" in the human mind as the motor command for moving the eyes to the left traverses the interhemispheric bridge (the corpus callosum) from the action hemisphere to its neighboring counterpart (minor hemisphere); which in turn implements the commands issued in the former, moving the eyes to the left (in a real right handed person, see below). Thus, the longer route imposed on the command for moving the eyes to the left is interpreted by the person clocking the event as a longer "time," and by the subject as additional "space" (hence the over-estimation) [1, 10-12]. Accordingly, for the person who initiates looking to the left, the automatic "time-stamping" of the event begins after the emanation of the related commands from the major hemisphere [1, 13]. Because of the additional callosum-width routing imposed, however, an asymmetry occurs in the excursion of the two eyes, with those to the left coming out short (by an IHTT). In moving the extremities, on the other hand, when exactly the same excursion is imposed on the right and left arm, the latter must add the additional (callosum-width) extent to its journey's length in order to accomplish the aim set forth by the decision-maker located in the major hemisphere [14].

The Line Bisection Test, Further Observations:

First described by Hall and Hartwell (1884) [15], deviations to the left in bisecting a line is seen in vast majority of normal right handed people. Only in more recent times the phenomenon was named "pseudoneglect," by Bowers and Heilman [16]. According to 1-Way callosal traffic circuitry underpinning lateralities of motor and sensory control, the reason for the left deviation in bisecting lines in visual paradigms is that all movements occurring on or towards the nondominant side of the body are bi-hemispherical events requiring callosal participation. This delay applies to all movements of or toward the nondominant side, including the saccades (gaze) or those of the diaphragms for breathing. Thus, while conjugate eye movements to the dominant side do not require callosal participation, moving them to the left demands an intact corpus callosum [8, 9, 13]. Similarly, sensing from the nondominant side of the body requires callosal participation to convey those signals arising from the nondominant side of the body that have reached the right hemisphere to the left hemisphere before they are consciously apprehended. Accordingly, left sided movements incur a delay equal to the interhemispheric transfer time (IHTT) as the motor signals move from the left to the right hemisphere and the sensory signals incur a similar delay in the opposite direction [17].

Therefore, as expected, when drawing two straight lines with both hands simultaneously, the lines drawn by the two hands are unequal, with those by the dominant hand (the side in direct contact with the major hemisphere) being longer than those by the hand ipsilateral to the command center (regardless of the behavioral handedness of the person; see under fake (ostensible)-handedness for exceptions) [9].

To recapitulate, in assessing the middle of a line, it takes longer for a right handed person to move his or her eyes to the left than moving them to the right by an amount equal to IHTT. As a result of an automatic trading of space and time an overestimation of the left side of the line will occur, giving rise to the left deviation of the marking [1]. Similarly, in the somatosensory realm, an overestimation of the size of an object occurs which is reported when judging the same size disks manipulated between our thumbs and index fingers simultaneously by both hands (with eyes closed). The disk on the left is judged as bigger due to automatic "time stamping" of motor events in measuring as the subject manipulates the disks [18, 19]. Here, therefore, the left deviation in "visual" line bisection has its "appendicular" counterpart, similar to the tau phenomenon described by Helson (1930) [20]. Turning to the abovementioned paper and pencil test, if the hands holding the pencils are moved from the side of the page to the middle, the lines drawn will meet on the left of the midline, reproducing the results of the line bisection test. Those who remain skeptical of the above explanation may take solace by attempting to draw two separate lines of the same length with each hand with their eyes closed as they count to a certain arbitrary number while drawing a line. It will again be noted that the line drawn by the neurally dominant hand is longer than that by the other hand by an amount equal to IHTT, again reflecting the aforementioned inevitable "trading of time with space" in the human mind [1, 9, 21]. Similarly, the right visual field advantage described in tachistoscopic experiments on naming latency, the wider excursion of the eyes to the right (which is based on the same anatomy just mentioned) "becomes more pronounced with the number of letters in the word" [22].

In the past, these visuo-motor asymmetries were erroneously accounted for by the faster speed of signals traveling from the minor to the major hemisphere compared to those moving in the opposite way (i.e. assuming a Newtonian division of visual half fields between the hemispheres) [23, 24]. The fact remains, however, that in right handed subjects laterality of tachistoscopically presented stimuli in Poffenberger paradigm is irrelevant to the reaction time of the dominant hand and the performances of both hands remain unchanged in response to stimuli appearing in the left visual field [23, 24].²³ Finally, the absence of a role for the corpus callosum in vision is

indicated by the fact that the asymmetries described above persist while performing the bimanual drawing task blindfolded [9, 21]. Additional observations point to the same conclusion [25-27].

Initial Visual Exploration (IVE), Optimal Viewing Position (OVP) and Point of Subjective Equality (PSE):

According to the abovementioned scheme (i.e. 1-way callosal traffic circuitry), what appears to be the center of a scene to an observer incurs a slight leftward shift compared to the veridical center of that scene. For example, in a paradigm employing targets made from texture numbers scattered to the right and left of the midline, 65 % of normal controls “started exploration in the left half of the arrays” and in a paradigm using overlapping figures, neurological control subjects (with lesions in the brainstem or below) displayed a strong tendency to identify first the parts of a composite diagram lying just to the left of the midline [2, 28]. On the other hand, in the studies mentioned, patients with right hemisphere damage displayed an initial directional bias to items ipsilateral to the damaged hemisphere, regardless of presence or absence of other signs of neglect (such as those seen in bilateral simultaneous stimulation, line cancellation and copying tasks); while measurements of the landing positions of the gaze while reading demonstrated a left of midline positioning of the gaze when viewing words (see Introduction). Lastly, using a speeded reaching task in thirteen right handed participants, Oliveira et al documented a leftward shift in their mean PSE in an experiment in which the participants chose the hand with which to respond to stimuli appearing in the right or left visual field [3]. The authors also documented faster response by the right hand to stimuli occurring in the right visual field compared to those of the left to stimuli appearing on the left side ($p=0.0499$). Nevertheless, the authors, following a conventional understanding of visual sense of space as well as that of motor control, the authors failed to provide a valid interpretation of their results (i.e. the callosum-width proximity of the dominant side of the body to the command center/macular vision in vast majority of right handed people) [9, 17, 26].

Asymmetry in Perceptual Span:

There is a substantial literature in which faster responses in moving the eyes to the right than moving them to the left, in vast majority of right handers is documented [11, 17, 22]. However, the relationship between this asymmetry to the asymmetry in visual perceptual span has never been explored [3, 11, 29-31]. Instead, the totality of the literature ascribes the right

visual field advantage (RVFA) in lexical decision and naming tasks to the “specialization” of the left hemisphere for speech, without proving comments or specifications as to the mechanism underpinning the same. In this respect, Orbach’s [31] and Bub and Lewine’s [22] articles, by demonstrating a wider perceptual span to the left in two groups of left handed participants, provide solid evidence in favor of directionality in callosal traffic by documenting that abovementioned asymmetry is indexed to a person’s laterality of motor control as espoused in this article; as does the demonstration of a faster verbal response to right sided visual stimuli in a group of right handers studied by Melamed et al who also noted that participants displayed wider excursions to the right (RVFA) in an experiments involving lexical decision task [33].

Scrotal, Galvanic Skin Response and H-reflex Asymmetries:

There are numerous autopsy reports of patients with unilateral supratentorial lesions involving the dominant hemisphere associated with bilateral Babinski signs or with bilaterally absent abdominal and cremasteric responses associated with bilateral up going toes [34-38]. Importantly, only one of the eleven patients reported in the aforementioned five articles (the case by Adams et al) involved the right hemisphere of the patients described. Thus, the knowledge that traffic between hemispheres is one-way and that all transcallosal influences are purely excitatory [39] allow the clinicians familiar with the concept of interhemispheric diaschisis (separation shock) to properly interpret the above described findings; i.e. the diaschitic paralysis of contralateral hemisphere in lesions affecting the action/major hemisphere associated with paralysis of the contralateral side of the body directly connected to the action hemisphere.

The largest series of similar cases, i.e. paralysis ipsilateral to a lesion affecting the major hemisphere, remains the classical study by Kernohan and Woltman where only one half (17 of the 35) of patients with supratentorial lesions displayed ipsilateral pyramidal signs, regardless of presence or absence of a Kernohan notch [40]. Failure to consider or understand the above-described mechanism (i.e. interhemispheric diaschisis) as the cause of ipsilateral paralysis in lesions affecting the major hemisphere has led to numerous expressions of bewilderment by some the most distinguished luminaries of clinical neurology upon confronting similar findings in their patients [38, 41], or has prevented well known clinicians from correctly interpreting the numerical results obtain in their otherwise excellent clinical research [14].

The circuitry described above also provides a plausible explanation for other laterality indexed findings

such as the asymmetrical positioning of testicles in response to gravity [42, 43] and the longer latencies of the nondominant side to electrical stimulation in Huffman and Galvanic skin response measurements [44, 45]; i.e. a callosum-width proximity of the body's dominant side to the command center located in the major hemisphere.

Fake (ostensible)-handedness:

The footprints of one-way callosal circuitry is visible in all circumstances in which laterality of motor control plays a role in our daily lives, as in dueling sports [26] and the laterality of seizure onset [46, 47]. There is a caveat, however, as follows: Statistically, it has been shown that one in five persons display a behavioral handedness opposite for which the person is wired (see above). Thus, about one half of the left handed people and 20 % of right handers are wired in the opposite direction as judged from the laterality of their speech or the speed of their movements [9, 48-50]. Neurologists have long known about this disparity starting with Bramwell's article on "crossed aphasia" in a left hander who lost his speech and became agraphic after a right sided hemiplegia [51]. According to observations supporting 1-way callosal traffic scheme, it is the higher speed of the side contralateral to the command center, relative to the side ipsilateral to the same, that unmistakably points to the laterality of motor control in any individual, regardless of his/her declared handedness [17, 52, 53]. Understandably, there are no data on manual reaction times of those who later become crossed aphasics. Nevertheless, since the incidence of crossed aphasia among the right handers is about 20 percent [48, 49] there is a likelihood of running into one or two of such person with "anomalous brain organization" in any gathering of 10 right handers; i.e. persons who react more quickly to a signal, or tap faster in a short span of time, with their left hand than their right hands despite their avowed handedness to the opposite [9, 12, 21, 50, 54-56]. Historically, the Imperial Counselor described by Liepmann [56], the famous neuroanatomist Alf Brodal, whose description of his own aphasic performance when writing with his right hand [58] and David Kinnebrook whose consistent 500-800 milliseconds tardiness in responding to events in his view finder compared to those of his superior (Nevil Maskelyne, Royal Astronomer) cost him his job as an assistant at Greenwich observatory [59, 60]. At the same time, it appears that ostensible right handers may have played a role in the early twentieth century drama unfolded after Pierre Marie's attack "on the basic tenets of Broca's aphasia," by the absence of "left third frontal lesion without l'Aphasie de Broca" [61] Finally, reference must be made to the syndromes of crossed nonaphasia [62, 63] and occurrences of right sided neglect

in ostensibly right handed subjects after insults to their left hemisphere to complete the list of behavioral surprises seen in those with "anomalous brain organization" [46, 47, 55, 62, 63].

Conclusion:

Much harm has come from the blind faith in the Newtonian hypothesis of contralateral representation of vision and its counterpart in the motor realm as opined by Valsalva. Neither the right visual field advantage in perceptual span nor the right hand advantages in timed movements can be explained by the Newtonian and Valsalva's schemes (leaving aside for now the issue of macular sparing in hemianopia). Nor can these dogmas account for the known asymmetries in initial visual search, line bisection or optimal viewing point as reviewed above. In this article I have reviewed critical studies that support the existence of 1-way callosal traffic circuitry, underpinning lateralities of motor and sensory controls as represented in our daily lives. Bimanual simultaneous drawing test reveals the laterality of motor control in all those who are able to perform the test by allowing the brain to speak for itself in a simple paper and pencil test. The test allows identification of members of that minority of humans who for over a century have wreaked havoc in our understanding of brain structures underpinning the laterality of motor control and consciousness.

Acknowledgments: This article is dedicated to the memory of my sister Farkhondeh and mother Rebecca Derakhshan, Melbourne, Australia. Their kindness and dedication was limitless.

References:

1. Derakhshan I. Overestimation of numerical distances in the left side of space. *Neurology*. 2005; 64:1822-1823.
2. Ebersbach G, Trottenberg T, Hättig H, Schelosky L, Schrag A, Poewe W. Directional bias of initial visual exploration. A symptom of neglect in Parkinson's disease. *Brain*. 1996;119:79-87.
3. Oliveira FT, Diedrichsen J, Verstynen T, Duque J, Ivry RB. Transcranial magnetic stimulation of posterior parietal cortex affects decisions of hand choice. *Proc Natl Acad Sci U S A*. 2010;107:17751-17756.
4. O'Regan JK, Lévy-Schoen A, Pynte J, Brugailière B. Convenient fixation location within isolated words of different length and structure. *J Exp Psychol Hum Percept Perform*. 1984;10: 250-257.

5. O'Regan JK, Jacobs AM. Optimal position effect in word recognition: A challenge to current theory, *J Exp Psychol Human Percept Perform.* 1992; 18: 185-197.
6. Brysbaert M, Vitu F, Schoyens W. The right visual field advantage and the optimal position viewing effect: On the relation between foveal and parafoveal word recognition *Neuropsychology.* 1996; 10: 385-395.
7. Laeng B, Park A. Handedness effects on playing a reversed or normal keyboard. *Laterality.* 1999; 4:363-377.
8. Preilowski BF. Possible contribution of the anterior forebrain commissures to bilateral motor coordination. *Neuropsychologia.* 1972;10:267-277.
9. Derakhshan I. Attentional asymmetry or laterality of motor control? Commentary on Buckingham et al. *Cortex.* 2011; 47:509-510.
10. Derakhshan I. In defense of the sinistrals: anatomy of handedness and the safety of prenatal ultrasound. *Ultrasound Obstet Gynecol.* 2003; 21:209-212.
11. Elias LJ, Bulman-Fleming MB, McManus IC. Visual temporal asymmetries are related to asymmetries in linguistic perception. *Neuropsychologia.* 1999;37:1243-1249.
12. Shen YC, Franz EA. Hemispheric competition in left-handers on bimanual reaction time tasks. *J Mot Behav.* 2005; 37:3-9.
13. Derakhshan I. How do the eyes move together? New understandings help explain eye deviations in patients with stroke. *CMAJ.* 2005 18; 172:171-173.
14. Mack L, Gonzalez-Rothi LJ, Heilman KM. Hemispheric specialization for handwriting in right handers. *Brain Cogn.* 1993; 21:80-86.
15. Hall GS, Hartwell EM. Bilateral asymmetry of function. *Mind* 9: 93-109, 1884
16. Bowers D, Heilman KM. Pseudoneglect: effects of hemispaces on a tactile line bisection task. *Neuropsychologia.* 1980;18:491-498.
17. Derakhshan I. Crossed-uncrossed difference (CUD) in a new light: anatomy of the negative CUD in Poffenberger's paradigm. *Acta Neurol Scand.* 2006; 113:203-208.
18. McPherson A, Renfrew S. Asymmetry of perception of size between the right and left hands in normal subjects. *Q J Exp Psychol.* 1953; 5: 66-74.
19. MacDonald PA, Paus T. The role of parietal cortex in awareness of self-generated movements: a transcranial magnetic stimulation study. *Cereb Cortex.* 2003; 13:962-967.
20. Helson H. The tau effect; an example of psychological relativity. *Science.* 1930; 71: 536-537.
21. Derakhshan I. Right sided weakness with right subdural hematoma: motor deafferentation of left hemisphere resulted in paralysis of the right side. *Brain Inj.* 2009;23:770-774.
22. Bub DN, Lewine J. Different modes of word recognition in the left and right visual fields. *Brain Lang.* 1988;33:161-188.
23. Bisiacchi P, Marzi CA, Nicoletti R, Carena G, Mucignat C, Tomaiuolo F. Left-right asymmetry of callosal transfer in normal human subjects. *Behav Brain Res.* 1994; 64:173-178.
24. Nougier V, Azemar G, Stein JF, Ripoll H. Covert orienting to central visual cues and sport practice: Relations in the development of visual attention. *J Exp Child Psychol.* 1992; 54: 315-333.
25. Berlucchi G, Heron W, Hyman R, Rizzolatti G, Umiltà C. Simple reaction times of ipsilateral and contralateral hand to lateralized visual stimuli. *Brain.* 1971; 94:419-430.
26. Derakhshan I. Handedness and macular vision: laterality of motor control underpins both. *Neurol Res.* 2004; 26:331-337.
27. Azémar G, Stein SF, Ripoll H. Effects of ocular dominance on eye-hand coordination in sporting duels. *Sci & Sports* 2008; 23:263-277.
28. Gainotti G, D'Erme P, Bartolomeo P. Early orientation of attention toward the half space ipsilateral to the lesion in patients with unilateral brain damage. *J Neurol Neurosurg Psychiatry.* 1991; 54:1082-1089.
29. Summers DC, Lederman SJ. Perceptual asymmetries in the somatosensory system: a dichhaptic experiment and critical review of the literature from 1929 to 1986. *Cortex.* 1990; 26:201-226.

30. Lindell AK, Nicholls ME, Kwantes PJ, Castles A. Sequential processing in hemispheric word recognition: the impact of initial letter discriminability on the OUP naming effect. *Brain Lang.* 2005; 93:160-172.
31. Rayner K, Well AD, Pollatsek A. Asymmetry of the effective visual field in reading. *Percept Psychophys.* 1980; 27:537-544.
32. Orbach J. Differential recognition of Hebrew and English words in right and left visual fields as a function of cerebral dominance and reading habits. *Neuropsychologia.* 1967; 5: 127-134.
33. Melamed F, Zaidel E. Language and task effects on lateralized word recognition. *Brain Lang.* 1993;45:70-85.
34. Allison RS, Morison JE. Cerebral vascular lesions and the tentorial pressure cone. *J Neurol Psychiatry.* 1941; 41: 1-10.
35. Riese W. Aphasia in brain tumors; its appearance in relation to the natural history of the lesion. *Confin Neurol.* 1949; 9:64-79.
36. Sherman IC, Krumholz S. A case of intradural hematoma with ipsilateral hemiplegia and ipsilateral third nerve palsy, *J Nerv Ment Dis.* 1942; 95: 176-182.
37. Peyser E, Doron Y. Ipsilateral hemiplegia in supratentorial space occupying lesions. *Int Surg.* 1966; 45:689-695.
38. Adams RD, Scully RE, Richardson EP. Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises, case 35. *N Engl J Med.* 1966; 275:325-331.
39. Lee H, Kydd RR, Lim VK, Kirk IJ, Russell BR. Effects of trifluoromethylphenylpiperazine (TFMPP) on interhemispheric communication. *Psychopharmacology (Berl).* 2011;213:707-174.
40. Derakhshan I. The Kernohan-Woltman phenomenon and laterality of motor control: fresh analysis of data in the article "Incisura of the crus due to contralateral brain tumor". *J Neurol Sci.* 2009; 287:296.
41. Haaland KY, Schaefer SY, Knight RT, Adair J, Magalhaes A, Sadek J, Sainburg RL. Ipsilesional trajectory control is related to contralesional arm paralysis after left hemisphere damage. *Exp Brain Res.* 2009;196:195-204.
42. Chang KS, Hsu FK, Chan ST, Chan YB. Scrotal asymmetry and handedness. *J Anat.* 1960; 94:543-548.
43. Bogaert AF. Genital asymmetry in men. *Hum Reprod.* 1997; 12:68-72.
44. Olex-Zarychta D, Koprowski R, Sobota G, Wróbel Z. Asymmetry of magnetic motor evoked potentials recorded in calf muscles of the dominant and non-dominant lower extremity. *Neurosci Lett.* 2009; 459:74-78.
45. Danilov A, Sandrini G, Antonaci F, Capararo M, Alfonsi E, Nappi G. Bilateral sympathetic skin response following nociceptive stimulation: study in healthy individuals. *Funct Neurol.* 1994; 9:141-151.
46. Derakhshan I. Anatomy of handedness and the laterality of seizure onset: surgical implications of new understandings in motor control. *Neurol Res.* 2005; 27:773-779.
47. Derakhshan I. Laterality of seizure onset and the simple reaction time: revamping the Poffenberger's paradigm for seizure surgery. *Neurol Res.* 2006; 28:777-784.
48. Mohr JP, Weiss GH, Caveness WF, Dillon JD, Kistler JP, Meierowksy AM, Rish BL. Language and motor disorders after penetrating head injury in Viet Nam. *Neurology.* 1980; 30:1273-1279.
49. Thomson AM, Taylor R, Whittle IR. Assessment of communication impairment and the effects of resective surgery in solitary, right-sided supratentorial intracranial tumours: a prospective study. *Br J Neurosurg.* 1998;12:423-429.
50. Buckingham G, Main JC, Carey DP. Asymmetries in motor attention during a cued bimanual reaching task: left and right handers compared. *Cortex.* 2011; 47:432-440.
51. Bramwell B. On "Crossed Aphasia" and the factors which go to determining whether the "leading" or "driving" speech centers shall be located in the left or the right hemisphere of the brain. *Lancet* 1899; 153: 1473-1479.
52. Satz P. Satz P. Correlation between assessed manual laterality and predicted speech laterality in a normal population. *Neuropsychologia* 1967; 5: 295-310.
53. Wyke M. Wyke M. Influence of direction on the rapidity of bilateral arm movements. *Neuropsychologia* 1969; 7:189-194.

54. McKeever WF, Hoff AL. Evidence of a possible isolation of left hemisphere visual and motor areas in sinistrals employing an inverted handwriting posture. *Neuropsychologia*. 1979;17: 445-455.
55. Kim M, Barrett AM, Heilman KM. Lateral asymmetries of pupillary responses. *Cortex*. 1998; 34:753-762.
56. Walsh RR, Small SL, Chen EE, Solodkin A. Network activation during bimanual movements in humans. *Neuroimage*. 2008; 43: 540-553. See page 7, EMG Results
57. Jeannerod M. The origin of voluntary action: history of a physiological concept. *C R Biol*. 2006; 329:354-362.
58. Brodal A. Self-observations and neuro-anatomical considerations after a stroke. *Brain*. 1973; 96:675-694.
59. Mollon JD, Perkins AJ. Errors of judgement at Greenwich in 1796. *Nature*. 1996; 380:101-102.
60. Derakhshan I. From Celestial to Terrestrial: A New Light on David Kinnebrook's Systematic Error of Judgment at Greenwich in 1796. 2010; BIOCOMP 2010: 666-670.
61. Mohr JP. Broca's Area and Broca's Aphasia. In H. Whitaker and H.A. Whitaker (Eds): *Studies in Neurolinguistics*, Vol 1, Academic Press, 1976 (p 218).
62. Hund-Georgiadis M, Zysset S, Weih K, Guthke T, von Cramon DY. Crossed nonaphasia in a dextral with left hemispheric lesions: a functional magnetic resonance imaging study of mirrored brain organization. *Stroke*. 2001; 32: 2703-2707.
63. Derakhshan I. Crossed nonaphasia in a dextral with left hemispheric lesions: handedness technically defined. *Stroke*. 2002; 33:1749-1750. Erratum in: *Stroke*. 2002 33: 2524.

In silico Docking Study of Active Constituents Identified in *Morinda Citrifolia* Linn as Enzyme targets of Alzheimer's Disease

J. Srikanth¹, S. Kavimani², and C. Uma Maheswara Reddy¹

¹Department of Pharmacology, Faculty of Pharmacy, Sri Ramachandra University, Porur, Chennai – 600 116, Tamil Nadu, India

²Dept of Pharmacology, Mother Theresa Post Graduate and Research Institute of Health Sciences, Puducherry - 605006, India

Abstract - Alzheimer's disease (AD) or Senile Dementia of the Alzheimer Type (SDAT) is an irreversible but progressive neurodegenerative disorder caused by the loss of neurons and synapses in the cerebral cortex and certain sub-cortical regions. Cholinesterases (ChEs) are family of enzymes that share extensive sequence homology (65%). ChEs in vertebrates have been classified into two types, acetylcholinesterase (AChE) and butyrylcholinesterase (BChE), on the basis of distinct substrate specificities and inhibitor sensitivities which serves as enzyme targets for AD. The search can be focused on plant natural products that may offer treatment for AD than currently used drugs. As an attempt to identify such natural alternates with cholinomimetic & neuroprotective activities, a set of 22 compounds identified from *Morinda citrifolia* fruit juice was docked against human AChE (PDB ID:1B41) / Butyrylcholine esterase (PDB ID: 2PM8) enzymes retrieved from protein data bank using Molegro Virtual Docker (MVD). Among the compounds analysed, five compounds, namely, (+)-3,3'-bisdemethyltanegool, 3,3'-bisdemethylpinoresinol, (-)-pinoresinol, isoamericanic acid A, quercetin are docked with a MolDock score of -124.227, -115.403, -107.812, -106.993, -106.634 respectively for AChE and (+) -3,3'-bisdemethyltanegool, (-)-pinoresinol, americanin A, Deacetylasperuloside, 3,3'-bisdemethylpinoresinol are docked with a MolDock score -132.26, -126.487, -115.81, -114.994, -109.8 respectively for BChE and all these phytoconstituents satisfies Lipinski's rule of '5' for drug likeliness property. The compounds were identified as potent and selective inhibitors of AChE/BChE compared to currently available drug molecules, tacrine, rivastigmine and huperazine A which showed inhibitory activity for AChE (MolDock score was -69.7799, -95.5779 & -72.1161) and for BChE (MolDock score was -70.3026, -91.32 & -68.5103). These phytoconstituents from *M. citrifolia* may serve as potential lead compound for developing new anti- alzheimer drug.

Keywords: *Morinda Citrifolia*, Docking, Acetylcholine esterase, Butyrylcholine esterase.

1 Introduction

Morinda citrifolia Linn (Rubiaceae) known commercially as Noni grows widely throughout the Pacific and is one of the most significant sources of traditional medicines among Pacific island societies. A number of phytoconstituents has been identified in the fruits of *Morinda citrifolia* such as Allantoin, Octanoic acid, Vanillin, n Decanoic acid, 1, 2-dihydroxy-anthraquinone, Hexoic acid, Isoscopoletin, Morindin, 1, 3-dimethoxy-anthraquinone, quercetin, scopoletin, kaempferol, Asperuloside,, americanin A, citrifolinin B, Dehydromethoxygaertneroside, (-)-pinoresinol, 3,3'-bisdemethylpinoresinol, (+)-3,3'-bisdemethyltanegool, Borreriagenin, Deacetylasperuloside, isoamericanic acid A [1-5]. Traditional synthesis of a series of new compounds utilizing combinatorial chemistry and high-throughput screening can be carried out at high cost and also are time consuming whereas on the other hand, docking various ligands to the protein of interest followed by scoring to determine the affinity of binding and to reveal the strength of interactions has become increasingly important in the contest of drug discovery. As the extracts and fruit juice of *M.citrifolia* have been shown to possess neuroprotective against alzheimer's disease in some earlier studies [6, 7], it was considered worthwhile to study the interaction of phytoconstituents identified with both AChE / BChE and compared with existing drug molecules by molecular docking studies.

2 Materials & Methods

2.1 Preparation of Ligand

We have collected the structures of phytoconstituents of *M.citrifolia* and currently available drug molecules from PubChem database (<http://pubchem.ncbi.nlm.nih.gov/>). Our AChE/ BChE inhibitor database comprises 22 bioactive compounds from *M. citrifolia*. The inhibitors were converted to .pdb format and optimized by means of ligand preparation using default settings in Molegro Virtual Docker (MVD-2010,4.2.0) [8]. The collected structures (ligands) were prepared for further studies.

2.2 Preparation of receptor

The X-ray crystal co-ordinates of AChE (PDB ID: 1B41) & BChE (PDB ID: 2PM8) were retrieved from protein data bank. Since ChEs have their crystal structure in a state that represent the pharmacological target for the development of new drugs to cure AD, these two PDBs were selected for modeling studies. It is well known that PDB files often have poor or missing assignments of explicit hydrogens, and the PDB file format cannot accommodate bond order information. Therefore, proper bonds, bond orders, hybridization and charges were assigned using the MVD. The potential binding sites of both ChE receptors were calculated using the built-in cavity detection algorithm implemented in MVD. The search space of the simulation exploited in the docking studies was studied as a subset region of 25.0 Angstroms around the active side cleft. The water molecules are also taken in to consideration and the replaceable water molecules were given a score of 0.50.

2.3 Molecular docking

2.3.1 MVDs docking search algorithms and scoring functions

Ligand docking studies were performed by MVD, which has recently been introduced and gained attention among medicinal chemists. MVD is a fast and flexible docking program that gives the most likely conformation of ligand binding to a macromolecule. MolDock software is based on a new heuristic search algorithm that combines differential evolution with a cavity prediction algorithm [9]. It has an interactive optimization technique inspired by Darwinian Evolution Theory (Evolutionary Algorithms - EA), in which a population of individuals is exposed to competitive selection that weeds out poor solutions. Recombination and mutation are used to generate new solutions. The scoring function of MolDock is based on the Piecewise Linear Potential (PLP), which is a simplified potential whose parameters are fit to protein-ligand structures and a binding data scoring function [10, 11] that is further extended in GEMDOCK (Generic Evolutionary Method for molecular DOCK) [12] with a new hydrogen bonding term and charge schemes.

2.4 Parameters for docking search algorithms

2.4.1 MolDock Optimizer

In MVD, selected parameters were used for the guided differential evolution algorithm: number of runs =5 by checking constrain poses to cavity option), population size=50, maximum interactions =2000,cross over rate=0.9,and scaling factor=0.5. A variance-based termination scheme was selected rather than root mean square deviation(RMSD).To ensure the most suitable binding mode in the binding cavity, Pose clustering was employed, which lead to multiple binding modes.

2.5 Parameters for scoring functions

2.5.1 MolDock score

They ignore-distant-atoms option was used to ignore atoms far away from the binding site. Additionally, hydrogen bond directionality was said to check whether hydrogen bonding between potential donors and acceptors can occur. The binding site on the protein was defined as extending in X, Y & Z directions around the selected cavity with a radius of 25 Angstroms.

2.6 Results & Discussions

2.6.1 Binding mode

The active site of AChE& BChE is subdivided into several subsites; the esteratic subsite, also called the catalytic triad (CT, Ser200, His440, Glu327), oxyanion hole (OH, Gly118, Gly119, Ala201), anionic subsite (AS, Trp84, Tyr121, Glu199, Gly449, Ile444), acyl binding pocket (ABP, Trp233, Phe288, Phe290, Phe292, Phe330, Phe331) and peripheral anionic subsite (PAS, Asp72, Tyr121, Ser122, Trp279, Phe331, Tyr334) are buried at the bottom of a 20 Å deep aromatic cleft. It was found out by ligand energy inspector that the phytoconstituents as well as the drug molecules were able to bind to the any one of the sub sites of AchE & BchE.

2.6.2 Predicted ADME properties

We analysed 22 physically relevant properties of bioactive compounds from *Morinda citrifolia*, among which were molecular weight, H-bond donors, H-bond acceptors and Log P (octanol/water), according to Lipinski's rule-of-five (Tables 1 & 2) by EPI suite software [13]. Lipinski's rule of 5 is a thumb to evaluate drug likeness, or determine if a chemical compound with a certain pharmacological or biological activity has properties that would make it a orally active drug in humans. The rule describes molecular properties important for a drug's pharmacokinetics in the human body, including its ADME. However, the rule does not predict if a compound is pharmacologically active. In this study, all the showed allowed

values for the properties analysed and exhibited drug-like characteristics based on Lipinski's rule-of-five. Four compounds of *Morinda citrifolia* namely Dehydromethoxygaertneroside, citrifolinin B, Asperuloside & Morindin deviate Lipinski's rule-of-five even though they had the maximum Moldock score [14, 15].

2.7 Tables

Table 1: Top 1 pose for each ligand based on Moldock score and applying Lipinski's rule of 5 on AChE (PDB ID: 1B41)

Ligand	MolDock Score	Re rank Score	H Bond	Molecular Weight [g/mol]	Log P	H-Bond Donor	H-Bond Acceptor
Dehydromethoxygaertneroside	-179.614	-109.844	-8.71295	240.21092	-0.1	5	14
Morindin	-168.25	-99.7315	-9.67862	332.30474	-0.6	8	14
Asperuloside	-135.728	-98.4434	-5.61205	286.2363	-2.4	4	11
Deacetylasperuloside	-130.287	-97.5299	-8.58053	144.21144	-3	5	10
citrifolinin B	-124.937	-85.3398	-7.6983	172.2646	-3.2	5	12
(+)-3,3'-bisdemethyltanegool	-124.227	-94.7819	-10.4588	192.16812	0.6	6	7
3,3'-bisdemethylpinoresinol	-115.403	-93.4557	-5.44479	268.26408	1.6	4	6
(-)-pinoresinol	-107.812	-83.2098	-2.5	192.16812	2.3	2	6
isoamericanoic acid A	-106.993	-84.3006	-8.97877	116.15828	1.7	3	7
quercetin	-106.634	-81.4073	-8.8934	328.31604	1.5	5	7
americanin A	-105.994	-74.0129	-7.85951	214.21516	1.7	3	6
kaempferol	-97.2375	-77.3282	-6.32952	158.11544	1.9	4	6
Rivastigmine	-95.5779	-75.4124	-1.78946	250.34	2.24	0	4
Borreriagenin	-88.1042	-71.2213	-5	152.14732	-1.5	3	5
Allantoin	-84.2071	-66.7818	-6.03864	576.50282	-2.2	4	3
n- Decanoic acid	-76.3079	-63.1102	-2.98015	414.36068	4.1	1	2
1, 2-dihydroxy-anthraquinone	-75.6724	-63.5793	-0.14404	418.34938	3.2	2	4
Huperazine A	-72.1161	-58.8041	-2.14292	242.32	1.54	0	3
Isoscopoletin	-71.1847	-60.2539	-1.41837	330.33192	1.5	1	4
1, 3-dimethoxy-anthraquinone	-70.5029	-65.5463	-2.52925	302.2357	2.8	0	4
scopoletin	-69.9682	-59.3564	-0.745131	358.38508	1.5	1	4
Tacrine	-69.7799	-64.0732	0	234.7246	2.71	2	0
Vanillin	-68.04	-56.2506	-4.89722	372.324	1.2	1	3
Octanoic acid	-66.7875	-55.739	-3.4205	564.49212	3	1	2
Hexoic acid	-60.5111	-48.9568	0	348.3472	1.9	1	2

Table 2: Top 1 pose for each ligand based on Moldock score and applying Lipinski's rule of 5 on BChE (PDB ID: 2PM8)

Ligand	MolDock Score	Re rank Score	H Bond	Molecular Weight [g/mol]	Log P	H-Bond Donor	H-Bond Acceptor
Dehydromethoxygaertneroside	-174.148	-119.883	-15.1737	576.5028	-0.1	5	14
citrifolinin B	-149.789	-93.8147	-11.8618	418.3494	-3.2	5	12
Asperuloside	-133.606	-94.1912	-8.84191	414.3607	-2.4	4	11
Morindin	-132.945	-2.96198	-9.31293	564.4921	-0.6	8	14
(+)-3,3'-bisdemethyltanegool	-132.26	-63.2503	-12.2203	348.3472	0.6	6	7
(-)-pinoresinol	-126.487	-80.7681	-5.041	358.3851	2.3	2	6
americanin A	-115.81	-86.7568	-5.72732	328.316	1.7	3	6
Deacetylasperuloside	-114.994	-81.5302	-11.4067	372.324	-3	5	10
3,3'-bisdemethylpinoresinol	-109.8	-81.0274	-13.6929	330.3319	1.6	4	6
isoamericanoic acid A	-109.506	-70.1125	-5.88322	332.3047	1.7	3	7
quercetin	-99.4821	-29.3219	-11.4941	302.2357	1.5	5	7
kaempferol	-95.065	-65.5896	-6.58358	286.2363	1.9	4	6
Rivastigmine	-91.32	-69.1692	0	250.34	2.24	0	4
Octanoic acid	-90.1963	-68.9598	-4.58812	144.2114	3	1	2
1, 2-dihydroxy-anthraquinone	-82.4872	-68.386	-6.08512	240.2109	3.2	2	4
Allantoin	-82.2958	-64.3674	-4.89054	158.1154	-2.2	4	3
Borreriagenin	-81.8557	-64.565	-5	214.2152	-1.5	3	5
1, 3-dimethoxy-anthraquinone	-80.0551	-67.1951	-0.63643	268.2641	2.8	0	4
Isoscopoletin	-78.277	-61.4975	-4.46503	192.1681	1.5	1	4
scopoletin	-78.135	-59.0394	-2.42021	192.1681	1.5	1	4
n-Decanoic acid	-75.7889	-60.7082	-2.5	172.2646	4.1	1	2
Vanillin	-71.7359	19.2429	-5.50472	152.1473	1.2	1	3
Tacrine	-70.3026	-55.7944	-1.46667	234.7246	2.71	2	0
Hexoic acid	-69.2025	-56.9533	-4.57399	116.1583	1.9	1	2
Huperazine A	-68.5103	-57.8567	-0.44966	242.32	1.54	0	3

3 Conclusions

Molecular docking studies revealed that the potential of plant phytoconstituents of *Morinda citrifolia* to inhibit ChE'S was attributable to cumulative effects of strong H₂-bonds, cationin- π , π - π interactions and hydrophobic interactions. A comparison of the docking results of selected phytoconstituents with standard drugs/molecules (Rivastigmine, Tacrine, Huperazine A) was found to have better affinity. This study has revealed the fact that herbal medicinal plants identified in Indian systems of Medicine are more efficacious compared to allopathic system of medicine but it draws back due to the difficulty in standardization and lack of literature. These modern techniques and analysis will be helpful in evaluating and documenting these herbal compounds identified in the Indian system of medicine as potent compounds for treatment for various ailments.

4 References

- [1] Wang Mian-Ying, Brett J West, C Jarakae Jensen, Diane Nowicki, Su Chen, Afak Palu, Gary Anderson. *Morinda citrifolia* (Noni): A literature review and recent advances in Noni research. *Acta Pharmacol Sin* 2002 Dec; 23 (1 2): 1127 - 1141.
- [2] Deng S, Palu K, West BJ, Su CX, Zhou BN, Jensen JC. Lipoxygenase inhibitory constituents of the fruits of noni (*Morinda citrifolia*) collected in Tahiti. *J Nat Prod.* 2007 May;70(5):859-62. Epub 2007 Mar 23.
- [3] Siddiqui BS, Sattar FA, Ahmad F, Begum S, Isolation and structural elucidation of chemical constituents from the fruits of *Morinda citrifolia* Linn. *Arch Pharm Res.* 2007 Aug;30(8):919-23.
- [4] Siddiqui BS, Sattar FA, Ahmad F, Begum S, Isolation and structure determination of two new constituents from the fruits of *Morinda citrifolia* Linn. *Nat Prod Res.* 2008;22(13):1128-36.
- [5] Lin CF, Ni CL, Huang YL, Sheu SJ, Chen CC, Lignans and anthraquinones from the fruits of *Morinda citrifolia*. *Nat Prod Res.* 2007 Nov;21(13):1199-204.
- [6] Muralidharan P, Kumar VR, Balamurugan G, Protective effect of *Morinda citrifolia* fruits on beta-amyloid (25-35) induced cognitive dysfunction in mice: an experimental and biochemical study, *Phytother Res.* 2010 Feb;24(2):252-8.
- [7] Pachauri SD, Tota S, Khandelwal K, Verma PR, Nath C, Hanif K, Shukla R, Saxena JK, Dwivedi AK. Protective effect of fruits of *Morinda citrifolia* L. on scopolamine induced memory impairment in mice: a behavioral, biochemical and cerebral blood flow study. *J Ethnopharmacol.* 2012 Jan 6;139(1):34-41. Epub 2011 Nov 15.
- [8] Thomsen R, Christensen MH: MolDock: A new technique for high-accuracy docking. *J Med Chem* 2006, 49:3315-3321.
- [9] Storn R, Price K: Differential evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces; Technical report. International Computer Science Institute: Berkley, CA; 1995.
- [10] Gehlhaar DK, Verkhivker G, Rejto PA, Fogel DB, Fogel LJ, Freer ST: Docking conformationally flexible small molecules into a protein binding site through evolutionary programming. In *Proceedings of the Fourth International Conference on Evolutionary Programming*: 1-3 March 1995; San Diego Edited by: John R McDonnell, Robert G Reynolds, David B Fogel. MIT Press; 1995:615-627.
- [11] Gehlhaar DK, Bouzida D, Rejto PA, Eds: Fully automated and rapid flexible docking of inhibitors covalently bound to serine proteases. In *Proceedings of the Seventh International Conference on Evolutionary Programming*: 25-27 March 1998; San Diego Edited by: William Porto V, Saravanan N, Donald E Waagen, Eiben AE. Springer; 1998:449-461.
- [12] Yang JM, Chen CC: GEMDOCK: A generic evolutionary method for molecular docking. *Proteins* 2004, 55:288-304.
- [13] The Estimation Programs Interface (EPI) Suite TM. Copyright 2000-2011 United States Environmental Protection Agency for EPI Suite TM and all component programs except BioHCWIN and KOAWIN.
- [14] Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 23, 1997, 3-25.
- [15] Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharm. Tox. Meth.* 44, 2000, 235-24.

Archaeopteryx Dethroned?

Jack K. Horner
 P.O. Box 266
 Los Alamos NM 87544 USA
 jhorner@cybermesa.com

Abstract

It has recently been argued that, based on a maximum parsimony analysis of a broad set of oviraptorosaur, archaeopterygid, and basal deinonychosaur morphological data, Archaeopteryx is not on the main line of avian evolution, and instead is more similar in general morphology to the oviraptorosaurs than to the archaeopterygids and basal deinonychosaurs. A Bayesian phylogenetic analysis does not sustain this view.

Keywords: Archaeopteryx, maximum parsimony Bayesian phylogenetic, paleo-ornithology

1.0 Introduction

Archaeopteryx is widely accepted as the most basal bird discovered to date, and thus has been central to our understanding of avialan origins ([8],[9]). It has recently been argued ([4]) that, based on a maximum parsimony phylogenetic assessment ([6]) of oviraptorosaur, archaeopterygid, and basal deinonychosaur morphological data, together with the discovery of a new *Archaeopteryx*-like theropod, *Xiaotingia zhengi*, *Archaeopteryx* is *not* on the main line of avian evolution, and instead is more similar in general morphology to the oviraptorosaurs than to the archaeopterygids and basal deinonychosaurs. These relationships are depicted in Figure 1.

2.0 Method

The taxon descriptors used in [5] were converted from PDF to MS-DOS text format using the *deskUNPDF Standard Version 3.1* software ([7]). The resulting text

file was reformatted under Microsoft *Notebook* to be compatible with [1]. There are 8 two-valued character-positions in among the taxon descriptors in data in [5], yielding $2^8 = 256$ distinct phylogenetic data sets. Each of these data sets was derived from the original data, using an ad hoc batch editing program written in *Mathematica* ([13]). Each data set was inserted, one data set per execution, in the data matrix block of the script schematized in Figure 2. The character-legend and references for [5] were then inserted as comments in the template shown in Figure 2. The resulting script was then executed under a Bayesian phylogenetic ([2]) software package (*MRBAYES*, [1]). The software was run on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 clocked at 2.33 GHz, with 8.00 GB RAM, under *Windows Vista Home Premium/SP2*. The above experiment was repeated with *X. zhengi* removed from the taxon set in [5], and the phylogenetic trees generated by [1] with, and without (not shown), *X. zhengi* were compared

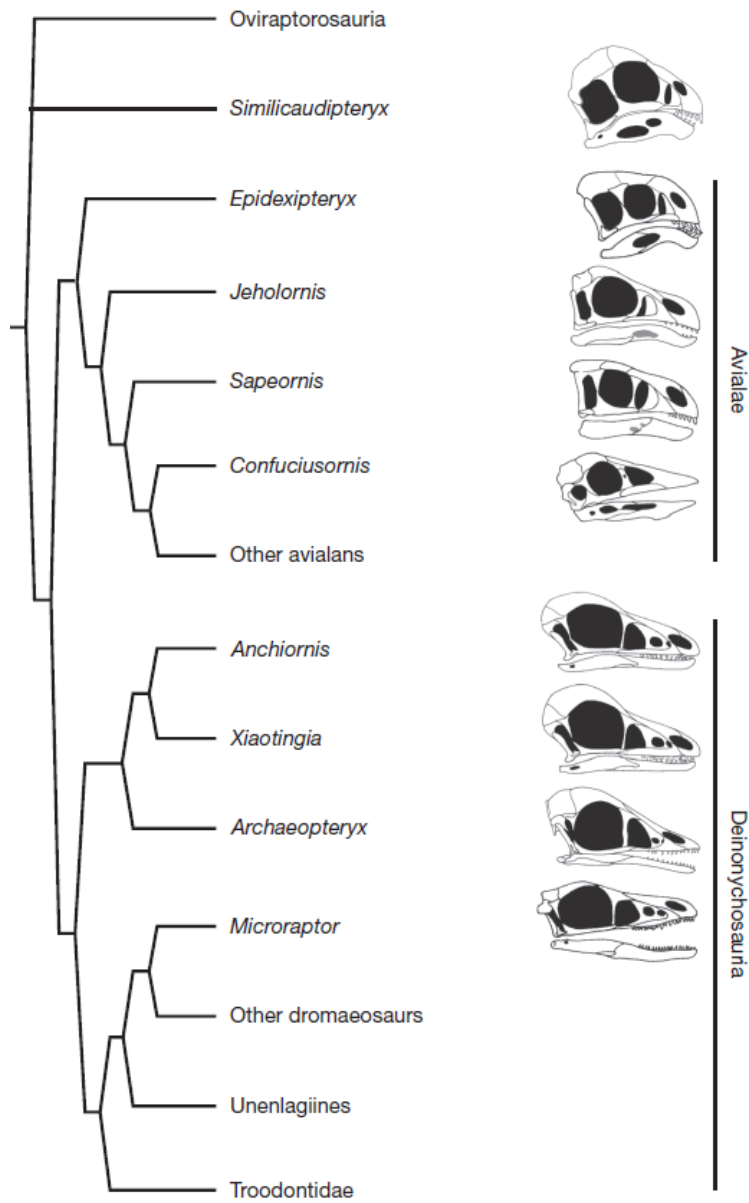


Figure 1. Simplified cladogram from [4] (p. 469) showing the systematic position of *Xiaotingia* and *Archaeopteryx*.

```

begin data;
  dimensions ntax=89 nchar=374;
  format datatype=Standard gap=- missing=?;

  matrix

[data matrices adapted from [5] go here, not shown]
;
end;

begin mrbayes;
  log start filename=archae4_log.log replace;
  set autoclose=yes;
  mcmcp nruns=2 ngen=3000000 printfreq=100
    samplefreq=100 nchains=4 savebrlens=yes
    filename=archae;
  mcmc;
  plot filename=archae.run1.p;
  plot filename=archae.run2.p;
  sumt filename=archae burnin=10000 contype=halfcompat;
  log stop;
end;

```

Figure 2. Template of the *MRBAYES* script [1]) used in this study. The script creates 3000000 (*ngen*) Markov Chain ([10]) generations, (Monte Carlo, [11]) sampling every 100 (*samplefreq*) generations. The first 10000 (*burnin*) trees are discarded. Partial tree consensus (*contype*) is allowed. For definitions of other parameters used in this script, see [1].

3.0 Results

Figure 3 (which includes *X. zhengi*) is representative of the trees output by the script shown in Figure 2. The time to

produce each tree on the platform described in Section 2.0 was about 3 hours, for a total of ~1540 hours to produce trees for the entire collection of phylogenetic data sets derived from [5].

```

/- Allosaurus_fragilis (1)
|
|- Sinraptor (2)
|
|   /- Dilong_paradoxus (3)
|   |
|   |-- Eotyrannus_lengi (4)
|   |
|   | / Tanycolagreus_topwilsoni (7)
|   | /-+
|   || \ Coelurus_fragilis (8)
|   ||
|   ||   /- Ornitholestes_hermani (9)
|   ||   |
|   ||   | /---- Falcarius_utahensis (22)

```


produced. Bayesian methods have the distinct theoretical advantage, however, that if the sample selected is large enough, the Central Limit Theorem ([12], Chap. 7) guarantees the solution based on the sample will converge to the population distribution of trees; heuristic MP cannot be guaranteed to satisfy this criterion.

5.0 Acknowledgements

This work benefited from discussions with Tony Pawlicki, with Town Peterson and Kris Krishtalka of the University of Kansas Biodiversity Institute, and with Joan Hunt of the University of Kansas Medical Center. For any problems that remain, I am solely responsible.

6.0 References

- [1] Ronquist F and Huelsenbeck JP. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19 (2003), 1572-1574.
- [2] Felsenstein J. *Inferring Phylogenies*. Sinauer Associates. 2004.
- [3] Lee MSY and Worthy TH. Likelihood reinstates Archaeopteryx as a primitive bird. *Biology Letters*. <http://dx.doi.org/10.1098/rsbl.2011.0884>. 2011.
- [4] Xu X, You H, Du K, and Han F. An *Archaeopteryx*-like theropod from China and the origin of Avialae. *Nature* 475 (28 July 2011), 465-470.
- [5] Supplementary Information for [4]. doi:10.1038/nature10288. <http://www.nature.com>.
- [6] Goloboff PA, Farris J, and Nixon KC. TNT, a free program for phylogenetic analysis. *Cladistics* 24, 774-786 (2008).
- [7] Docudesk Corporation. *Docudesk deskUNPDF Standard 3.1*. Build 3.1.111111. <http://www.docudesk.com/>. 2011.
- [8] Feduccia A. *The Origin and Evolution of Birds*. Second Edition. Yale. 1999.
- [9] Zhou Z-H. The origin and early evolution of birds: discoveries, disputes, and perspectives from fossil evidence. *Naturwissenschaften* 91 (2004), 455-471.
- [10] Gilks WR, Richardson S, and Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. Chapman and Hall. 1996.
- [11] Liu JS. *Monte Carlo Strategies in Scientific Computing*. Springer. 2001.
- [12] Chung KL. *A Course in Probability Theory*. Third Edition. Academic Press. 2001.
- [13] Wolfram Research. *Mathematica Home Edition v8.0.4*. Available at [http://www.wolfram.com/mathematica-home-edition/?src=google&129+\[mathematica+home\]&gclid=COyNhdaWvqwCFYUUbQgodFHZfoA](http://www.wolfram.com/mathematica-home-edition/?src=google&129+[mathematica+home]&gclid=COyNhdaWvqwCFYUUbQgodFHZfoA).

Bioinformatics Web Services

Mohamad Ibrahim Ladan

Computer Science Department, Haigazian University, *Beirut – LEBANON*

Abstract - *Bioinformatics is emerging as a new major or emphasis of study or work as a merge between biology and information technology majors or field of work. In addition, Web Services have emerged as a new Web-based technology paradigm for exchanging information on the Internet using platform-neutral standards, such as XML and adopting Internet-based protocols. This has helped in the birth of what is called Bioinformatics Web Services. In this paper, I will introduce bioinformatics web services, and survey the different existing tools and mechanisms available to develop such systems.*

Keywords: Bioinformatics, Web Services, Bioinformatics Web Services.

1 Introduction

These The recent advances in the field of molecular biology and genomic sequences technologies have resulted in flood of data and biological information from the research community. In order to utilize this huge volume of data in an efficient way, there was a need for the use of information technology and computerized tools to store, manage, view, index, and analyze this volume of data. This has led to the birth of what is called bioinformatics. It is a new science field in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of this field is to create a global perspective from which unifying principles in biology can be determined [1]. To accomplish this goal, there is a clear need for a technology environment or system that links together and make use of data and tools in different formats and shape found at different computers in different locations to create workflows that can be used by biologists from anywhere at anytime. Web Services technology is the right platform to be used to fulfill this requirement.

Web services represent a new programming approach based on a document-oriented model designed for interoperability at a document, typically XML, level. They are modular, self-describing, self-contained applications that are based on open standards and can be published, located, and invoked across the Internet/Web. Web services are a distributed computing technology that provides software services over the web and enable us to build Web-based applications using any platform, object model, and programming language that we may require [2]. Because of its

features, Web Services is the perfect choice for bioinformatics applications developments.

The rest of this paper is organized as follows: Section 2 introduces and discusses the bioinformatics field. Section 3 introduces the Web Services technology and environment. Section 4 introduces and discusses the Bioinformatics Web Services in general, and surveys and discusses the different types of existing Bioinformatics Web Services in particular with their benefits and shortcomings. Finally, section 5 concludes the paper. Instructions for authors

2 Bioinformatics

Please Bioinformatics is the analysis of biological information using computers and information technology. According to Oxford dictionary, bioinformatics is conceptualizing biology in terms of molecules and applying information technologies to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications. The National Center for Biotechnology Information defines bioinformatics as [3]: "Bioinformatics is the field of science in which biology, computer science, and information technology merges into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."

In the past, the main concerns of bioinformatics were storing, managing, analyzing volume of biological information, and development of complex interfaces to access this information and submit new and updated information by different researchers. In February 2001, the scientists have mapped the human genome, the complete set of genes. The process is called sequencing. It is an overwhelming process requiring complex analytical tools and techniques, and it was considered as the greatest success of bioinformatics tools [4]. With time, the bioinformatics field has evolved and is currently using different computational techniques which includes besides sequencing and structural alignment, database design and data mining, macromolecular geometry,

prediction of protein structure and function, gene finding, and expression data clustering.

In brief, the main objectives of bioinformatics can be stated as follows: The creation and maintenance of a database to store biological information, the development of complex interfaces for researchers to access and update existing data, and to develop tools and computational techniques for analyzing and interpreting the various types of data..

3 Web Services

Web Services are based on a collection of standards and protocols that allow us to make processing requests to remote systems by speaking a common, non-proprietary language and using common transport protocols such as HTTP and SMTP. Web services represent a new programming approach based on a document-oriented model designed for interoperability at a document, typically XML, level. They are modular, self-describing, self-contained applications that are based on open standards and can be published, located, and invoked across the Internet/Web. Web services enable us to build Web-based applications using any platform, object model, and programming language that we may require. In addition, they are implemented using a collection of several related, established and emerging technologies and communication protocols that include HTTP, XML, Simple Object Application Protocol (SOAP), Universal Description Discovery and Integration (UDDI), Web Services Description Language (WSDL), Common Object Request Broker Architecture (CORBA), Java Remote Method Invocation (RMI), and .NET [2].

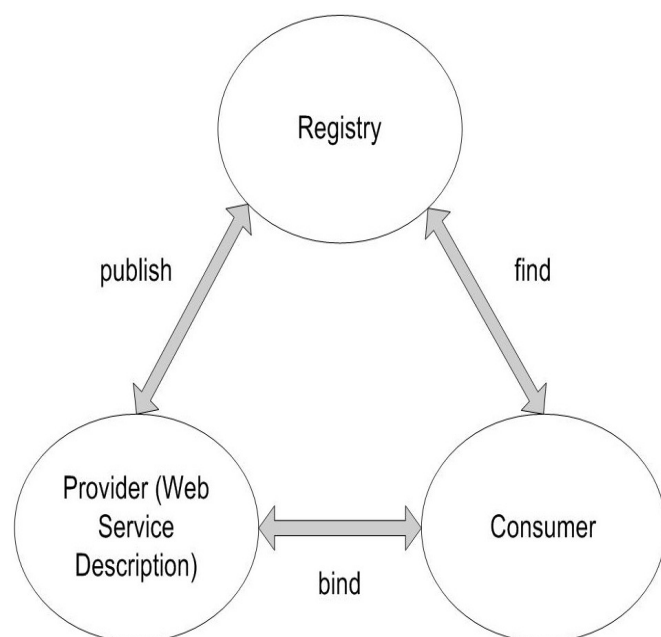


Figure 1. The web service model

The web service model consists of three entities, the service provider, the service registry and the service consumer. Figure 1 shows a graphical representation of the traditional web service model. The service provider creates or simply offers the web service. The service provider needs to describe the web service in a standard format, which in turn is XML and publish it in a central Service Registry. The service registry contains additional information about the service provider, such as address and contact of the providing company, and technical details about the service. The Service Consumer retrieves the information from the registry and uses the service description obtained to bind to and invoke the web service.

Web Services have several benefits and can offer solutions to several problems faced in bioinformatics. Web Services can make it possible for scientists to access biological data and analysis applications residing at different servers in different labs all over the world as if they were installed on their laboratory computers. In addition, Web Services can provide easier integration and interoperability between bioinformatics applications and the data they require from different locations. In the following section, I will be discussing and surveying some of the well known Bioinformatics Web Services.

4 Bioinformatics Web Services

This Web Services features and environment turned out to be the solution to some of the challenges faced in bioinformatics, in terms of integration and automation. Web Services can combine different types of bioinformatics tools available at different location on the Internet into one comprehensive set of bioinformatics services accessible from anywhere at any time. In addition, they provide easier integration and interoperability between bioinformatics applications and the data they require

Web Services technology enables scientists to access biological data and analysis applications as if they were installed on their local laboratory computers. Similarly, it enables programmers to build complex applications without the need to install and maintain the databases and analysis tools. Using Web Services users can browse various data resources and invoke analysis tools available on different computers/servers at different locations from anywhere in the world. In their simplest form, Web Services can provide a middle layer between a database and the user interface. This layer analyzes the user submitted data by intelligent computing or searching against certain databases, and finally provides user the domain knowledge as shown in Fig. 2 [5].

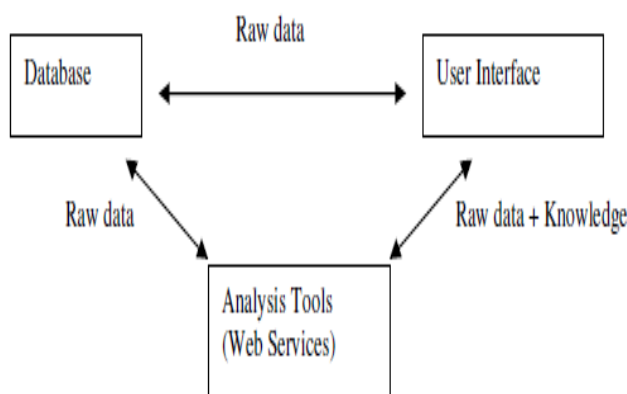


Figure 2. Web Services as a layer of data analysis.

Over the past decade many tools have been generated for the bioinformatics field; however most of these tools are web HTML forms-based tools. This is because it is easy for developers to develop an interface for their program that can be accessed using a web browser than to develop an interface for specific platform. In addition, the use of web browsers as the interface for bioinformatics services makes the development of simple graphical user interfaces relatively easy. Although these tools are very popular, they have a serious disadvantage which is the difficulty of integrating different tools and using different data from different sources to create workflows and data analysis. To overcome this difficulty, the bioinformatics community has generated several tools simplify the developing of workflows using Open Source libraries, such as BioPerl, BioJava and BioRuby [6]. These reusable procedures in different languages allow developers to develop systems for automatic generation of wrappers around web form-based tools to ease the integration of workflows and data from different sources [7, 8]. An example of a web form-based tool that does not need programming skills to use is the Sight project [9]. It is advertised as 'Automatic genomic data-mining without programming skills'. It is a web form analyzer that extracts data from a web form and presents it to the user. The user can then select the data of interest and create an agent from this selection. To create a workflow, it simply connects. However, the main disadvantage of this kind of tools is that each time a service provider updates its interface; the web form analyzer has to be used to reanalyze the interface and fix the corresponding agent.

More advanced tools are required to overcome the integration problems of web form-based tools. The introduction of Extensible Markup Language (XML) provided the solution for simplifying the application integration process. XML is a meta language that has a well-defined syntax and semantics [10]. It is used in the Web Services architecture as the format for transferring information/data between a Web Services provider application and a Web Services client application. It enables developers to separate

the content of data exposed over the Web from its presentation. More importantly, XML has been widely accepted as the universal language of choice for exchanging information over the Web and is not the proprietary product of any company. As a result, researchers in bioinformatics can develop new standards for specific functions based on XML. They can define new tags like gene names and biology-specific names and tags.. This main property and others made XML very popular in Bioinformatics Web Services.

In addition to XML, SOAP (*Simple Object Access Protocol*) gained a lot of popularity in the bioinformatics web services community. It is an XML-based protocol for exchanging information in a decentralized, distributed environment [10]. It defines a mechanism to pass commands and parameters between clients and servers. The main reason for its popularity is its simplicity in using the Hyper Text Transfer Protocol (HTTP) for transporting data as messages instead of defining any new protocols. This use of HTTP ensures that Bioinformatics Web Services provider's applications and client applications can communicate using the Internet.

The numbers of Bioinformatics Web Services being developed are increasing every day. At the time of writing this paper, the number of such services listed in the BioCatalog (<http://www.biocatalogue.org/>) is 2278 services [11]. Most of these services provide programmatic access to data sources and/or algorithmic implementations to analyze biomedical data. These data and the corresponding analysis tools are mainly accessed using browser-based interfaces. They can efficiently answer specific data extraction and analysis needs. However, biomedical problems such as characterizing a gene in terms of a sequence, its translation, expression profile, function and structure requires accessing widely distributed services, exploring and globally evaluating the numerous available data, and the integration and linking of several database information retrieval and analysis services [12]. This tedious task can be achieved using Web Services technologies.

The European Bioinformatics Institute (EBI) has been using Web Services technology to enhance and ease the use of the bioinformatics resources it provides [13, 14] Currently, the European Bioinformatics Institute provides access to more than 200 databases and to about 150 bioinformatics applications.

Some of the well known Bioinformatics Web Services include the followings:

- *ToolBus* is an integrated environment in which bioinformatics data and tools can be interoperable and accessible in an open and flexible manner [5]. It is developed at The Cyber infrastructure Group (CIG) at the Virginia Bioinformatics Institute.

- *Distributed Annotation System (DAS)* is open source software from biodas.org that provides access to complete genome annotations using a SOAP web interface [15, 16].
- *BLAST*, Basic Local Alignment Search Tool, is a Web Service family of applications that allow biologists and scientists to easily identify and find homologues of an input sequence in DNA and protein sequence libraries [17]. Many genomics laboratories provide a Web-based BLAST interface to their sequence databases for this purpose [18].
- *Pathway Database System* is an integrated system of a set of software tools for modeling, storing, analyzing, visualizing, and querying biological pathways data at different levels of genetic, molecular, and biochemical detail [19].
- *KEGG*, Kyoto Encyclopedia of Genes and Genomes, API was initiated by the Japanese human genome programme in 1995. It uses SOAP based interface to provide access to a collection of [online databases](#) dealing with genomes, [enzymatic pathways](#), and biological chemicals [20].
- *PDBML*, Protein Data Bank Markup Language, is an XML-based schema for the data in the Protein Data Bank (PDB) [21, 22]. The PDB is a repository for the 3-D structural data of large biological molecules, such as [proteins](#) and [nucleic acids](#). One of the members of the PDB organization, Protein Data Bank Japan (PDBj), has developed a tool called xPSSSS that provides a SOAPbased service to retrieve PDBML data [21].
- *MAGE-ML Server* is a tool to map proprietary database schemas for storage of microarray data into Microarray And Gene Expression Markup Language (MAGE-ML) and make them accessible using SOAP [23]. The main objective was to have a standardized Extensible Markup Language format for describing microarray experiments and their results.
- *AGML Central* provides access to databases containing proteomics information in Annotated Gel Markup Language (AGML) using a SOAP interface [24]. It is a web-based open-source public infrastructure for dissemination of two-dimensional Gel Electrophoresis (2-DE) proteomics data in AGML format. It includes a growing collection of converters from proprietary formats to AGML format. A JAVA applet visualizer was developed to visualize the AGML data with cross-reference links. In order to facilitate automated access a SOAP web service is also included in the AGML Central infrastructure.
- *EMBOSS*, European Molecular Biology Open Software Suite, is an Open Source analysis software suite that contains over 200 bioinformatics applications [25]. *Jemboss* is a graphical user interface for the *EMBOSS*, it consists of a client and server both written in Java [26]. The client communicates using SOAP with a Tomcat server that passes requests to the Jemboss server. The Jemboss server can then indirectly execute *EMBOSS* applications. This Jemboss server could easily be used to provide access via SOAP to other clients than the Jemboss GUI by describing and publishing the interface in WSDL.
- *BioMOBY* is an Open Source project that aims at providing a system for the discovery and processing of biological data using web services [27, 28]. It is emerging as the standard of fact for data exchange and web services inter-communication in bioinformatics. BioMOBY is actually two projects in one: there is Semantic MOBY (S-MOBY) and MOBY Services (MOBY-S). MOBY-S tries to solve the interoperability problem by specifying the syntax and messaging layer to link clients and service providers via information in a central registry. MOBY Services uses SOAP for communication between client, central registry and services. Semantic MOBY takes a little different approach. It tries to solve the interoperability problem by providing a way to clients and providers to describe their data and identify the data relevant to them.
- *MOWServ* is the bioinformatic platform offered by the Spanish National Institute of Bioinformatics to provide integrated access to databases and analytical tools [29]. It is a BioMoby-based web client that enables the secure and integrated analysis of data and straightforward access to databases, services and computational resources.
- *jORCA* is a desktop client aimed at facilitating seamless integration of Web Services [30]. It does so by making a uniform representation of the different web resources, supporting scalable service discovery, and automatic composition of workflows.
- *myGrid* is a project from the UK e-Science Programme funded by the Engineering and Physical Sciences Research Council (EPSRC). All myGrid components are developed in Java and its code base is available as Open Source [31]. It can access several types of services using Java and SOAP. The tool to create workflows for myGrid is called Taverna [32], which can be used to integrate several types of services including web services described by a WSDL document, SOAPlab services, and local applications. To describe a workflow, Taverna uses a custom XML-based language called simple conceptual unified flow language (Scufl).

- *caCORE* is a project developed by the National Cancer Institute Center for Bioinformatics and Information Technology (NCI CBIIT) to provide building blocks for development of interoperable information management systems and aimed at integrating bioinformatics services to support research in cancer biology and medicine [33]. It is an interconnected set of software and services. Enterprise Vocabulary Services (EVS) provide controlled vocabulary, dictionary and thesaurus services. The Cancer Data Standards Repository (caDSR) provides a metadata registry for common data elements. Cancer Bioinformatics Infrastructure Objects (caBIO) implements an object-oriented model of the biomedical domain and provides Java, Simple Object Access Protocol and HTTP-XML application programming interfaces. *caCORE* has been used to develop scientific applications that bring together data from distinct genomic and clinical science sources.
- *Mummer* is a Web service for genome wide sequence comparison to find Maximum Unique Matches between two sequences. MUMmer 3 is the latest version according to its web site: (<http://mummer.sourceforge.net/>). It is an open source project based on the mummer algorithm which is a suffix tree algorithm designed to find maximal exact matches of some minimum length between two input sequences [34]. The match lists produced by mummer can be used alone to generate alignment dot plots, or can be passed on to the clustering algorithms for the identification of longer non-exact regions of conservation. These match lists have great versatility because they contain huge amounts of information and can be passed forward to other interpretation programs for clustering, analysis and searching.

The above list is not a complete list of available Bioinformatics Web Services. I have only mentioned some of these services and tools to give an idea about the importance of this new field that combines computer science, information technology, biology, and the internet. In addition, it shows the continuous progress and advancement in this area starting from pure bioinformatics to HTML-based interfaces to bioinformatics arriving to more powerful and beneficial systems of what is called Bioinformatics Web services.

5 Conclusions

Researcher, in general, can easily publish their research results on the internet, compare their findings with others, and building on existing results to make new or more advanced progress. This is true in all fields of research, in general, and in the field of biology in particular. The value of accessing data from other institutions and the relative ease of disseminating this data has increased the opportunity for multi-institution collaborations, which produce dramatically

larger data sets than were previously available and require advanced data management techniques for full utilization.

The rapidly emerging field of bioinformatics promises to lead to advances in understanding basic biological processes and, in turn, advances in the diagnosis, treatment, and prevention of many genetic diseases. Bioinformatics has transformed the discipline of biology from a purely lab-based science to an information science as well. Increasingly, biological studies begin with a scientist conducting vast numbers of database and Web site searches to formulate specific hypotheses or to design large-scale experiments. Users can access all data and applications as if they were installed in their local machines, providing seamless integration between disparate services and allowing the construction of workflows to perform complex tasks. However, these benefits come with some unresolved difficulties.

One of these difficulties is the service quality management. Several groups might offer same type of service for redundancy or load balancing, but they may be inconsistent or out of synchronization. In other words, some of these services may be out-of-date. It is currently not possible to discover the most up-to date service. Several servers may host different versions of the database and there might be changes in the available data. To deal with such issues, information about the quality of services needs to be implemented in the tools to handle and query the service directories. As an example, BioMOBY requires web service providers to register their services in a central repository. Service providers are expected to make sure that the information for their services is kept up to date.

Another difficulty has to do with standardization. Most, if not all Bioinformatics Web Services take advantages of the extensibility of XML to define their own tags to describe biology data. This extensibility feature of XML turns out to have a fire back effect in a sense that it enables scientists to describe every piece of data in the bioinformatics domain in XML by choosing different extensions for the same type of data. The problem surfaces when linking and integrating different services to form workflows to analyze collected data, and these set of data need to be converted from one XML schema into another. Therefore there is a need for standards in the area of bioinformatics or some kind of code of conduct for service providers such as the one proposed in [35] to prevent unnecessary and inefficient conversions between different data formats or tags. Although Web Services main feature is the ability to integrate and link data and tools with different formats. But this will only be efficient if the bioinformatics developers can reach consensus on one or couple of standards to describe bioinformatics data.

6 References

- [1] A. Labarga, F. Valentin, M. Anderson and R. Lopez, Web Services at the European Bioinformatics Institute, Nucleic Acids Research, Web Server issue, Vol. 35, 2007.
- [2] Mohamad Ladan, "Web Services: Technologies and Benefits" *Journal of Communication and Computer*, ISSN 1548-7709, Number 6, Volume 7, 2010.
- [3] National Center for Biotechnology Information. <http://www.ncbi.nih.gov/>
- [4] A Science Primer. Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources. <http://www.ncbi.nlm.nih.gov>, March 29, 2004
- [5] B. Yang, J Eckart, E. Nordberg and B. Sobral. ToolBus: An Interoperable Environment for Biological Researchers, Proceedings of The 2005 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '05), p274, Las Vegas, NV, June 2005.
- [6] Open Bioinformatics Foundation. <http://www.open-bio.org/>
- [7] Rocco, D. and Critchlow, T. (2003), 'Automatic discovery and classification of bioinformatics web sources', *Bioinformatics*, Vol. 19, pp. 1927–1933.
- [8] Kossenkov, A., Manion, F. J., Korotkov, E. et al. (2003), 'ASAP: Automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database', *Bioinformatics*, Vol. 19, pp. 675–676.
- [9] Meskauskas, A., Lehmann-Horn, F. and Jurkat-Rott, K. 'Sight: Automating genomic data-mining without programming skills', *Bioinformatics*, Vol. 20, pp. 1718–1720, 2004.
- [10] Mohamad Ladan, "An Overview of XML and a Comparison with HTML and SGML", International Conference on Research Trends in Science and Technology, RTST 2002, Beirut and Byblos, Lebanon. March 4 - 6, 2002.
- [11] Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orłowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., Goble, C.A.: BioCatalogue: a universal catalogue of web services for the life sciences, *Nucl. Acids Res.*, 2010.
- [12] Marco Masseroli, Giorgio Ghisalberty, and Stefano Ceri, Bio Search Computing: Bioinformatics web service integration for data-driven answering of complex Life Science questions. *Procedia Computer Science*, Volume 4, 2011, Pages 1082-1091.
- [13] Pillai, S., Silventoinen, V., Kallio, K., Senger, M., Sobhany, S., Tate, J., Velankar, S., Golovin, A., Henrick, K. et al. (2005) SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res.*, 33, 25–28.
- [14] Harte, N., Silventoinen, V., Quevillon, E., Robinson, S., Kallio, K., Fustero, X., Patel, P., Jokinen, P. and Lopez, R. Public web-based services from the European Bioinformatics Institute. *Nucleic Acids Res.*, 32, 3–9. 2004.
- [15] Dowell, R., Jokerst, R. M., Day, A. et al. 'The distributed annotation system', *BMC Bioinformatics*, Vol. 2, p. 7. 2001.
- [16] BioDAS. <http://biodas.org/>
- [17] Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215 (3), 403–410. 1990
- [18] Gish, W. (2002). BLAST. <http://blast.wustl.edu/>
- [19] Krishnamurthy, L., Nadeau, J., Ozsoyoglu, G. et al. 'Pathways database system: An integrated system for biological pathways', *Bioinformatics*, Vol. 19, pp. 930–937. 2003.
- [20] Kawashima, S., Katayama, T., Sato, Y. and Kanehisa, M., 'KEGG API: A web service using SOAP/WSDL to access the KEGG system', *Genome Informatics*, Vol. 14, pp. 673–674. 2003.
- [21] Westbrook, J., Ito, N., Nakamura, H. et al., 'PDBML: The representation of archival macromolecular structure data in XML', *Bioinformatics*, Vol. 21, pp. 988–992. 2005.
- [22] Bernstein, F. C., Koetzle, T. F., Williams, G. J. et al., 'The Protein Data Bank: A computer-based archival file for macromolecular structures', *J. Mol. Biol.*, Vol. 112, pp. 535–542. 1977.
- [23] Tjandra, D., Wong, S., Shen, W. et al., 'An XML message broker framework for exchange and integration of microarray data', *Bioinformatics*, Vol. 19, pp. 1844–1845. 2003.
- [24] Stanislaus, R., Chen, C., Franklin, J. et al., 'AGML central: Web based gel proteomic infrastructure', *Bioinformatics*. 2005.
- [25] Rice, P., Longden, I. and Bleasby, A., 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends Genet.* Vol. 16, pp. 276–277. 2000.
- [26] Carver, T. and Bleasby, A., 'The design of Jembooss: A graphical user interface to EMBOSS', *Bioinformatics*, Vol. 19, pp. 1837–1843. 2003.

[27] Wilkinson, M. D. and Links, M., 'BioMOBY: An open source biological web services proposal', *Brief. Bioinform.*, Vol. 3, pp. 331–341. 2002.

[28] Wilkinson, D., Gessler, D., Farmer, A. and Stein, L.. 'The BioMOBY Project explores open-source, simple, extensible protocols for enabling biological database interoperability', in 'Proceedings of the Virtual Conference on Genomics and Bioinformatics', Vol. 3, pp. 17–27 , 2003.

[29] Sergio Ramírez, Antonio Muñoz-Mérida, Johan Karlsson, Maximiliano García, Antonio J. Pérez-Pulido, M. Gonzalo Claros, Oswaldo Trelles, MOWServ: a web client for integration of bioinformatic resources, *Nucleic Acids Res.*, Web Server issue 38, July 1, 2010.

[30] Martín-Requena V, Ríos J, García M, Ramírez S, Trelles O. *orca: easily integrating bioinformatics web services*. *Bioinformatics*; 26:553-559, 2010.

[31] Stevens, R. D., Robinson, A. J. and Goble, C. A., 'myGrid: Personalised bioinformatics on the information grid', *Bioinformatics*, Vol. 19. 2003.

[32] Oinn, T., Addis, M., Ferris, J. et al. , 'Taverna: A tool for the composition and enactment of bioinformatics workflows', *Bioinformatics*, Vol. 20, pp. 3045–3054. 2004.

[33] Covitz, P. A., Hartel, F., Schaefer, C. et al., 'caCORE: A common infrastructure for cancer informatics', *Bioinformatics*, Vol. 19, pp. 2404–2412. 2003.

[34] Bin Hu, Gary Xie, Chien-Chi Lo, Shawn R. Starkenburg, and Patrick S. G. Chain Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics *Briefings in Functional Genomics*, 10(6): 322-333, 2011.

[35] Stein, L., 'Creating a bioinformatics nation', *Nature*, Vol. 417, pp. k119–120. 2002.

Evaluating Spirometric Trends in Cystic Fibrosis Patients

S. Zarei*, M. Abouali*, A. Mirtar†, J. Redfield‡, D. Palmer§ and D. J. Conrad¶ P. Salamon‡,

*Computational Science Research Center, San Diego State University

†Electrical and Computer Engineering Department, University of California San Diego

§Department of Biology, San Diego State University

¶Division of Pulmonary and Critical Care Medicine, University of California San Diego

‡Department of Mathematics and Statistics, San Diego State University

Abstract—In this research, we applied both supervised and unsupervised machine learning methodologies to spirometric data from patients with cystic fibrosis (CF). We developed an ensemble of neural networks to evaluate the severity of chronic CF within an individual, given the appropriate clinical input data, and a series of reference equations to describe the CF patient's pulmonary function at different ages, heights, and sex groups in order to determine longitudinal spirometric trends. The neural networks were able to be eighty-eight percent accurate when evaluating chronic disease severity and our regression analysis revealed several trends, such as in females with CF, obstruction and functional airflow movement within the lungs generally tends to deteriorate at an accelerated rate compared to males with CF. Our findings have the potential to serve as useful reference tools to physicians in the diagnosis and treatment of cystic fibrosis.

Keywords: Cystic fibrosis, neural networks, best-fit regression analysis, spirometric trends

I. INTRODUCTION

Cystic fibrosis is an inherited chronic disease that affects the lungs, digestive system, and even the circulatory system of CF patients. Most commonly, CF is characterized by both chronic airway inflammation and recurrent infections, typically leading to permanent structural lung changes and a progressive decline in lung function. Spirometric testing, commonly used by pulmonologists to assess pulmonary function, involves taking measurements to quantify the degree of airway obstruction. Oftentimes, CF patients will go through a series of different tests and treatments in hopes of alleviating their chronic disease symptoms; this results in the accumulation of vast quantities of longitudinal spirometric data and makes this research possible. Defined as the volume of air a patient can forcibly exhale in one second, FEV1 is one of the most significant parameters obtained by spirometry, since it identifies both restrictive and obstructive respiratory symptoms. It is also a powerful predictor of increased risk of lung cancer and other obstructive lung diseases [Miller et al. 2005; Pierce 2004; Wagner et al. 2006]. All CF spirometry data was collected at the Adult Cystic Fibrosis Clinic (ACFC) and Pulmonary Function Laboratory (PFL) at the Veteran's Affairs Medical Center in La Jolla CA, and was made available by Dr. Douglas Conrad, M.D., the ACFC and PFL director.

In different research studies, pulmonary measurements have been collected from healthy individuals of both sexes across

range of ages. They provide reference equations for the different aspects of the spirometry test, including FEV1 measurements for males and females of various race/ethnic groups [Hankinson et al. 1999]. On the other hand, CF investigators do not have sufficient statistical analyses for assessing the relationship between pulmonary function outcomes and predictor parameters of interest [Edwards 2000]. The aim of this paper is to discuss some of the important features of statistical analysis on the CF patient FEV1 database. This will include grouping CF patients based on their longitudinal FEV1 data through the use of a clustering method and an evaluation of a regression analysis on the FEV1 data. Next, we find the corresponding reference equations that can be used in predicting the CF patients FEV1 value. Furthermore, we present an ensemble of artificial neural networks to predict the severity of chronic cystic fibrosis within an individual by comparing against fifty patients ranked ordinally by increasing disease severity.

II. SPIROMETRY REGRESSION AND CLUSTERING ANALYSIS

The two most important values from a spirometric test are FVC and FEV1. The forced vital capacity (FVC) indicates the maximum volume of air that can be forcibly expired from the lungs. FEV1 represents the forced expiratory volume in the first second. According to the research that was conducted by Hankinson and Odencrantz the best fit regression equation describing the lung function parameter FEV1 was in terms of age and height for different age groups in both males and females of different ethnicities. The general form of their regression equation for Caucasians is as follows:

$$FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height^2 \quad (1)$$

For research purposes, we use only the corresponding regression equation for Caucasians. This is due to the fact that of all ethnic groups, they hold the highest inherited risk for CF, where approximately 1 in every 25 Caucasians is a carrier for this recessive condition, and 1 in 2,500 are clinically affected [Tsui et al. 1997]. Table I illustrates the corresponding regression equations among healthy Caucasian individuals. In this study, we used the 2004-2009 spirometry test results of patients from the University of California San Diego Adult Cystic Fibrosis Center (UCSD-ACFC), which

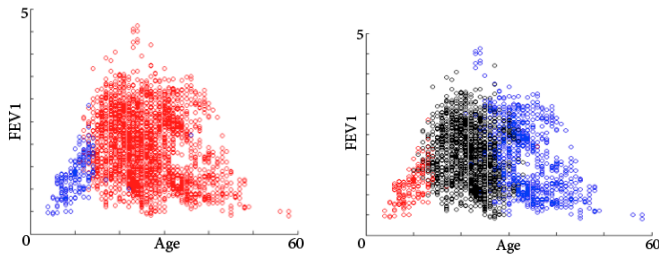


Fig. 1. Age vs. FEV1 with 2 clusters. Fig. 2. Age vs. FEV1 with 3 clusters.

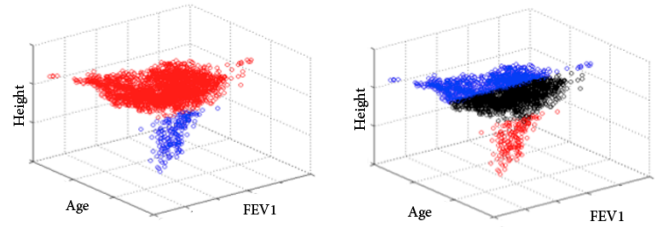


Fig. 3. Age vs. Height vs. FEV1 with 2 clusters. Fig. 4. Age vs. Height vs. FEV1 with 3 clusters.

TABLE I
FEV1 REGRESSION EQUATIONS FOR HEALTHY INDIVIDUALS

Sex	Caucasian < 20 years of age	Caucasian > 20 years of age
Female	$FEV_1 = -0.8710 + 0.06537 \times Age + 0.00011496 \times Height^2$	$FEV_1 = 0.4333 - 0.00361 \times Age - 0.000194 \times Age^2 + 0.00011496 \times Height^2$
Male	$FEV_1 = -0.7453 - 0.04106 \times Age + 0.0004477 \times Age^2 + 0.00014098 \times Height^2$	$FEV_1 = 0.5536 - 0.01303 \times Age - 0.000172 \times Age^2 + 0.00014098 \times Height^2$

includes approximately a total of 6,000 samples. Our first attempt was to find the possible clusters within our samples to help us group CF patients based on their lung function. Figures 1 through 4 are the cluster plots using the K-mean method. In data mining, K-means clustering is an algorithm for partitioning (or clustering) N data points into K disjoint subsets S_j containing N_j data points in a way that minimizes the following sum-of-squares criterion.

$$J = \sum_{j=1}^K \sum_{n \in S_j} |x_n - \mu_j|^2 \quad (2)$$

where x_n is a vector representing the n th data points, and μ_j is the geometric centroid of the data points in S_j [Bishop1995]. This will result in K clusters in which each observation belongs to the cluster with the nearest mean. As depicted in Figure (1), the FEV1 values of CF patients can be divided into two different groups one above and one below age 15. On the other hand, Figure (2) shows that when we used three clusters the additional group formed between ages 30 to 60. According to the CF foundation, more than 45% of the CF patient population is age 18 or older, additionally the predicted median age of survival for a person with CF is 37 years [Cystic Fibrosis Foundation]. Therefore, we can refer to the blue (darker) part of Figure (2) as the survival group. In Figures (3) and (4) we plot age vs height vs FEV1 values. As both figures show, age 25 is the main separation line regardless of the number of clusters. This could be due to the growth in height during that age span. In Figure (4), even though we have three clusters, two have similar heights and the one showing age 25 and under falls into a separate cluster, which is easily seen in Figure (3) with only two clusters. By clustering our CF FEV1 data we realized that as the height of the CF patients increases throughout their developing years, between the ages of 0-25, their FEV1 values increase.

III. REGRESSION ANALYSIS OF LONGITUDINAL CF FEV1

In this section, we find the best fit equation for different sex and age groups. According to Figure (1) and (2), the cut-off age should be 15, yet due to the fact that we are trying to compare lung function between healthy individuals and CF patients, we preferred to use the same age grouping as Hankinson and Odencrantz [1999]; that is before and after age 20 which is close to our cut-off age. For each group we conducted a regression analysis to identify the best-fit equation among the following four equations:

$$\text{Reg 1) } FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Height$$

$$\text{Reg 2) } FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height$$

$$\text{Reg 3) } FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height^2$$

$$\text{Reg 4) } FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height + b_4 \cdot Height^2$$

In order to find the best fit equation for each group, we considered corresponding residual plots, normal plot of residuals, coefficient of determination as well as Akaike's information criterion values. Therefore, we selected a model with the following characteristics:

Lowest RSS (Residual Sum of Square) value:

$$RSS = \sum_{i=1}^n \epsilon_i^2 \quad (3)$$

Highest coefficient of determination:

$$R^2 = 1 - \frac{RSS}{SS_{tot}} \quad \text{where} \quad SS_{tot} = \sum_i (\bar{y} - y)^2 \quad (4)$$

Lowest Akaike information criterion (AIC):

$$AIC = 2K + n \cdot \left[\ln\left(2 \cdot \pi \cdot \frac{RSS}{n}\right) + 1 \right] \quad (5)$$

where k is the number of parameter and n is the number of sample. Generally, having more parameters in a regression equation will result in a higher R^2 , and lower RSS . However, the optimal model is one consisting of only necessary parameters. Traditionally we would add a parameter to our model only if it increases the R^2 value by a minimum of 5%. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being considered to be the best. The AIC methodology attempts to find the model that best explains the data with the minimum number of free parameters. Table II displays

our regression results for females over the age of 20. Based

TABLE II
FEV1 REGRESSION VALUES FOR FEMALES OVER AGE OF 20

Model	RSS	R^2	AIC
Reg1	0.9318	0.88591	-16.6689
Reg2	0.9169	0.88774	-15.1704
Reg3	0.9137	0.88813	-15.2779
Reg4	0.6325	0.92257	-24.6821

on Table II, residual plots and normal plots show the best fitted equation for females above the age of 20 is $FEV1 = b_0 + b_1 \cdot Age + b_2 \cdot Age^2 + b_3 \cdot Height + b_4 \cdot Height^2$, since it captures the lowest AIC value, highest regression coefficient and lowest RSS values. The same model was selected as the best fit equations for females under the age of 20 as well.

Female > 20 :

$$FEV1 = 664.0178 - 0.0232 \cdot age - 0.0010 \cdot age^2 - 8.1568 \cdot Height + 0.0251 \cdot Height^2 \quad (6)$$

Female < 20 :

$$FEV1 = 0.00008 + 0.49077 \cdot age - 0.1219 \cdot age^2 + 0.02319 \cdot Height + 0.00004 \cdot Height^2 \quad (7)$$

We repeated the same regression analysis on the male sample data. Using the Table III regression results and their corresponding residual normal plots, we selected the second regression model as the best to define the FEV1 for male CF patients over, as well as under the age of 20.

TABLE III
FEV1 REGRESSION VALUES FOR MALES OVER AGE OF 20

Model	RSS	R^2	AIC
Reg1	1.5006	0.75899	-13.8053
Reg2	1.4055	0.77427	-14.4240
Reg3	1.4057	0.77423	-14.4176
Reg4	1.4029	0.77468	-12.4970

Male > 20 :

$$FEV1 = 9.6464 - 0.0634 \cdot age + 0.0004 \cdot age^2 - 0.0332 \cdot Height \quad (8)$$

Male < 20 :

$$FEV1 = -0.119 - 0.0339 \cdot age + 0.0054 \cdot age^2 - 0.0094 \cdot Height \quad (9)$$

IV. HEALTHY VS. CF FEV1 COMPARISONS

After finding the best fit equation of FEV1 for Cystic Fibrosis patients, using the same ages and heights, we found the FEV1 value for healthy individuals by using the lung function parameter equation found by Hankinson and Odencrantz [1999]. We then compared the ratio of the two regression equations. As Figures (5) and (6) depict, the FEV1 values for healthy individuals are always higher than the FEV1 of those with Cystic Fibrosis. As shown in Figure (7), average females with CF begin with almost 75 percent lung function at age 8

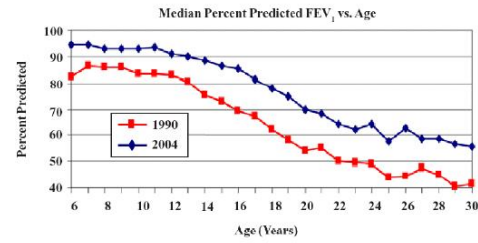


Fig. 9. Median percent predicted FEV1 vs. Age for years 1990 and 2004.

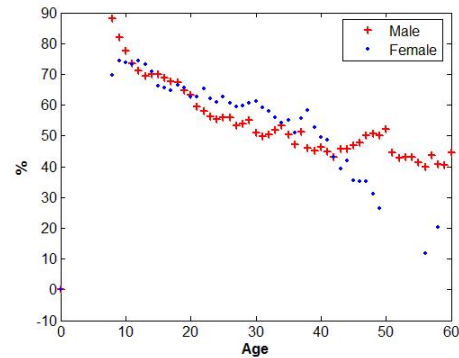


Fig. 10. Median percent predicted FEV1 vs. Age for year 2008.

and that number drops down to only 10 percent by age 60 (for the survival group). On the other hand, Figure (8) shows FEV1 values for men with CF begin with almost 90 percent lung function at age 8 and only 40 percent of lung function by age 60. Figures (9) and (10) represent the functional lung volume of CF patients as the percents of the normal lung for the years 1990, 2004 and 2008 respectively. We can see a significant improvement compared to year 1990 when at age 30, the average CF patient had only 40 percent normal lung function compared to 2004 and 2008 where this value has increased to 55 percent. This may be due to the advanced treatments developed since 1990 that help CF patients to control the progress of the disease.

V. ARTIFICIAL NEURAL NETWORKS

In this section we develop an ensemble of artificial neural networks (ANNs) to predict chronic disease severity within cystic fibrosis patients as an experienced pulmonary physician would. ANNs are a form of machine learning algorithm based on the functionality and structure of a biological neural network, as observed in the brain. Used to observe complex trends and patterns in a set of data, they are capable of applying sets of non-linear equations to inputs to achieve a desired outcome. These equations can be used to apply to further data. The data was collected from the Adult Cystic Fibrosis Clinic (ACFC) and Pulmonary Function Laboratory (PFL) at the Veteran's Affairs Medical Center. Fifty patients were selected and ordinally ranked by Dr. Douglas J. Conrad, director at the ACFC and PFL, in order of increasing disease severity ranking 1 to 50 as a training dataset for the ANNs. The 50 patients and their corresponding 14 variables were compiled as a matrix, along with their actual rankings, and

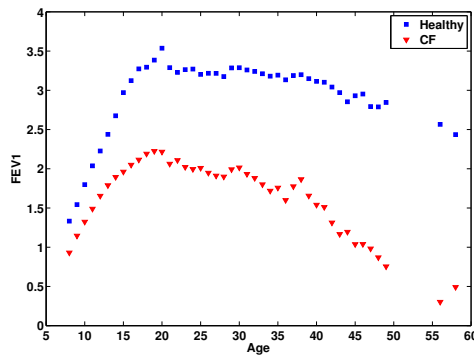


Fig. 5. CF FEV1 compare to healthy FEV1 (female).

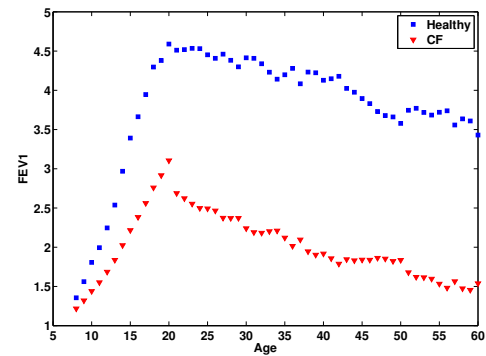


Fig. 6. CF FEV1 compare to healthy FEV1 (male).

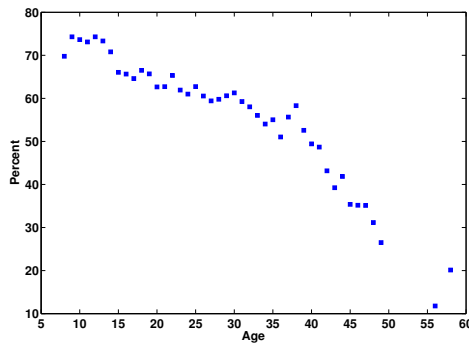


Fig. 7. Functional lung volume of CF patients as a percent of the healthy lung (female).

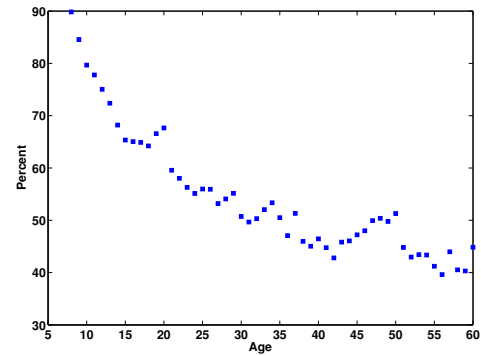


Fig. 8. Functional lung volume of CF patients as a percent of the healthy lung (male).

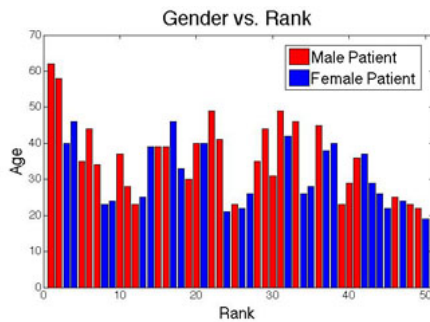


Fig. 11. Age vs. Rank vs. Gender for the 50 patient data set

imported to Matlab. Such variables included results from lung function tests (FEV1, FVC, FEV1/FVC), physical descriptions (age, height, weight, gender, BMI) and longitudinal regression values based on FEV1 vs. time graphs (m, b, r^2, se_m, se_b). For each patient, only the best FEV1 value from the previous year was considered.

VI. ANN TRAINING

To obtain the artificial neural networks, a progression of training, validating and testing steps were taken to develop the ability to predict with an acceptable amount of error. In training, each network is supplied with a set of data as inputs, and through a series of equations, returns an answer. Once the tested outputs of the network accurately reflect the

answers provided in the training data, the ANNs can then be used to classify future data. For an ANN to follow the trends in cystic fibrosis data, the variables were run through a series of equations, deemed “layers.” Four matrices of random numbers were generated, two representing weights and two being biases. Let w_1 and w_2 denote the weight matrices, b_1 and b_2 the biases, and “in” represents the vector of one patient’s inputs. The layout of a single-hidden layer ANN is as follows:

$$\text{hidden layer} = \text{squash}((in \times w_1) - b_1) \quad (10)$$

$$\text{CF severity} = (\text{hidden layer}) \times w_2 - b_2 \quad (11)$$

$$\text{squash}(x) = \frac{1}{1 + e^{-x}} \quad (12)$$

Once the severity has been run through the above layers for each patient, the ANN returns its predicted severity, and the error is calculated between its prediction and the actual answer. The weights and biases are adjusted using Matlab’s `fminsearch` optimization function, and after each adjustment, the squared error is recalculated between the ANN outputs and the actual provided patient severity. `fminsearch` is set to repeat these adjustments until a minimum in the total calculated error is found. However, if the entire set of 50 patients were to be used in training a network, there would be no unknowns upon which to test its accuracy. For this reason, only thirty of the fifty

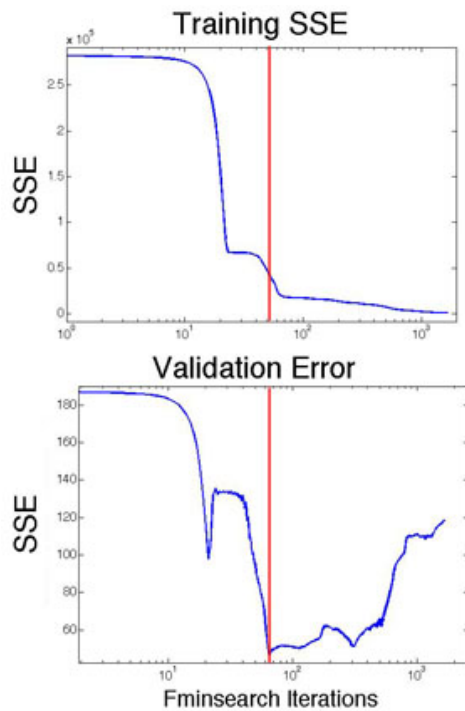


Fig. 12. Observed SSE during Training and Validation

patients, or 3/5 of the original data set, were randomly selected and used to train each ANN. The remaining 20 patients are randomized and split evenly into validation and testing groups. In validation, the purpose is to halt the fminsearch function once a network begins to over-train. Between the fminsearch iterations of the training set, the squared error is calculated and recorded for the ten validation patients. Once the weights and biases have become overly specific for the training set, the validation error will increase and halt the network training, as shown in Figure (12). The test set for the network is recorded along with the adjusted parameters.

VII. INITIAL ANN TESTING

A set of twenty networks' parameters and test sets was compiled. Each patient defined to be in a test set was run through the layers using the corresponding network's weights and biases, and averaged with the other participating networks. Thus, each ANN only "voted" on the severity of inputs that were not used in its training or validation processes. The averaged output consists of fifty patients, to be compared against the actual severities provided on the original fifty-patient data sheet (Figure (14)).

VIII. SUBSEQUENT INPUTS

Following the testing of the original networks, the importance of the 14 training variables was determined. Using the computational program R and the randomForest toolbox, FVC, FEV1, and obstruction ratio were found to hold the strongest predictive power for CF severity (Figure (13)). A new ensemble of 50 ANNs were trained and tested using only the FVC, FEV1, and obstruction ratio as training features (see Figure 15). A new dataset was then assembled including

First Inputs	Node Purity	Second Inputs	Node Purity
FVC	3018.8	Multiproduct	2308.5
FEV1	1960.8	FEV1	1217.0
Obstruction Ratio	1086.6	FVC	960.5
Age	616.3	PowerProduct	832.5
BMI	593.3	Obstruction Ratio	622.3
Weight	452.0	Brasfield	563.9
m	381.0	Cystic	481.3
r ²	377.9	Age	456.1
b	299.5	BMI	293.1
Patient ID	290.1	Overall	254.1
Height	258.6	Patient ID	160.1
sey	219.2	Height	145.4
seb	201.0	Linear	131.4
sem	155.7	Exp	73.1
Gender	21.5	Gender	16.5
		LgLS	13.2

Fig. 13. Inputs used in both matrices

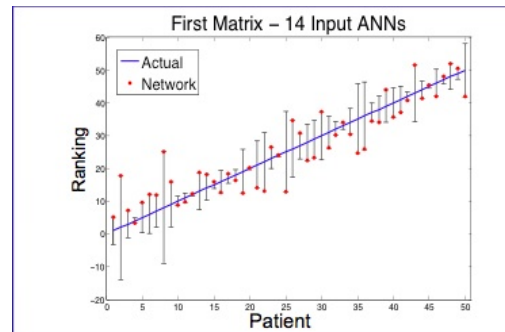


Fig. 14. Severity prediction from ANNs

several new inputs. Multiproduct and powerproduct attempt to place emphasis on age, and were generated from the equations:

$$\text{multiproduct} = \text{Age} \times \text{FEV1}\% \quad (13)$$

$$\text{powerproduct} = \text{FEV1}\% \times e^{\frac{\text{Age}}{10}} \quad (14)$$

Other variables included the Brasfield score and its components. The randomForest toolbox predicted multiproduct, FEV1, FVC, powerproduct, obstruction ratio, and the overall Brasfield score as the most important variables of the new set (Figure (13)). An ensemble of ANNs were trained from these six variables and tested for their performance (Figure (16)). For each of the 3 sets of inputs, the R^2 values and ranking accuracies were calculated, shown in Table IV. The ranking accuracy is defined by the ability of the ANNs to identify the more severe case of CF for any two patients.

TABLE IV
RESULTS OF ANN VOTING

Dataset	Inputs	Train Time (min)	R^2	Ranking Accuracy %
1	14	360	0.8261	86.67
1	3	10	0.7845	85.14
2	6	30	0.8109	88.48

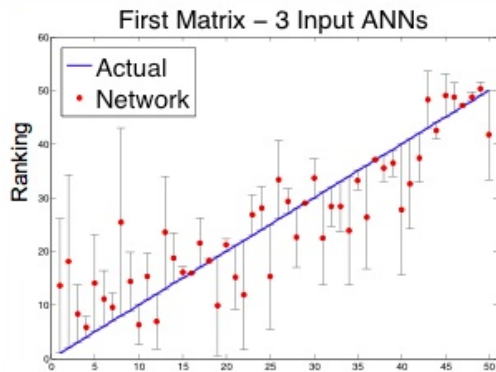


Fig. 15. Severity prediction from ANNS

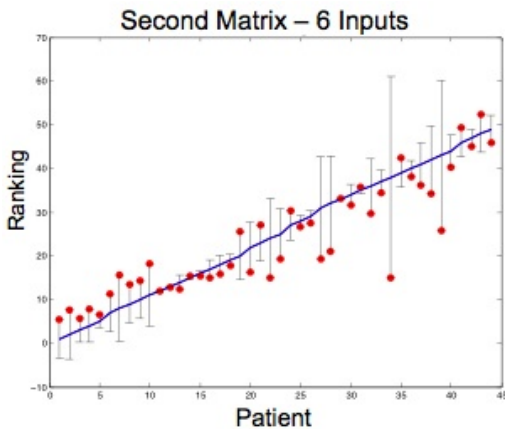


Fig. 16. Severity prediction from ANNS

IX. CONCLUSION

In conclusion, we were able to find the best fit equations for identifying lung function parameters (FEV1) for both male and female among CF patients. Using these equations we were able to see the overall trend of reduction in their lung function as a over time. Females' FEV1 and FVC values decline faster than males when afflicted with CF. The overall trends of CF lung function have improved due to advanced treatments discovered in more recent years. Using our reference equations, clinicians can predict CF patients' FEV1 and use it as a reference tool for evaluating their treatments. Ensembles of neural networks were able to be trained from the provided inputs and accurately vote upon unseen CF patients. The variables FEV1, FVC, and obstruction ratio appeared to hold the greatest ability to train the ANNs from the original list of inputs. Of the second list of inputs, the Brasfield Index, multiproduct, and powerproduct were also found to be useful in ANN training. The patient data provided has shown potential leeway for training ANNs to perform other medical analysis, such as predicting CF exacerbations per year or the severity of a given patient in five years. These ANNs can be programmed into a GUI available for practitioner use.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Institutes of Health under Grant No. ECF-56586B and by NSF under Grant No. UBM-0827278. The author wishes to

acknowledge the helpful comments and suggestions made by Rojeen Zarei. We would like to also thank the SDSU UBM and Cystic Fibrosis Groups for their guidance and many helpful discussions.

REFERENCES

- [1] Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford, England: Oxford University Press, 1995.
- [2] Cystic Fibrosis Foundation. [Online] [Cited: May 27, 2012.] <http://www.cff.org>.
- [3] Edwards, L. J. *Modern statistical techniques for the analysis of longitudinal data in biomedical research* 30: 330-344, 2000.
- [4] Eigen, H., H. Bieler and D. Grant, Spirometric pulmonary function in healthy preschool children. *Am J Respir Crit Care Med* 163: 619-623, 2001.
- [5] Glindmeyer, H. W., J. J. Lefante, C. McColloster, R. N. Jones, H. Weill. Blue-collar normative spirometric values for Caucasian and African-American men and women aged 18 to 65. *Am J Respir Crit Care Med* 151: 412-422, 1995.
- [6] Glenny, R. W., S. L. Bernard, and H. T. Robertson. Pulmonary blood flow remains fractal down to the level of gas exchange. *J Applied Physiology* 89: 742-748, 2000.
- [7] Hankinson, J. L., J. R. Odencrantz and K.B. Fedan. Spirometric Reference Values from a Sample of the General U.S Population. *Am. Jr. Respi. Crit. Care Me* 159: 179-187, 1999.
- [8] Miller, M. R., J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C.P.M. van der Grinten, P. Gustafsson, R. Jensen, D. C. Johnson, N. MacIntyre, R. McKay, D. Navajas, O.F. Pedersen, R. Pellegrino, G. Viegi, J. Wanger. Standardisation of spirometry. *European Respiratory Journal* 26: 319-338, 2005.
- [9] Nelson, S. B., R. M. Gardner, R. O. Crapo and R. L. Jensen. Performance evaluation of contemporary spirometers. *Chest* 97: 288-297, 1990.
- [10] O'Donnell, D. E., M. Lam, K. A. Webb, M. Lam and K. A. Webb. Spirometric correlates of improvement in exercise performance after cholinergic therapy in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 160: 524-549, 1999.
- [11] Pierce, R. Spirometer: An essential clinical measurement. *Australian Family Physician*: 34, 535-539, 2004.
- [12] Smith, J. J., S.M. Travis, E.P. Greenberg and M.J. Welsh. Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. *Cell* 85: 229-236, 1996.
- [13] Tsui, L. C., P. Durie. Genotype and phenotype in cystic fibrosis, *Hosp Prac* 32: 115-142, 1997.
- [14] Wagner, N. L., W. S. Beckett and Steinberg, R. Using Spirometry results in occupational medicine and research: common errors and good practice in statistical analysis and reporting. *Indian Journal of Occupational Environmental Medicine* 10: 5-10, 2006.

Computer Algebra in Pharmaceutical Engineering

Juan Carlos Guerrero Sierra
Logic and Computation Group
Physics Engineering Program
School of the Science and Humanities
EAFIT University
Medellin, Colombia
jguerre7@eafit.edu.co

Abstract

A cylindrical drug delivery device was analyzed using a Laplace transform-based method. The two-dimensional model represented a pharmaceutical agent uniformly distributed in a polymeric matrix, which was surrounded by an impermeable layer. Molecules could only be transferred through a thin ring located on the lateral surface of the device. A closed-form solution was obtained to help study the effects of design parameters and geometries on the cumulative amount of drug released. The latter variable increased with the mass transfer and diffusion coefficients and decreased with any increment in the device's length. The delivery rate was described by an effective time constant calculated from Laplace transforms. Reducing the height of thin ring would delay transport of the medication. Simplified expressions for the release profile and the time constant were derived for special design cases.

Keywords: Mathematical model; Diffusion; Drug transport; Cylindrical matrix device; Laplace transforms; Residue theorem; Effective time constant; Computational drug discovery; Biomedical engineering.

1. Introduction

At present the pharmaceutical industry seeks permanent improvement of methods of supply of medicines through computational methods that can contribute to the development of new technologies associated with mathematical models [1,2,3]. Recently computer algebra methods [4,6] have been applied to predict the evolution of the profiles of the active agents both in vivo and in vitro situations. In the present work is performed the analysis of a cylindrical device for delivering pharmaceutical agents using the method of the Laplace transform. The cylindrical device is a polymeric matrix for which the transfer of the active agent will be done only by a thin ring located on the side surface of the device, which will allow the controlled passage of the pharmaceutical agent.

2. Problem

The system studied is a cylindrical monolithic structure inside of which an active agent, A , is dissolved or dispersed uniformly. The matrix device is covered by an impermeable coating substance. The drug is released through a thin ring located on the lateral surface of the device such as is showed in Fig. 1.

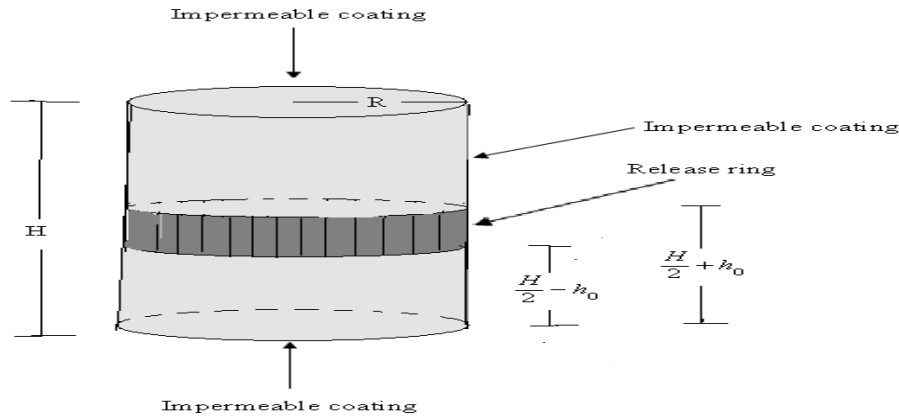


Fig 1. Geometry and Design of the Cylindrical Drug Delivery Device with Radial Discharge through a Thin Ring

A balance on component A is given by the diffusion equation which takes the following form in cylindrical coordinates:

$$\frac{\partial}{\partial t} C_A(t, r, z) = \frac{\eta_A \left(\left(\frac{\partial}{\partial r} C_A(t, r, z) \right) + r \left(\frac{\partial^2}{\partial r^2} C_A(t, r, z) \right) \right)}{r} + \eta_A \left(\frac{\partial^2}{\partial z^2} C_A(t, r, z) \right) \quad (1)$$

where $C_A(t, r, z)$ denotes the concentration of A located at the point with coordinates (r, z) and η_A is the drug diffusion coefficient in the matrix. The initial condition for (1) is:

$$C_A(0, r, z) = c_{AS} \quad (2)$$

where c_{AS} is the saturated concentration of A in the matrix. The boundary conditions are

$$\left(\frac{\partial}{\partial z} C_A(t, r, z) \right) \Big|_{z=0} = 0 \quad (3), \quad \left(\frac{\partial}{\partial z} C_A(t, r, z) \right) \Big|_{z=H} = 0 \quad (4);$$

$$\left(\frac{\partial}{\partial r} C_A(t, r, z) \right) \Big|_{r=R} = \begin{cases} -\frac{k_m C_A(t, R, z)}{\eta_A} & z \leq \frac{H}{2} + h_0 \text{ and } \frac{H}{2} - h_0 \leq z \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

These boundary conditions Eqs. (3)-(5) correspond to the previously described situation according to which the drug is released only through the thin ring of radius R located between the planes $z = H/2 - h_0$ and $z = H/2 + h_0$. The parameter k_m is a boundary-layer mass transfer coefficient. A high k_m is indicative of a low mass transfer resistance due to factors, such as vigorous mixing, that can reduce the thickness of the layer. Low k_m values would decrease the rate at which drugs leave the device.

The equations (1)-(5) can be written in the following dimensionless form:

$$\frac{\partial}{\partial \tau} C(\tau, \rho, \zeta) = \frac{\partial}{\partial \rho} C(\tau, \rho, \zeta) + \left(\frac{\partial^2}{\partial \rho^2} C(\tau, \rho, \zeta) \right) + \left(\frac{\partial^2}{\partial \zeta^2} C(\tau, \rho, \zeta) \right) \quad (6)$$

$$C(0, \rho, \zeta) = 1 \quad (7), \quad \left(\frac{\partial}{\partial \zeta} C(\tau, \rho, \zeta) \right) \Big|_{\zeta=0} = 0 \quad (8), \quad \left(\frac{\partial}{\partial \zeta} C(\tau, \rho, \zeta) \right) \Big|_{\zeta=\frac{H}{R}} = 0 \quad (9);$$

$$\left(\frac{\partial}{\partial \rho} C(\tau, \rho, \zeta)\right)\Big|_{\rho=1} = \begin{cases} -Sh C(\tau, 1, \zeta) & \zeta \leq \frac{H}{2R} + \frac{h_0}{R} \text{ and } \frac{H}{2R} - \frac{h_0}{R} \leq \zeta \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

The dimensionless magnitudes which appear in the equations (6)-(10) are defined as follows:

$$\rho = \frac{r}{R}, \zeta = \frac{z}{R}, \tau = \frac{t \eta_A}{R^2}, C(\tau, \rho, z) = \frac{C_A(\tau, \rho, z)}{c_{AS}}, Sh = \frac{R k_m}{\eta_A} \quad (11)$$

The Sherwood number Sh is a dimensionless number that represents the ratio of convective to diffusive mass transfer. This parameter is directly proportional to k_m and inversely proportional to η_A .

3. Analytical Solution

The analytical solution to the dimensionless problem, Eqs. (6)-(10), can be derived using the Laplace transform technique with the Bromwich integral and the residue theorem. The procedure is implemented in Maple (Waterloo Software Inc.). Taking the Laplace transform of Eq. (6) gives

$$s\bar{C}(s, \rho, \zeta) - 1 = \frac{\partial \bar{C}(s, \rho, \zeta)}{\partial \rho} + \frac{\partial^2 \bar{C}(s, \rho, \zeta)}{\partial \rho^2} + \frac{\partial^2 \bar{C}(s, \rho, \zeta)}{\partial \zeta^2} \quad (12)$$

after using the initial condition. The general solution of Eq. (12) is

$$\bar{C}(s, \rho, \zeta) = \left[C_1 \sin(\sqrt{-s+c_1}\zeta) + C_2 \cos(\sqrt{-s+c_1}\zeta) \right] \left[C_3 J_0(\sqrt{-c_1}\rho) + C_4 Y_0(\sqrt{-c_1}\rho) \right] + \frac{1}{s} \quad (13)$$

where J_0 and Y_0 are Bessel functions of the first kind and second kind, respectively; C_1, C_2, C_3, C_4 and c_1 are constants to be determined using the boundary conditions. Because $Y_0(x)$ is singular around $x = 0$, the constant C_4 should be zero for a finite solution at $\rho = 0$. Equation (13) becomes:

$$\bar{C}(s, \rho, \zeta) = \left[C_1 \sin(\sqrt{-s+c_1}\zeta) + C_2 \cos(\sqrt{-s+c_1}\zeta) \right] C_3 J_0(\sqrt{-c_1}\rho) + \frac{1}{s} \quad (14)$$

Without any loss of generality, C_3 is set equal to 1, which yields:

$$\bar{C}(s, \rho, \zeta) = \left[C_1 \sin(\sqrt{-s+c_1}\zeta) + C_2 \cos(\sqrt{-s+c_1}\zeta) \right] J_0(\sqrt{-c_1}\rho) + \frac{1}{s} \quad (15)$$

After applying Eq. (8) and Eq. (9), we obtain

$$C(s, \rho, \zeta) = C_2 \cos\left(\frac{n \pi R \zeta}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2} \rho}{H}\right) + \frac{1}{s} \quad (16)$$

where n is an integer from 0 to ∞ . Application of the superposition principle results in

$$C(s, \rho, \zeta) = A_0 I_0(\sqrt{s} \rho) + \frac{1}{s} + \left(\sum_{n=1}^{\infty} A_n \cos\left(\frac{n \pi R \zeta}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2} \rho}{H}\right) \right) \quad (17)$$

Now, applying the boundary condition given by Eq. (10) to Eq. (17), the following equality is obtained:

$$A_0 I_1(\sqrt{s}) \sqrt{s} + \left(\sum_{n=1}^{\infty} \frac{A_n \cos\left(\frac{n \pi R \zeta}{H}\right) I_1\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2}}{H}\right) \sqrt{n^2 \pi^2 R^2 + s H^2}}{H} \right) =$$

$$\begin{cases} -Sh \left(A_0 I_0(\sqrt{s}) + \frac{1}{s} + \left(\sum_{n=1}^{\infty} A_n \cos\left(\frac{n \pi R \zeta}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2}}{H}\right) \right) \right) & \zeta \leq \frac{H}{2R} + \frac{h_0}{R} \text{ and } \frac{H}{2R} - \frac{h_0}{R} \leq \zeta \\ 0 & \text{otherwise} \end{cases}$$

(18)

Now, integrating with respect to ζ both sides of Eq. (18) from 0 to H/R leads to

$$\frac{A_0 I_1(\sqrt{s}) \sqrt{s} H}{R} = - \frac{2 Sh \left(A_0 I_0(\sqrt{s}) + \frac{1}{s} \right) h_0}{R}$$

$$- Sh \left(\sum_{n=1}^{\infty} \left(\frac{2 A_n H \cos\left(\frac{n \pi}{2}\right) \sin\left(\frac{n \pi h_0}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2}}{H}\right)}{n \pi R} \right) \right)$$

(19)

Isolating A_0 from (19) we obtain

$$A_0 = \frac{-Sh \left(\sum_{n=1}^{\infty} \left(\frac{2 A_n H \cos\left(\frac{n \pi}{2}\right) \sin\left(\frac{n \pi h_0}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2}}{H}\right)}{n \pi R} \right) \right) s R - 2 Sh h_0}{I_1(\sqrt{s}) s^{(3/2)} H + 2 Sh h_0 I_0(\sqrt{s}) s}$$

(20)

On the other hand, multiplying both sides of Eq. (18) by $\cos\left(\frac{m \pi R \zeta}{H}\right)$ and integrating the results from 0 to H/R lead to the following system composed of an infinite number of equations:

$$\frac{1}{2} \frac{A_m I_1\left(\frac{\sqrt{m^2 \pi^2 R^2 + s H^2}}{H}\right) \sqrt{m^2 \pi^2 R^2 + s H^2}}{R} =$$

$$\frac{2 Sh H \cos\left(\frac{m \pi}{2}\right) \sin\left(\frac{m \pi h_0}{H}\right) (A_0 I_0(\sqrt{s}) s + 1)}{s m \pi R}$$

$$- Sh \left(\sum_{n=1}^{\infty} A_n \int_{\frac{H}{2R} - \frac{h_0}{R}}^{\frac{H}{2R} + \frac{h_0}{R}} \cos\left(\frac{n \pi R \zeta}{H}\right) \cos\left(\frac{m \pi R \zeta}{H}\right) d\zeta I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2}}{H}\right) \right)$$

(21)

which can be solved for A_n when Eq. (20) is replaced in Eq. (21). The subscript m is an integer that varies from 1 to ∞ . Expressions for A_0 and the first M A_i coefficients (i.e., $i = 1, \dots, M$) can be replaced in Eq. (17) to yield the

transform $\bar{C}(s, \rho, \zeta)$. The parameter M is selected to achieve a desired degree of accuracy. A formal expression for the dimensionless concentration in the time domain can be derived by taking the inverse Laplace transform of $\bar{C}(s, \rho, \zeta)$ using the Bromwich integral and the residue theorem. The inverse Laplace transform of Eq. (17) is given by

$$C(\tau, \rho, \zeta) = C_1(\tau, \rho, \zeta) + C_2(\tau, \rho, \zeta) + C_3(\tau, \rho, \zeta) + C_4(\tau, \rho, \zeta) \tag{22}$$

where

$$C_1(\tau, \rho, \zeta) = \sum_{p=1}^{\infty} \left(\frac{2 Sh \sum_{n=1}^{\infty} \left(\frac{2 P_n(-\alpha_p^2) H \cos\left(\frac{n\pi}{2}\right) \sin\left(\frac{n\pi h_0}{H}\right) I_0\left(\frac{\sqrt{-(\alpha_p H - n\pi R)(\alpha_p H + n\pi R)}}{H}\right)}{Q_n(-\alpha_p^2) n\pi R} \right)}{J_0(\alpha_p)(\alpha_p^2 H^2 + 4 Sh^2 h_0^2)} \right) \alpha_p^2 R J_0(\alpha_p \rho) e^{(-\alpha_p^2 \tau) H} \tag{23}$$

$$C_2(\tau, \rho, z) = \sum_{q=1}^{\infty} \left(\frac{Sh \sum_{n=1}^{\infty} \left(\frac{2 P_n(S_q) H \cos\left(\frac{n\pi}{2}\right) \sin\left(\frac{n\pi h_0}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + S_q H^2}}{H}\right)}{\left(\frac{d}{d\sigma} Q_n(\sigma)\right) \Big|_{\sigma=S_q} n\pi R} \right)}{I_1(\sqrt{S_q}) S_q \left(\frac{3}{2}\right) H + 2 Sh h_0 I_0(\sqrt{S_q}) S_q} \right) S_q R I_0(\sqrt{S_q} \rho) e^{(S_q \tau)} \tag{24}$$

$$C_3(\tau, \rho, \zeta) = \sum_{p=1}^{\infty} \left(\frac{4 Sh h_0 J_0(\alpha_p \rho) e^{(-\alpha_p^2 \tau) H}}{J_0(\alpha_p)(\alpha_p^2 H^2 + 4 Sh^2 h_0^2)} \right) \tag{25}$$

$$C_4(\tau, \rho, \zeta) = \sum_{q=1}^{\infty} \left(\sum_{n=1}^{\infty} \frac{I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2} \rho}{H}\right) P_n(s) \cos\left(\frac{n\pi R \zeta}{H}\right) e^{(s \tau)}}{\left. \frac{d}{ds} Q_n(s) \right|_{s=S_q}} \right) \tag{26}$$

Being α_p the roots of the equation

$$J_1(\alpha_p) = \frac{2 Sh h_0 J_0(\alpha_p)}{\alpha_p H} \tag{27}$$

and being S_q the roots of the equation $Q_n(s)=0$.

THE SPEED OF THE CUMULATIVE RATIO OF DRUG RELEASE PROFILE

From the concentration $C_A(t, r, z)$ it is possible to derive the time that the cylindrical matrix device takes to release the total amount of active agent initially dissolved in the device. The method proposed by Simon is applied.¹⁵ By definition, the cumulative amount of the active agent released to the environment at time t , denoted $M(t)$, is the difference between the amount of active agent initially dissolved in the device and the amount of the active agent remaining at time t :¹⁰

$$M(t) = c_{AS} \pi R^2 H - 2 \pi \int_0^H \int_0^R C_A(t, r, z) r dr dz \quad (28)$$

Equation (28) is rewritten in terms of dimensionless variables as

$$M(\tau) = c_{AS} \pi R^2 H - 2 \pi R^3 c_{AS} \int_0^{\frac{H}{R}} \int_0^1 C(\tau, \rho, \zeta) \rho d\rho d\zeta \quad (29)$$

The normalized form of Eq. (29) is

$$\frac{M(\tau)}{M(\infty)} = 1 - \left(\frac{2R}{H} \int_0^{\frac{H}{R}} \int_0^1 C(\tau, \rho, \zeta) \rho d\rho d\zeta \right) \quad (30)$$

where $M(\infty) = c_{AS} \pi R^2 H$. The Laplace transform of Eq. (30) is

$$\frac{M(s)}{M(\infty)} = \frac{1}{s} - \frac{2R}{H} \int_0^{\frac{H}{R}} \int_0^1 \bar{C}(s, \rho, \zeta) \rho d\rho d\zeta \quad (31)$$

or

$$\frac{M(s)}{M(\infty)} = \frac{1}{s} - \frac{2R}{H} \int_0^{\frac{H}{R}} \int_0^1 \left(A_0 I_0(\sqrt{s} \rho) + \frac{1}{s} + \left(\sum_{n=1}^{\infty} A_n \cos\left(\frac{n \pi R \zeta}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2} \rho}{H}\right) \right) \right) \rho d\rho d\zeta \quad (32)$$

After computing the integrals in Eq. (32), the following equation is obtained:

$$\frac{M(s)}{M(\infty)} = - \frac{2 A_0 I_1(\sqrt{s})}{\sqrt{s}} \quad (33)$$

and finally

$$\frac{M(s)}{M(\infty)} = - \frac{\left(-2 Sh \left(\sum_{n=1}^{\infty} \left(\frac{2 A_n H \cos\left(\frac{n \pi}{2}\right) \sin\left(\frac{n \pi h_0}{H}\right) I_0\left(\frac{\sqrt{n^2 \pi^2 R^2 + s H^2}}{H}\right)}{n \pi R} \right) \right) s R - 4 Sh h_0 \right) I_1(\sqrt{s})}{\left(I_1(\sqrt{s}) s \left(\frac{3}{2} \right) H + 2 Sh h_0 I_0(\sqrt{s}) s \right) \sqrt{s}} \quad (34)$$

Now according to the work by Simon and Collins, the effective relaxation time is defined by [1,5]:

$$t_{eff} = \lim_{s \rightarrow 0} \left(\frac{\psi_{ss}}{s^2} + \frac{d\bar{\psi}(s)}{ds} \right) \left[\lim_{s \rightarrow 0} \left(\frac{\psi_{ss}}{s} - \bar{\psi}(s) \right) \right]^{-1} \tag{35}$$

where ψ_{ss} is the steady-state value and $\bar{\psi}$ is the Laplace transform of ψ . When Eq. (35) is applied to Eq. (34) we obtain the effective time given by

$$\tau_{eff} = - \left(\frac{\partial^2}{\partial s^2} - \frac{\sqrt{s} \left(-2Sh \sum_{n=1}^{\infty} \left(\frac{2A_n H \cos\left(\frac{n\pi}{2}\right) \sin\left(\frac{n\pi h_0}{H}\right) I_0\left(\frac{\sqrt{n^2\pi^2 R^2 + sH^2}}{H}\right)}{n\pi R} \right) \right)_{sR-4Shh_0} I_1(\sqrt{s})}{I_1(\sqrt{s}) s \left(\frac{3}{2} \right) H + 2Shh_0 I_0(\sqrt{s}) s} \right) \Big|_{s=0} \tag{36}$$

4. Conclusions

An analytical and computational study of a cylindrical matrix device for controlled drug release was conducted. Expressions for the time constant and the cumulative amount of drug released were provided. The transient two-dimensional model was studied using Laplace transform techniques, the Bromwich integral, and the residue theorem. When the release ring and the cylinder were of equal height, the total amount of drug delivered was an increasing function of the mass transfer and diffusion coefficients and a decreasing function of the device length. The release rate increases with the mass transfer coefficient. A similar conclusion was reached when the length of the ring was marginally smaller than that of the system. However, the expressions obtained in this case were more computationally demanding when compared to the equal-height design specification. The last case study showed that the methodology was well-suited for any size requirement but came at the expense of more elaborate calculations. A potential application of the method is the prediction of the time to reach a desired plasma drug concentration following the application of the device.

5. References

1. Simon, L. 2011. A computational procedure for assessing the dynamic performance of diffusion-controlled transdermal delivery devices. *Pharmaceutics* 3: 485-496.
2. Tojo, K, Miyanami, K. 1983. Controlled Release from a Cylindrical Matrix Device. *Bulletin of University of Osaka Prefecture. Series A, Engineering and Natural Sciences* 31: 149-157.
3. Tojo, K. 1984. Prolonged drug release from a simple cylindrical device with a small hole. *Chem Eng Commun* 30: 311-322.
4. http://en.wikipedia.org/wiki/Computer_algebra_system
5. Collins, R. 1980. The choice of an effective time constant for diffusive processes in finite systems. *J Phys D Appl Phys* 13: 1935-1947.
6. Maple (www.maplesoft.com)

Analytical Model of the Brain Vascular System for Estimation of the Arterial Input Function (AIF) at the Tissue Level

Siamak P. Nejad Davarani^{1,2}, Hassan Bagher-Ebadian^{2,3,4}, James R. Ewing^{2,3}, Michael Chopp^{2,3}, Douglas Noll¹, Quan Jiang^{2,3}

¹Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI

²Department of Neurology, Henry Ford Hospital, Detroit, MI

³Department of Physics, Oakland University, Rochester, MI

⁴Department of Nuclear Engineering, Shiraz University, Shiraz, Iran

Abstract- *In Magnetic Resonance Dynamic Contrast Enhanced (MR DCE) studies, one of the key elements is estimating the AIF at the tissue level. So far, no analytical model has been implemented to address dispersion at all vascular branching levels down to the tissue, to give a realistic profile of the blood flow at the this level. Here, we introduce a model that we have proposed using laws of fluid dynamics and morphology of the vascular structure and have employed that to find dispersion of the contrast agent profile in the brain vessels and also the transfer function of the vessels at different levels.*

Keywords: Arterial Input Function, Vascular Modeling, Dynamic Contrast Enhanced MRI, Contrast Agents.

1 Introduction

Estimating the AIF of a Contrast Agent (CA), the time-concentration curve in plasma, has long presented a problem in MR-DCE and Dynamic Susceptibility Contrast (DSC) imaging studies. The AIF is used in estimating Mean Transit Time (MTT), Cerebral Blood Flow (CBF), Cerebral Blood Volume (CBV), vascular forward transfer rate constant (K^{trans}), vascular volume fraction (v_D), and extracellular-extravascular space (v_e) in DSC and DCE studies[1, 2]. Inaccurate estimation of the AIF for evaluation of permeability and perfusion could substantially increase bias in the estimated hemodynamic and permeability maps. This is one of the main reasons for finding the correct Arterial Input Function (AIF) at the tissue level (which we will refer to as “Tissue Input Function” or TIF). One of the first steps toward this goal is modeling the vascular system in the brain and using that to find the blood flow at the capillary (tissue) level.

Some of the earliest studies to understand and quantitate the morphology of the vascular system and dynamics of blood flow was done by Cecil Murray[3, 4] in the early 20th century where the relationship between the rate of blood flow and the volume of the vessel and

also his well-known arterial branching rule[4, 5] were interpreted. In the past couple of decades, many researchers have attempted to model vasculature for applications in DSC and DCE studies using different approaches. In one study Calamante et al used Independent Component Analysis (ICA) in perfusion studies as a tool to define a local AIF for obtaining more accurate quantification of CBF in DSC-MRI studies[6] based on a semi-manual approach. In another study, Mouridsen et al used Bayesian methods for estimation of cerebral perfusion [7]. The model that they designed for capillaries is basically a set of parallel delay lines each representing an arteriole and a capillary and each having a different transit time. The assumption for the AIF in this work was having a gamma-variate PDF which is a simplified form of the actual AIF in vasculature; this function or the exponential decay function for the local AIF are assumptions that have been used in other studies as well[8, 9]. In another study, Kazan et al have modeled the effects of laminar dispersion in Arterial Spin Labeling[10]; however, in their work they have not considered the effects of multiple pathways of flow through the vasculature for modeling the overall dispersion. Cebral et al used noninvasive methods to develop detailed assessment of blood flow patterns from direct in vivo measurements of vessel anatomy and flow rates using finite element methods[11]. In a recent work, Li et al created a method for tracking the AIF in DCE-MRI images of the breast[12]. However, this method was only focused on finding the voxels in the images that showed characteristics of being representative of the AIF and the goal was not finding a TIF. Another approach has been modeling the blood circulatory system of the whole body and finding the flow at different locations in the vascular system. In this category, Sherwin et al built a one dimensional network based on space-time variables and linear and non-linear modeling[13]. Another modeling approach is 3D-1D coupled models [14]. In other studies, Bagher-Ebadian et al suggested models based on the blood-circulatory system for estimating the CA time-concentration curve in arterial plasma after an intravenous bolus injection[15-17]. These methods show

different models for the AIF but these models either represent the input function only at the level of the major arteries (such as the carotid artery) or if they have an estimation of the input function at a lower level, the model does not represent all the major parameters that affect the AIF at the capillary level. In a recent study, Gall et al simulated delay and dispersion in a vascular tree model by calculating dispersion in a single vessel due to laminar flow and also by using a scaling rule for modeling the arterial tree[18]. This model has mainly been used for ASL and DSC bolus measurements[19, 20]. Here we have used a similar approach to this problem but have applied different rules for modeling the structure of the vascular tree and we also present a model for the TIF. For validating our model, we have used Dynamic Contrast Enhanced Computed Tomography (DCE-CT) images.

2 Methods and materials

2.1 Parametric expression of dispersion in a single vessel

The largest vessel entering the brain is the Carotid Artery which has a Reynolds Number less than 4000[21], therefore we exclude the possibility of having turbulent flow in any of the vessels in the brain. Based on this, we will focus on the arteries, arterioles, veins and venules in the brain where flow is laminar and will calculate the dispersion due to laminar flow. Figure 1 shows the effect of laminar flow on the shape of the profile of the contrast agent along the vessel. In Laminar flow the velocity of the fluid is dependent on the radial distance to the center of the tube [21]:

$$v = v_0 \left(1 - \frac{r^2}{R^2}\right) \quad (1)$$

where v_0 is the velocity of blood along the central axis of the vessel (maximum velocity) and v is the blood velocity at the radial distance r from this axis (with a radius of R). The fact is that at every point in time, we can assume that flow has reached a steady state and the velocity in the vessel has a magnitude that is dependent on the overall structure of the vascular system and therefore from this point on, we will only deal with v_0 .

As shown in Figure 1-a, we introduce a contrast agent to the entrance of the vessel in the form of a step function. Our goal is to find an equation that shows the concentration of the Contrast Agent (CA) in the volume enclosed by the two planes at D_0 and $D_0 + \Delta D$ as a function of time. We will use this equation to find a transfer function for this vessel.

First, we consider the situation as in Figure 1-b that CA enters this space, but does not pass through the second plane. We consider the time taken for the CA to reach this position is t , therefore D_0 and t are related through:

$$D_0 = v_0 \left(1 - \frac{r^2}{R^2}\right) t \quad (2)$$

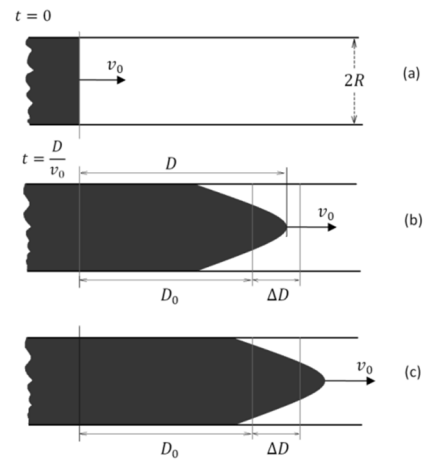


Figure 1. a) Introduction of contrast agent in the form of a step function to a vessel with laminar flow. b) The parabolic form of the CA after flowing the distance of D_0 in the vessel at time t while entering the volume enclosed by planes at D_0 and $D_0 + \Delta D$. c) The next step where the tip of the parabola exits the enclosed volume.

We define the time for the bolus to reach the D_0 plane as t_0 :

$$t_0 = \frac{D_0}{v_0} \quad (3)$$

Based on this definition, and also by dividing the volume of the CA by the enclosed volume, we can calculate the concentration of the CA between the two planes:

$$CA_c = 0 \quad \text{for} \quad t < t_0$$

$$CA_c = \frac{v_0 t}{2\Delta D} \left(1 - \frac{t_0}{t}\right)^2 \quad \text{for} \quad t \geq t_0 \text{ and } t < t_0 + \frac{\Delta D}{v_0}$$

$$CA_c = \frac{(2D_0\Delta D - \Delta D^2 + 2\Delta D v_0 t)}{2\Delta D v_0 t} \quad \text{for} \quad t \geq t_0 + \frac{\Delta D}{v_0} \text{ and } t < \frac{D_0 + \Delta D}{v_s} \quad (4)$$

where v_s is the slip velocity. These equations show the CA concentration with respect to time for unit step function as the input. For each of these cases if the time derivative is calculated, the response to the delta function or transfer function can be calculated. If we consider a case where ΔD is very large and also the slip velocity being zero, this will result in the second and third cases to be the same. In this case to calculate the CA concentration we have:

$$CA_c = \frac{1}{2} \left(1 - \frac{D_0}{D}\right) = \frac{1}{2} \left(1 - \frac{t_0}{t}\right) \quad (5)$$

Now by calculating the derivative with respect to time, the impulse response or transfer function of a single vessel can be found. This is a time varying function and is dependent on t_0 . Considering the fact that the no

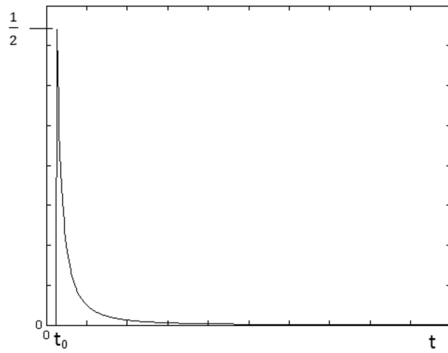


Figure 2. Plot of the transfer function of a single vessel. t_0 represents the time the tip of the CA wave reaches the end of the vessel.

contrast agent passes through the D_0 plane before time t_0 , the transfer function of a single vessel is as follows:

$$h(t) = \begin{cases} 0 & t < t_0 \\ \frac{t_0}{2t^2} & t \geq t_0 \end{cases} \quad (6)$$

Figure 2 shows the plot of the transfer function of a single vessel. If we consider the time that the contrast agent takes to reach the end of the vessel, D_0 would basically be the length of the vessel and v_0 the maximum velocity of blood in that vessel. This is the building block of our vascular model.

2.2 Morphological model of the vasculature

After calculating the transfer function of a single vessel, the next step is finding the transfer function (and distortion) of a cascade of branching vessels such that the morphological model of the brain vasculature is implemented and the flow and distortion at each level (and the overall distortion) is calculated. For this purpose, we designed a model of the vessels for finding the flow and dispersion from an artery to arterioles and all the way to capillaries and from those to the venules and veins. Figure 3 shows the 3D representation of this model. In this model, the artery at the first level is assigned a diameter and a flow rate, close to the same values of the carotid artery. At the end of this vessel, bifurcation happens and two daughter vessels are created with their radii following Murray's branching law of vessels [3, 5]:

$$r_p^3 = r_{d1}^3 + r_{d2}^3 \quad (7)$$

where r_p is the radius of the parent vessel and r_{d1} and r_{d2} are the radii of the two daughter vessels. In our model, first r_{d1} is selected randomly as a fraction of r_p and next, r_{d2} is calculated based on Murray's law. The length of the daughter vessels are also selected randomly

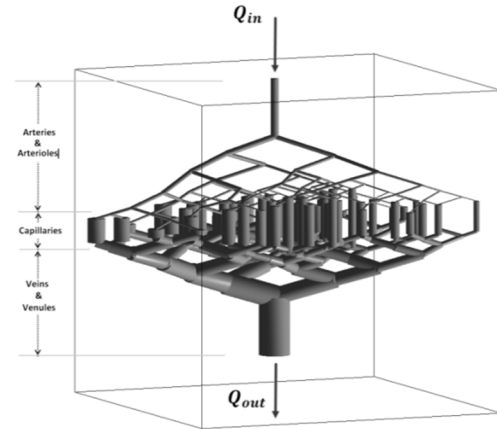


Figure 3. Morphological structure of the vascular model. Branching of arteries and arterioles has been done down to six levels. As seen here, the veins and venules have a larger volume and diameter compared to arteries and arterioles. The volumetric flow rate of blood entering this model equals the efferent flow. Every segment of the capillary bed is modeled as a single tube vessel in which the flow is non-laminar.

within a range as a fraction of the length of the parent vessel. This procedure is done recursively to create many levels of branching to get to the last level of arterioles [reference]. The next level in the vascular system following the arterioles is the capillary bed.

Here, we have assumed that every section of the capillary bed is fed by only one arteriole and the efferent blood is collected by only one venule. Also, we have modeled the whole capillary network between the arteriole and venule as a single vessel of larger diameter with non-laminar flow. Finally, we model the veins and venules. The overall volume of veins and venules in the body is about 4 times that of arteries and arterioles[21] and we have considered that in our model.

After implementing the morphological model of the vascular structure, the next step is calculating the flow rate in the branches which can be calculated as follows[21]:

$$Q_p = Q_1 + Q_2 \quad (8)$$

In our model, we assume that the blood flow is divided between the two daughter vessels based on their cross sectional areas. Using this, the velocity in all the branches and sub branches can be found.

2.3 Dispersion in a cascade of vessels

The next step in implementing the vasculature model is finding the transfer function of the vessels between the input artery and nodes in the structure. Based on this, the transfer function for a vessel at each level can be found as follows:

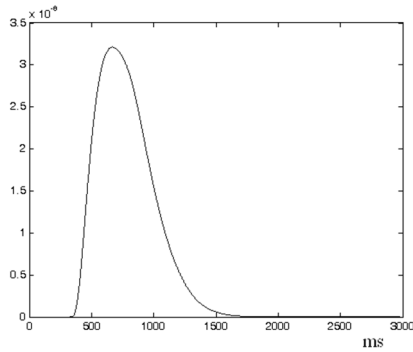


Figure 4. The transfer function of vessels from the input artery to the 6th level of the sub-branches

$$h(t)_n = \frac{t_{0n}}{2(t - t_{01} - t_{02} - \dots - t_{0n-1})^2}$$

for $t \geq t_{01} + t_{02} + \dots + t_{0n}$ (9)

In this function, t_{01} through t_{0n} are the time delays of each vessel. The transfer function of the vessels from the input to the n^{th} level of sub-branches is:

$$h(t)_{1 \text{ to } n} = h(t)_1 * h(t)_2 * \dots * h(t)_n \quad (10)$$

3 Results

3.1 Simulation results

After implementing the vascular model described in the previous section, by assigning a flow to the input artery, the flow to all the vessels in the model can be found. This, along with the physical characteristics of the vessels in the model can all be used to find the parameters of the transfer function (which are basically the time delays of the segments). Figure 4 shows the transfer function between the first node and a node at level 6 which has been analytically found using our model. This transfer function along with the AIF input to the main artery can be used for finding the AIF at different branching levels and the tissue.

This model can be validated using DCE Computed Tomography (CT) experimental data of the brain obtained with a bolus injection of Iodine. The advantage of using CT images instead of MR images is the higher time resolution of the dynamic series (0.55 sec vs. 5sec) and the linearity of the intensity of the CT images with respect to the CA concentration. Our validation method is based on measuring the AIF in every voxel of the image (or in other words, the TIF) and also the AIF of the main input artery that is seen in the image volume. Using parametric equations of the transfer function of the vascular structure (with different layers of branching), the goal is to find the best function that would distort the main AIF in a way to obtain the TIF for that voxel (Figure 5)

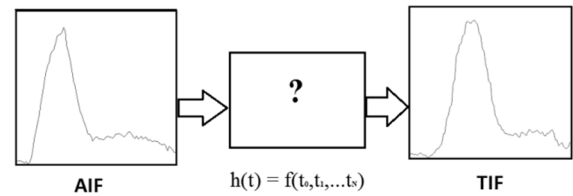


Figure 5. The parameters of the best fit transfer function describe the characteristics of the vascular structure between the main artery and the tissue.

4 Conclusion

We have designed a parametric model of the vasculature in the brain which is based on laws of fluid dynamics and laws governing the morphology of the vascular structure. Not all real life parameters have been addressed since that would make the model too complicated, beyond the needs of our applications. One advantage of our model is that it incorporates all the levels of the brain vascular system (capillaries, arteries, arterioles, veins and venules). At this point the proposed model lacks the results of validation with experimental data which is being worked on as the continuation of this work.

5 References

- [1] A. A. Chan and S. J. Nelson, "Simplified gamma-variate fitting of perfusion curves," *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, vol. 2, p. 4, 2004.
- [2] I. Nestorov, "Whole-body physiologically based pharmacokinetic models," *Expert Opin Drug Metab Toxicol*, vol. 3, pp. 235-49, Apr 2007.
- [3] C. D. Murray, "The Physiological Principle of Minimum Work: I. The Vascular System and the Cost of Blood Volume," *Proc Natl Acad Sci U S A*, vol. 12, pp. 207-14, Mar 1926.
- [4] C. D. Murray, "The Physiological Principle of Minimum Work Applied to the Angle of Branching of Arteries," *J Gen Physiol*, vol. 9, pp. 835-41, Jul 20 1926.
- [5] T. F. Sherman, "On connecting large vessels to small. The meaning of Murray's law," *J Gen Physiol*, vol. 78, pp. 431-53, Oct 1981.
- [6] F. Calamante, M. Morup, and L. K. Hansen, "Defining a local arterial input function for perfusion MRI using independent component analysis," *Magn Reson Med*, vol. 52, pp. 789-97, Oct 2004.
- [7] K. Mouridsen, K. Friston, N. Hjort, L. Gyldensted, L. Ostergaard, and S. Kiebel, "Bayesian estimation of cerebral perfusion using a physiological model of microvasculature," *Neuroimage*, vol. 33, pp. 570-9, Nov 1 2006.
- [8] F. Calamante, D. G. Gadian, and A. Connelly, "Delay and dispersion effects in dynamic susceptibility contrast MRI: simulations using singular value

decomposition," *Magn Reson Med*, vol. 44, pp. 466-73, Sep 2000.

[9] F. Calamante, D. G. Gadian, and A. Connelly, "Quantification of bolus-tracking MRI: Improved characterization of the tissue residue function using Tikhonov regularization," *Magn Reson Med*, vol. 50, pp. 1237-47, Dec 2003.

[10] S. M. Kazan, M. A. Chappell, and S. J. Payne, "Modeling the effects of flow dispersion in arterial spin labeling," *IEEE Trans Biomed Eng*, vol. 56, pp. 1635-43, Jun 2009.

[11] J. R. Cebal, P. J. Yim, R. Lohner, O. Soto, and P. L. Choyke, "Blood flow modeling in carotid arteries with computational fluid dynamics and MR imaging," *Acad Radiol*, vol. 9, pp. 1286-99, Nov 2002.

[12] X. Li, E. B. Welch, L. R. Arlinghaus, A. B. Chakravarthy, L. Xu, J. Farley, M. E. Loveless, I. A. Mayer, M. C. Kelley, I. M. Meszoely, J. A. Means-Powell, V. G. Abramson, A. M. Grau, J. C. Gore, and T. E. Yankeelov, "A novel AIF tracking method and comparison of DCE-MRI parameters using individual and population-based AIFs in human breast cancer," *Phys Med Biol*, vol. 56, pp. 5753-69, Sep 7 2011.

[13] S. J. Sherwin, V. Franke, J. Peiró, and K. H. Parker, "One-dimensional modelling of a vascular network in space-time variables," *J Eng Math*, vol. 47, pp. 217-250, 2003.

[14] P. J. Blanco, M. R. Pivello, S. A. Urquiza, and R. A. Feijoo, "On the potentialities of 3D-1D coupled models in hemodynamics simulations," *J Biomech*, vol. 42, pp. 919-30, May 11 2009.

[15] H. Bagher-Ebadian, K. Jafari-Khouzani, H. Soltanian-Zadeh, and E. J. R., "A Blood Circulatory Model to Estimate the Arterial Input Function in MR Brain Perfusion Studies," in *International Society of Magnetic Resonance in Medicine 16*, Toronto, Canada, 2008.

[16] A. Noorizadeh, H. Bagher-Ebadian, R. Faghihi, J. Narang, R. Jain, and E. J. R., "Input Function Detection in MR Brain Perfusion Using a Blood Circulatory Model Based Algorithm," presented at the International Society of Magnetic Resonance in Medicine, Stockholm, Sweden, 2010.

[17] H. Bagher-Ebadian, S. P. Nejad-Davarani, R. Paudyal, T. N. Nagaraga, S. Brown, R. Knight, J. D. Fenstermacher, and J. R. Ewing, "Construction of a Model-Based High Resolution Arterial Input Function (AIF) Using a Standard Radiological AIF and the Levenberg-Marquardt Algorithm," presented at the International Society of Magnetic Resonance in Medicine 19, Montreal, Canada, 2011.

[18] P. Gall, E. T. Petersen, X. Golay, and V. Kiselev, "Delay and Dispersion in DSC Perfusion Derived from a Vascular Tree Model Predicts ASL Measurements," presented at the ISMRM 16, Toronto, Canada, 2008.

[19] P. Gall, M. Guether, and V. Kiselev, "Model of Blood Transport Couples Delay and Dispersion and

Predicts ASL Bolus Measurements," presented at the ISMRM 18, Stockholm, Sweden, 2010.

[20] P. Gall and V. Kiselev, "On the Form of the Residue Function for Brain Tissue," presented at the ISMRM 18, Stockholm, Sweden, 2010.

[21] G. A. Truskey, F. Yuan, and D. F. Katz, *Transport phenomena in biological systems*, 2nd ed. Upper Saddle River, N.J.: Pearson Prentice Hall, 2009.

Protein Torsion Angle Class Prediction by a Hybrid Architecture of Bayesian and Neural Networks

Zafer Aydin¹, James Thompson², Jeffrey Bilmes³, David Baker² and William Stafford Noble⁴

¹ Department of Electrical and Electronics Engineering,
Bahcesehir University, Besiktas, Istanbul 34353 Turkey

² Department of Biochemistry, University of Washington, Seattle, WA 98195 USA

³ Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA

⁴ Department of Genome Sciences, Department of Computer Science and Engineering,
University of Washington, Seattle, WA 98195 USA

Abstract—*Protein torsion angles provide essential information about the three-dimensional structure of a protein. Accurate prediction of backbone angles can enhance the quality of tertiary (3D) structure prediction, sequence alignment and fold recognition. In this paper, we introduce a machine learning classifier that is able to predict the torsion angle category of an amino acid with high accuracy. Our method combines dynamic Bayesian networks with a neural network and is capable of incorporating information from multiple input representations such as position specific scoring matrices (PSSM) derived using sequence alignment methods. We show that 3D structure prediction accuracy of the widely used Rosetta program improves in the ab initio setting when the predicted torsion class information is used during the fragment selection step.*

Keywords: torsion angle prediction, fragment selection, dynamic Bayesian network, neural network, protein structure prediction

1. Introduction

Protein structure prediction is one of the most fundamental problems in computational molecular biology. Structure prediction is important because the biological functions of proteins are dependent on their 3D structures. Therefore, accurate prediction of the structure provides information on the functional role of the protein. Furthermore, structure prediction is necessary because experimental methods that solve structure are time consuming and cannot be easily applied to some classes of proteins. Finally, knowledge of protein structure enables us to design novel proteins and drugs, which is a fundamental task on the path toward treating diseases.

In tertiary structure prediction, the goal is to estimate the three dimensional coordinates of the atoms in a given target protein. Methods developed for structure prediction can be grouped into two main categories: template-based modeling and free modeling. In template-based modeling, the protein structure is built by matching the target to a template protein, which can be applied when structurally related templates are available in the Protein Databank (PDB). When such templates cannot be found, structure prediction is performed by free modeling. In this paper, we concentrate on free modeling (*i.e.*, the *ab initio* setting), in which we first select

a set of short amino acid fragments with known structures at overlapping segments of the target and then determine the tertiary structure of the target by assembling these fragments while minimizing an energy function [1]–[3].

Tertiary structure prediction greatly benefits from information such as secondary structure, solvent accessibility, torsion angles, and residue interactions, which are projections of the 3D structure to less complex representations. Therefore, instead of directly solving the 3D structure of a protein, which is a challenging task, an alternative approach is to predict these structural attributes and combine them to predict the full 3D coordinates of the atoms.

Torsion (*i.e.*, dihedral) angles contain important information for characterizing the three dimensional structure of a protein. The structure of an amino acid molecule can be defined with high precision by the torsion angles between three successive chemical bond vectors. Compared to other structure representations such as secondary structure, solvent accessibility or residue interactions, torsion angles provide not only complementary information but also a deeper insight into the 3D structure of a protein. Therefore, accurate prediction of torsion angles will significantly contribute to the accuracy and quality of 3D structure prediction.

Over the past couple of years, there has been a growing interest in predicting torsion angle information of a given amino acid sequence. Several methods have been proposed that concentrate on different sets of structural torsion states [4]–[13]. These methods exhibit a wide range of diversity. For instance, some of these methods predict real-valued torsion angles [11], while others predict discrete torsion labels [4], obtained by categorizing or clustering real-valued angles. Moreover, some methods rely on the availability of experimentally obtained NMR chemical shift data [13] and are therefore significantly more accurate than the ones which do not use such information. However, because NMR data is not available for all proteins, these methods should be considered in a separate category.

In this paper, we selected a previously defined 5-state torsion alphabet [14], and we concentrated our efforts on improving *ab initio* 3D structure prediction accuracy using torsion angle class predictions. Our torsion angle class prediction method is a hybrid architecture of a dynamic Bayesian network and a neural network. The model is capable of incorporating information from multiple position

specific scoring matrices (PSSM) derived using sequence alignment methods such as PSI-BLAST [15] and HH-MAKE [16]. In comparison with other predictors that use a similar alphabet, we obtain highly accurate torsion class predictions. Furthermore, we demonstrate that we are able to improve *ab initio* prediction of protein 3D structure by incorporating the torsion class predictions into Rosetta.

2. Methods

2.1 Problem Definition

An amino acid has three associated torsion angles as shown in Fig. 1. The angle ϕ denotes rotation about the C_α -N bond of the amino acid, ψ denotes rotation about the bond linking C_α and the carbonyl carbon, and ω denotes rotation about the bond between the carbonyl carbon of the current residue and the nitrogen of the next residue. We compute ϕ , ψ , and ω from the 3D coordinate information in PDB. Each of these angles is constrained to the range $[-180, 180]$.

Following [14], we subdivided the amino acids into five torsion angle classes, which represent the major clusters observed in the PDB (see the Ramachandran plot in Fig. 2). The resulting five labels are described in Table 1.

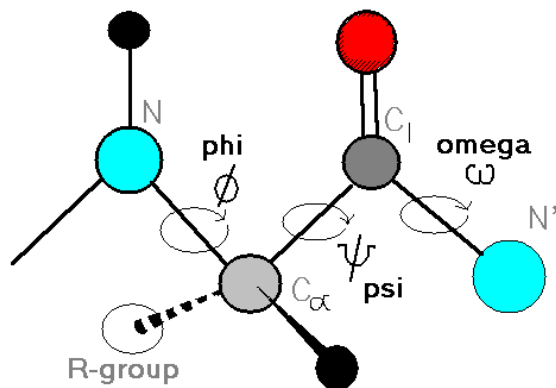


Fig. 1: Torsion angles of an amino acid. Image obtained from <http://www.bmb.uga.edu/wampler/tutorial/prot2.html>. Courtesy of Prof. John E. Wampler, University of Georgia, Athens, GA USA.

Table 1: The five torsion angle classes, their definitions, and the percent of amino acids assigned to each class in the 90% identity data set.

Label	Definition	Percent
A	$ \omega \geq 90, \phi < 0, -125 < \psi \leq 50$	50.22
B	$ \omega \geq 90, \phi < 0, \psi \leq -125$ or $\psi > 50$	42.23
E	$ \omega \geq 90, \phi \geq 0, \psi > 100$	1.94
G	$ \omega \geq 90, \phi \geq 0, \psi \leq 100$	4.73
O	$ \omega < 90$	0.88

Based on this definition, our 5-state torsion angle class prediction problem can be stated as follows. For a given protein, the goal is to assign to each amino acid a torsion angle label from the alphabet $\{A, B, E, G, O\}$ as shown in Fig. 3.

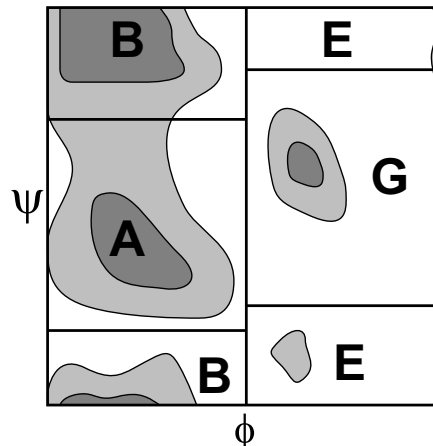


Fig. 2: Torsion angle classes obtained by partitioning the space of real-valued angles into discrete labels as in Table 1. The image is a high resolution version of Fig. 1(b) in [17]. Courtesy of Ben Blum, UC Berkeley, CA, USA.

amino acid sequence: LWGLVKQGLKCEDCGMNVHHKCREKVANLC
torsion angle labels: BBEABGABBBBAAAGBBBBBAAAAAABBABO

Fig. 3: 5-state torsion angle class prediction problem. The torsion labels are defined according to Table 1.

2.2 Prediction Model

Our *ab initio* torsion class predictor is a hybrid architecture, in which several dynamic Bayesian network (DBN) models are combined with a neural network. In this architecture, we first generate marginal *a posteriori* probability distributions of protein torsion angle classes using the DBNs. We then concatenate these distributions with the PSSM data and use as input features in the neural network as explained in Section 2.2.2. For completeness, we start with a brief description of our DBN model, which was previously introduced in [18] for protein structure prediction.

2.2.1 A dynamic Bayesian network model for torsion angle class prediction

We implemented the DBN shown in Fig. 4, which is similar to the model proposed by [19]. Each node in a DBN model represents a random variable. Our model contains five types of random variables: *state*, *state class history*, *state count down*, *change state*, and *amino acid profile*. These variables are observed during training, because the true torsion angle labels are available. During testing, only the *amino acid profile* is observed, and the other variables are hidden. An example showing the values of *state*, *state count down*, and *change state* is given in Fig. 5. The variables are briefly explained as follows:

- The *state* variable models the torsion label of an amino acid, as defined in Section 2.1.
- The *amino acid profile* variable models the observation data, which is a 20-dimensional vector of PSSM scores (*i.e.*, a column of the PSSM) derived by running a sequence alignment software against a protein database

(see Section 2.4).

- The *state class history* variable keeps track of the current and preceding torsion labels. This variable is represented as a tuple with $L_T + 1$ elements, where $L_T + 1$ is the size of the label dependency window, including the current label.
- The *state count down* variable models the length of a torsion label segment. When the length of a segment is less than or equal to D_{max} , then the value of *state count down* is the number of residues between the current position and the end of the segment. If the length of a segment is greater than D_{max} by k residues, then *state count down* is set to D_{max} for the first $k + 1$ residues and is set to $D_{max} - 1, D_{max} - 2, \dots, 1$ for the remaining residues in that segment.
- The *change state* variable also models the length of a torsion label segment. It simply signals when a transition to a new segment should be made. It is set to 1 if *state count down* is 1 and 0 otherwise.

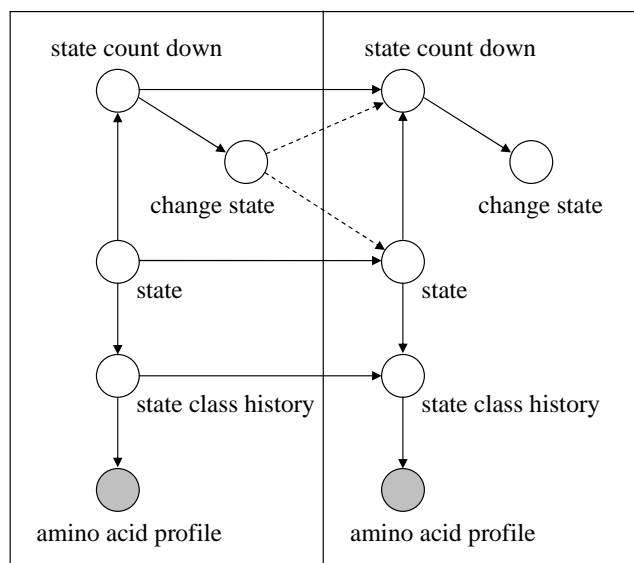


Fig. 4: A dynamic Bayesian network for torsion angle class prediction. The first column shows the variables of the *prologue* (models the first amino acid) and the second column shows the variables of the *chunk* (models the second up to the last amino acid).

```

state:      AAAAABBBEEOBBBBBBBEEAAAA
state count down: 543213212117765432114321
change state: 000010010110000000110000

```

Fig. 5: An example *state* sequence and the values of *state count down* and *change state* variables for $D_{max} = 7$.

A DBN models the generation of observation data for all possible values of hidden variables in a probabilistic framework. The relations among discrete variables in the DBN are defined by conditional probability distributions

(CPDs), and continuous variables are modeled by probability density functions. For instance, the state transition distribution assigns probabilities to transitions from one torsion angle state to another; distributions related to the lengths of the segments assign probability values for all possible lengths of torsion label segments, and the observation density models the generation of the observed data. Because of the dependencies among adjacent amino acids, the first amino acid is modeled slightly differently than the rest of the amino acids. Therefore, in Fig. 4(A), the first column (*prologue*) shows the nodes for the first amino acid, and the second column (*chunk*) is a model for the rest of the amino acids. By extending the *chunk* $N - 1$ times to the right, we obtain the full network structure, where N is the number of amino acids in the protein. Detailed formulations for the CPDs that define the relations among discrete nodes can be found in [19] and the probability density function that models the generation of observation data can be found in [18].

Our DBN has the following hyper-parameters: a one sided PSSM window size (L_A), a one sided torsion angle label window (L_T), a segment length threshold (D_{max}), a diagonal covariance component regularizer weight (ω) and a PSSM contribution weight (α). Detailed description of these parameters can be found in [18]. In this work, we set the parameters as: $L_A = 9$, $L_T = 4$, $D_{max} = 13$, $(\omega, \alpha) = (0.05, 0.5)$ when PSI-BLAST PSSM data is used as input features and $(\omega, \alpha) = (0.035, 0.4)$ when HHMAKE PSSM data is used (see Section 2.2.2). We implemented the model shown in Fig. 4 using the Graphical Models Toolkit (GMTK) [20], a C++ package for DBNs and other dynamic graphical models.

2.2.2 Combining multiple DBNs by a neural network classifier

Motivated by previous work [18], [19], we make our predictions by combining the results from multiple DBN models. In the first model, we allow dependencies from past positions only. Conversely, in the second model, we reverse the PSSM profiles as well as the torsion angle class labels and use the same model depicted in Fig. 4. Effectively, the second model only allows dependencies from future positions [18]. In both models, we use PSI-BLAST's PSSMs [15] as the observation data. Additionally, we implement a similar pair of DBNs characterizing past and future dependencies for PSSM profiles derived using HHMAKE (see Section 2.4). As a result, we have a total of four DBNs. Each model produces a marginal *a posteriori* distribution over torsion labels for each amino acid. In this work, we combine these distributions using a neural network classifier, which is a multi-layer perceptron as shown in Fig. 6. The input units of this network represent the elements of a rich feature set. For each amino acid position, we use a symmetric window of PSSM vectors derived from PSI-BLAST and HHMAKE as well as a window of marginal *a posteriori* probabilities that are generated from the four DBNs. We set the lengths of the PSSM and the posterior probability windows to be 15. Our feature set contains the following *a posteriori* distributions: (1) average of *a posteriori* probabilities from the four DBNs, (2) average of *a posteriori* probabilities from the DBNs

that use PSI-BLAST PSSMs, (3) average of *a posteriori* probabilities from the DBNs that use HHMAKE PSSMs. This gives a total of 825 features. Hence, our input layer has 825 input units. For positions at which the feature window extends beyond the boundaries of a protein (*i.e.*, those that are close to the N- or C-terminus), we include zeros in the feature set. Our neural network has a single hidden layer with 75 hidden units and an output layer with 5 output units. We use gradient descent to learn the parameters of this neural network where we optimized the mean square error by setting the number of iterations to 50, and the learning rate to 0.005. In the hidden layer, we use the hyperbolic tangent (*i.e.*, tanh) and in the output layer, we use the softmax transformation as the activation function. Our neural network predicts the torsion angle label of the amino acid at the center of the feature window by selecting the particular label with maximum score at the output layer. We implemented our neural network classifier using Torch5 [21].

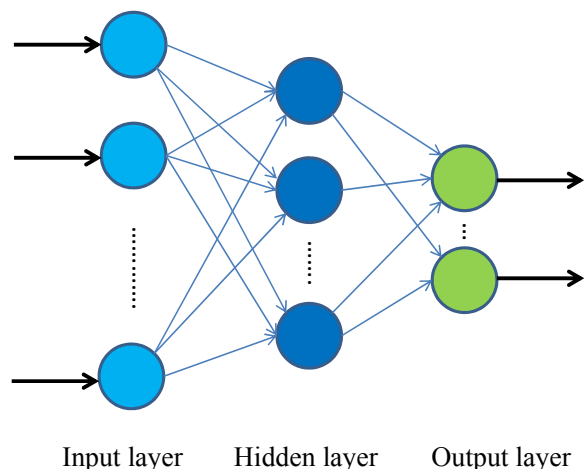


Fig. 6: A multi-layer perceptron for predicting the torsion angle class prediction. The neural network contains the input, hidden and output layers. A softmax transformation is used as the activation function at the output layer to estimate the torsion angle class probabilities.

2.3 Fragment based prediction of protein tertiary structure

We used the Rosetta software to select the amino acid fragments and to predict the 3D structure of proteins in our test set. We generated two versions of the fragment selection module of Rosetta. The first version utilizes the PSI-BLAST PSSM, predicted secondary structure and the rama score as input features and the second version utilizes the same features as well as the predicted torsion angle class information. Further details about the fragment selection in Rosetta can be found in [22]. Detailed description of the structure prediction experiments can be found in Section 3.

2.4 Generating position-specific scoring matrices for torsion angle class prediction

We use PSSMs generated by the PSI-BLAST [15] and HHMAKE [16] algorithms as input features. Detailed

descriptions of these PSSMs can be found in [18]. In this work we used BLAST version 2.2.20 and the NCBI's non-redundant (NR) database dated June 2011 to generate PSI-BLAST PSSMs. The command line we used to derive the profiles was: `./blastpgp -i protein.fasta -o protein.align -Q protein.pssm -j 3 -e 0.001 -h 1e-10 -d nr.filtered`. The PSI-BLAST software can be downloaded from the help section of <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

We derive HHMAKE PSSMs from HMM-profiles created using the HHMAKE algorithm, which is the first step of the HHsearch method [16]. To obtain the HMM-profiles with HHMAKE, we used the following pair of command lines: `./buildali.pl protein.fasta` followed by `./hhmake protein.a3m`. In this work, we used HHsearch version 1.5.1 to generate profiles. The recommended database for HHMAKE is the NRE database, which is a combination of the NR and the ENV databases. In this work, we used NRE90 and NRE70, which are the filtered versions of the NRE database at 90% and 70% identity thresholds, respectively. The binaries used for generating the HMM-profiles can be obtained from <ftp://toolkit.lmb.uni-muenchen.de/HH-suite/>.

Previous work suggests the utility of scaling the PSSM values by applying a transforming function [19], [23]. In this work, we employ the following sigmoidal transformation to scale the PSI-BLAST and HHMAKE PSSMs:

$$f_{sigmoid}(x) = \frac{1}{1 + \exp(-x)}. \quad (1)$$

The sigmoid transforms the PSSM values into the range $[0, 1]$. Presumably, one of the benefits of the sigmoidal transform is that it maps PSSM values in $(-\infty, \infty)$ to $[0, 1]$, which normalizes the variance.

2.5 Datasets

2.5.1 PDB-PC90 dataset

To obtain the PDB-PC90 dataset, we used the PISCES server [24] with the following set of criteria: percent identity threshold of 90%, resolution cutoff of 2.5 Å, and R-value cutoff of 1.0. We also used PISCES to filter out non-X-ray and $C\alpha$ -only structures and to remove short (< 30 amino acids) and long (> 10000 amino acids) chains. This dataset contained 17056 chains.

2.5.2 Training and test set for evaluating torsion angle class prediction accuracy

We randomly selected 5161 proteins from the PDB-PC90 dataset. Among those, we randomly selected 994 proteins to form our first test set, which is used to evaluate the torsion angle class prediction accuracy of our method. From the set of 5161 proteins, we then removed those proteins that are similar to the set of 994 proteins using a 10% sequence identity threshold. The remaining set contained 4205 chains, which is used to train our torsion angle class prediction method.

2.5.3 Training and test set for evaluating the improvement in 3D structure prediction accuracy

We generated the 3D structure prediction models on a benchmark set of 61 proteins that is commonly used for ab-initio structure prediction assessment [22]. To generate torsion angle class predictions for this test set, we compiled a second training set as follows. We started from the PDB-PC90 dataset and removed proteins that are similar to the benchmark set of 61 proteins at 10% sequence identity threshold. The remaining set contained 17018 proteins and is used to train our method to generate torsion angle predictions on the benchmark with 61 proteins.

2.6 Accuracy Measures

To assess the accuracy of torsion angle class predictions, we used the amino acid level accuracy [25], segment overlap score [26] and Matthew's correlation coefficients (MCC) [27]. All three of these metrics are widely used in protein secondary structure prediction. The amino acid level accuracy is computed as the total number of amino acids with correctly predicted torsion labels divided by the total number of amino acids in the test set. The segment overlap score measures how well the predicted torsion label segments match the true segments and is biologically more meaningful than the amino acid level accuracy. Matthews correlation coefficient is a correlation score between the observed and predicted binary classifications. It is a balanced measure which can be used even if the individual classes are of very different sizes. For the amino acid level accuracy, segment overlap score, and Matthew's correlation coefficient, a high score indicates a more accurate prediction as compared to a low score.

To evaluate the accuracy of fragment selection, we used the CRMSD measure, which is the root mean square deviation between the C_α atoms of the native (true) amino acid segment and a fragment selected from the library [22]. A low CRMSD score represents a more accurate fragment as compared to a high CRMSD score. Finally for 3D structure prediction, we used the GDTMM score, which is a variant of the Global Distance Test measure. GDTMM represents the percentage of residues superimposable to the experimentally determined native structure calculated across a number of different distance thresholds [28]. A high GDTMM score is more accurate than a low GDTMM score in predicting the 3D structure of a protein.

3. Results

3.1 Torsion class prediction accuracy

We evaluated the torsion class prediction performance of our classifier on our large benchmark dataset with 994 proteins where we trained our method on the set of 4205 proteins (see Section 2.5.2). We randomly split our training set into two and used the first half to train the DBNs and the second half to train the neural network. Table 2 shows the confusion matrix and Table 3 includes the amino acid level accuracy, segment overlap measure and Matthew's correlation coefficient measures for the 5-state torsion angle class prediction.

Table 2: Confusion matrix for the 5-state torsion class prediction on the set of 994 proteins.

True Pred	A	B	E	G	O	Row Sum
A	105914	12488	178	1123	8	119711
B	12968	81124	460	1013	72	95637
E	613	1320	1722	774	13	4442
G	2319	1976	563	6136	12	11006
O	177	539	37	27	1078	1858
Column Sum	121991	97447	2960	9073	1183	232654

Table 3: Confusion matrix for the 5-state torsion class prediction on the set of 994 proteins.

Acc(%)	SOV(%)	MCC _A	MCC _B	MCC _E	MCC _G	MCC _O
84.23	78.65	0.74	0.73	0.47	0.60	0.73

At this point, we are not able to directly compare our results to other methods because none of the torsion prediction methods available in the literature used the same 5-state mapping defined in Table 1. The closest torsion alphabet contains grid-defined four states (A, B, G, E) [29] excluding the O state in our 5-state alphabet. In this 4-state representation, Bystroff *et al.* [30] obtained 74% amino acid level accuracy, Kuang *et al.* [8] achieved 77% and Faraggi *et al.* [11] reached 84%. Note that, the dataset utilized by Faraggi *et al.* was obtained using a 25% sequence-identity threshold, which is higher than the threshold we used to compile our train/test sets. In other words, the rules we used to generate our train/test sets are significantly more stringent than Faraggi *et al.*, which makes our testing conditions more difficult. Furthermore, we evaluated our performance on a 5-state torsion angle alphabet, which is more difficult than a 4-state prediction because as the number of torsion states increases the torsion angle prediction accuracy decreases [4]. Therefore, we claim that our 5-state torsion class prediction accuracy is potentially at a level comparable to the state of the art. However, we are more interested in improving the 3D structure prediction using the torsion class information than in out-performing existing torsion angle prediction methods. Therefore, in the next section, we analyze the effect of incorporating torsion class information into Rosetta.

3.2 Fragment selection using torsion class prediction

In this section, we analyze the quality of fragments selected by Rosetta when we use the torsion class predictions in the fragment picker. We considered two possible scenarios in the score function of the fragment picker: (1) torsion score component is off (the control group) (2) torsion score component is on. We predicted 5-state torsion classes of the benchmark set of 61 proteins using our method. For this purpose, we used the dataset with 17018 proteins to train our models (see 2.5.3). Similar to Section 3.1, we randomly split our training set into two and used the first half to train the DBNs and the second half to train the neural network. The accuracy measures for these predictions is summarized in Table 4.

We selected 200 fragments at each fragment window position using Rosetta on the same benchmark. Fig. 7 summarizes the fragment quality measures. Fig. 7(a) shows

Table 4: Accuracy measures for the 5-state torsion class prediction on the benchmark set of 61 proteins.

Acc(%)	SOV(%)	MCC _A	MCC _B	MCC _E	MCC _G	MCC _O
88.34	82.90	0.87	0.51	0.68	0.83	0.80

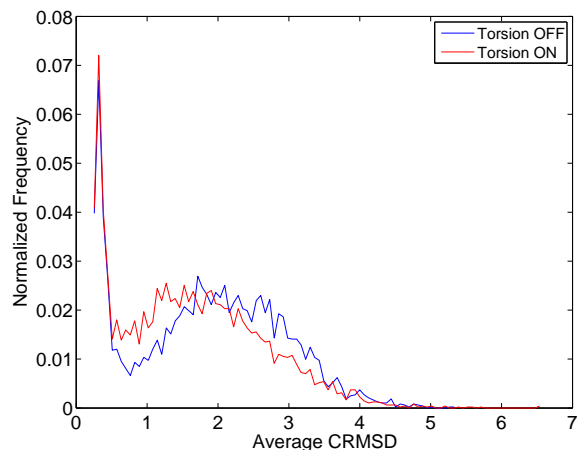
the histogram of the average CRMSD measures such that the average is computed over the set of 200 fragments in a given fragment window. In this plot, the blue curve shows the histogram for the control group in which the torsion prediction information is not utilized during fragment selection and the red curve depicts the histograms when the torsion class predictions are considered in fragment picker. The mean and standard deviation of the histograms shown in Fig. 7(a) are tabulated in Table 5. As a lower CRMSD value indicates a better fragment quality, we are able to improve the mean of the average CRMSD measure by 0.23 CRMSD when we use the torsion prediction information during fragment selection. Fig. 7(b) illustrates the average CRMSD values for each fragment window such that the average is computed over the set of 200 fragments. In Fig. 7(b), the fraction of points below the diagonal line is 73.76%, which means that on average, we are able to improve the fragment quality on the majority of positions.

Table 5: Mean and standard deviation of the average CRMSD values of the fragments picked for the benchmark set of 61 proteins.

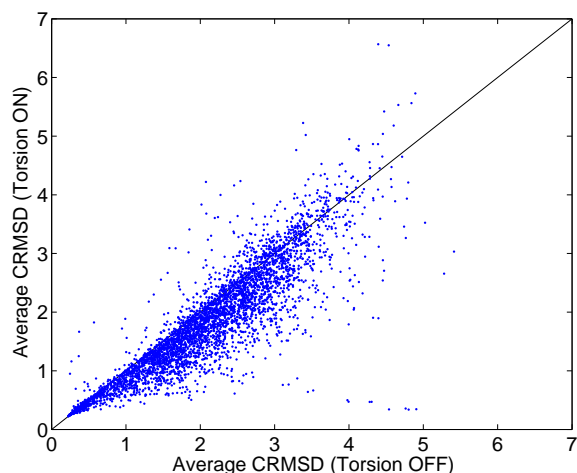
	Torsion OFF	Torsion ON
Mean/std of Average CRMSD	1.85/1.04	1.62/1.00

3.3 3D structure prediction improvement by torsion class prediction

In the next step, the 3D structures of the 61 benchmark proteins are predicted using Rosetta. In this experiment, the GDTMM scores of the 3D models predicted by Rosetta are computed for the two cases (torsion score on vs off during fragment selection), where a higher GDTMM score represents a closer match to the native structure. In both cases, PSI-BLAST profile similarity score, secondary structure similarity score, and rama score (obtained from constraints in Ramachandran space) are also used as input features of the fragment selection algorithm. Fig. 8 compares the accuracies of 3D structure prediction for the cases where torsion class predictions are used or not in fragment selection. Each box-plot shows the average, minimum, maximum, and standard deviation values of the GDTMM scores. Method 1 represents the case where torsion class predictions are used during fragment selection and Method 2 shows the case where torsion class predictions are excluded. Fig. 8 clearly shows that there is an improvement in the overall 3D structure prediction accuracy when the torsion class predictions are included during fragment selection. The GDTMM score improved in 44 out of 61 proteins and the average improvement is computed as 3.61%. These results demonstrate that more accurate 3D predictions can be obtained by improving the feature set of the fragment selector.



(a)



(b)

Fig. 7: Fragment quality improvement using 5-state torsion class prediction. (a) Histogram for the average CRMSD for all the fragments. The blue curves show the control group in which the torsion similarity score component is turned off and the red curves demonstrate the behavior when the torsion similarity score is included in fragment selection step of Rosetta. A negative shift of the control distribution indicates an improvement in the fragment quality. (b) Average CRMSD values for all the fragments evaluated for the benchmark set of 61 proteins. Average CRMSD values for each fragment window is depicted for the cases where the torsion prediction information is utilized and excluded. Each point below the $y=x$ axis represents an improvement in the fragment quality.

4. Conclusion

In this paper, we developed a machine learning classifier that is capable of predicting 5-state torsion angle classes with high accuracy, and we demonstrated that the resulting predicted torsion angles can be used to generate more accurate 3D structure models. This work suggests

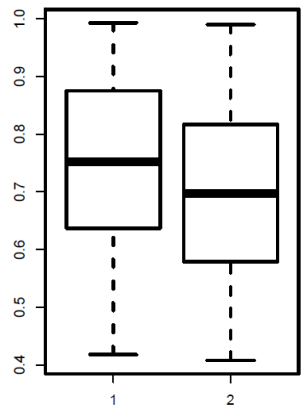


Fig. 8: Box-plots of GDTMM scores showing the average, minimum, maximum and standard deviations. The y-axis depicts the GDTMM scores of the 3D model predictions generated by Rosetta, and the x-axis shows the two cases where the torsion predictions are utilized or not during fragment selection. A higher GDTMM score represents a closer match with the native structure (*i.e.*, more accurate predictions). Method labeled 1 represents the case where the 5-state torsion predictions are incorporated into the fragment selection algorithm of Rosetta and the method 2 shows the statistics for the case where the 5-state torsion predictions are excluded. Employing torsion predictions shows a clear improvement in 3D structure prediction accuracy.

several directions for future research. First, it is possible to consider different torsion angle alphabets by clustering the real-valued torsion angles into discrete bins and analyzing which alphabet is most useful for 3D structure determination. Second, additional feature representations can be employed as input to our method so that the information contained in these representations is combined with that of the existing feature set by our DBN framework. Finally, the torsion class prediction method can easily be extended to operate in a comparative modeling setting, in which template proteins that are structurally similar to the target protein are used when available.

References

- [1] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions." *Journal of Molecular Biology*, vol. 268, pp. 209–225, 1997.
- [2] J. Lee, S. Y. Kimb, and J. Lee, "Protein structure prediction based on fragment assembly and parameter optimization," *Biophysical Chemistry*, vol. 115, no. 2-3, pp. 209–214, 2005.
- [3] J. B. Holmes and J. Tsai, "Some fundamental aspects of building protein structures from fragment libraries," *Protein Sci*, vol. 13, no. 6, pp. 1636–1650, 2004.
- [4] P. Kountouris and J. D. Hirst, "Prediction of backbone dihedral angles and protein secondary structure using support vector machines," *BMC Bioinformatics*, vol. 10, no. 437, 2009.
- [5] A. G. de Brevern, C. Etchebest, and S. Hazout, "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks," *Proteins*, vol. 41, no. 3, pp. 271–287, 2000.
- [6] O. Zimmermann and U. H. E. Hansmann, "Locustra: accurate prediction of local protein structure using a two-layer support vector machine approach," *J. Chem. Inf. Model*, vol. 48, no. 9, pp. 1903–1908, 2008.
- [7] Q. Dong, X. Wang, L. Lin, and Y. Wang, "Analysis and prediction of protein local structure based on structure alphabets," *Proteins*, vol. 72, pp. 163–172, 2008.
- [8] R. Kuang, C. S. Leslie, and A. S. Yang, "Protein backbone angle prediction with machine learning approaches," *Bioinformatics*, vol. 20, no. 10, pp. 1612–1621, 2004.
- [9] S. Wu and Y. Zhang, "Anglor: a composite machine-learning algorithm for protein backbone torsion angle prediction," *PLoS One*, vol. 3, no. 10, p. e3400, 2008.
- [10] E. Faraggi, B. Xue, and Y. Zhou, "Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network," *Proteins*, vol. 74, no. 4, pp. 847–856, 2009.
- [11] E. Faraggi, Y. Yang, S. Zhang, and Y. Zhou, "Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction," *Structure*, vol. 17, no. 11, pp. 1515–1527, 2009.
- [12] C. Mooney and G. Pollastri, "Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, pp. 181–190, 2009.
- [13] Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax, "Talos+: A hybrid method for predicting protein backbone torsion angles from nmr chemical shifts," *J. Biomol. NMR*, vol. 44, pp. 213–223, 2009.
- [14] B. Blum, M. Jordan, D. Kim, R. Das, P. Bradley, and D. Baker, "Feature selection methods for improving protein structure prediction with Rosetta," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 137–144.
- [15] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 1997.
- [16] J. Soding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, pp. 951–960, 2005.
- [17] B. Blum, M. I. Jordan, and D. Baker, "Feature space resampling for protein conformational search," *Proteins*, vol. 78, no. 6, pp. 1583–1593, 2010.
- [18] Z. Aydin, A. Singh, J. Bilmes, and W. S. Noble, "Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure," *BMC Bioinformatics*, vol. 12, p. 154, 2011.
- [19] X.-Q. Yao, H. Zhu, and Z.-S. She, "A dynamic bayesian network approach to protein secondary structure prediction," *BMC Bioinformatics*, vol. 9, no. 49, 2008.
- [20] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.
- [21] R. Collobert, "Torch," NIPS Workshop on Machine Learning Open Source Software, 2008, software available at <http://torch5.sourceforge.net/>.
- [22] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, and D. Baker, "Generalized fragment picking in Rosetta: Design, protocols and applications," *PLoS One*, vol. 6, no. 8, 2011.
- [23] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, vol. 292, pp. 195–202, 1999.
- [24] G. Wang and R. L. Dunbrack, Jr., "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, pp. 1589–1591, 2003, web server at <http://dunbrack.fccc.edu/PISCES.php>.
- [25] B. Rost and V. A. Eylich, "EVA: Large-scale analysis of secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 45, no. S5, pp. 192–199, 2002.
- [26] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment," *Proteins*, vol. 34, pp. 220–223, 1999.
- [27] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochim Biophys Acta*, vol. 405, no. 2, pp. 442–451, 1975.
- [28] A. Zemla, "LGA – a method for finding 3d similarities in protein structures," *Nucleic Acids Research*, vol. 31, pp. 3370–3374, 2003.
- [29] B. Olivia, P. A. Bates, E. Querol, F. X. Aviles, and M. J. Sternberg, "An automated classification of the structure of protein loops," *Journal of Molecular Biology*, vol. 266, no. 4, pp. 814–830, 1997.
- [30] C. Bystrhoff, V. Thorsson, and D. Baker, "HMMSTR: A hidden markov model for local sequence-structure correlations in proteins," *Journal of Molecular Biology*, vol. 301, pp. 173–190, 2000.

Cryptography and Information Protection in the Living World

Naya Nagy¹, Marius Nagy¹, and Paul Hodor²

¹ College of Computer Engineering and Science
Prince Mohammad Bin Fahd University, Al Khobar, KSA
{nnagy,mnagy}@pmu.edu.sa

² Booz Allen Hamilton, Rockville, MD, USA
hodor_paul@bah.com

Abstract. This paper explores parallels between concepts defined in cryptography and concepts of biology at different levels of organization. Cryptographic settings, including the presence of an eavesdropper are extensive in the realm of plants and animals. It also turns out that principles of information protection show strong similarities between the two disciplines: computer science and molecular biology. Biological information, as held by the DNA molecule, and digital information, as used in digital communication systems, are subject to analogous procedures of protection and repair when damaged.

Keywords: information protection, error correction, cryptography, mimicry, DNA, DNA repair

1 Introduction

Cryptography is a field that spreads *human* activities. The need to communicate privately, or secretly, enters various corners of human private and social life, such as financial transactions, personal privacy of communication, company secrets, information protection at the level of a country, a state, a group, or organization. In fact, the possibility to communicate privately with another human is considered to be an individual freedom. It has the flavor of a human right.

The idea behind this paper is that the need for secret communication, or more generally, the existence of cryptographic needs is not inherently pertaining to humans. Cryptographic settings and cryptographic solutions can be encountered throughout the living world. The point of view may be that of an information-carrying molecule, a cell, an entire organism, a population, or an ecosystem. This paper explores various scenarios in which encryption/decryption, and cryptographic identities are part of vital processes.

2 The Players in Cryptography and Their Interests

We may consider the classic model of a cryptographic setting to be sufficient for the biological realities to be discussed here. This model involves three entities:

two communication partners and a third malevolent party that intends to corrupt the communication. The corruption of the communication refers to its privacy and the reliability and truthfulness of the exchanged messages.

2.1 The Good Players

By standard now, it is Alice and Bob that intend to communicate secretly and reliably. Alice and Bob are usually equivalent partners. The communication is symmetric and as such, Alice's and Bob's points of view are identical. For the communication to comply to secrecy and reliability, Alice's expectations are the following:

1. Alice wants to get all messages from Bob. This means that the connection between Alice and Bob should be permanently working. Or else, if the connection is broken, both Alice and Bob should be aware of it. Any message *sent* by Bob should *reach* Alice.
2. Any message Alice gets from Bob is indeed sent by Bob. This is called authentication. Bob's message may carry Bob's unique signature. For logical completeness, any message that is *not* coming from Bob, is known to have another sender. That is, Alice recognizes the message to have a foreign sender.
3. For any message that Alice sends to Bob, Alice knows whether Bob has received the message. This is a handshake.
4. When Alice receives a message from Bob, the content of Bob's message is complete and unaltered. No parts of the message were lost, no meaning has been altered or twisted.
5. Bob's message is understandable to Alice. That is, they speak the same language, use the same alphabet, semantics, and syntax. Additionally, any new concept that Bob may start using, should first be defined to Alice, before its usage.

These expectations have been formulated for Alice. As Bob is an equivalent entity, all of the above items apply symmetrically to Bob.

2.2 The Bad Player

The third party, called Eve, makes every effort to meddle, interrupt, or attack the privacy of Alice's and Bob's communication. Thus any form of corrupting the transfer of messages is in the domain of Eve. Eve may listen to the communication, or break the connection, or may have any other destructive behavior. The interest of Eve may vary, depending on the practical setting and goals. Some of Eve's possible attacks may not be compatible with one other. For example, if Eve chooses to interrupt a conversation by severing the connection, this means she cannot gain any knowledge on what Alice and Bob would have communicated to each other. It means she obviously cannot eavesdrop on the conversation.

Consider the following list of attacks that Eve may plan on the communication:

1. **Eavesdropping.** Eve may listen to the communication channel and read the encrypted messages.
2. **Tampering.** Eve may tamper with the content of a message. For example, if the message is a string of characters, Eve deletes a substring from the message and/or inserts a substring of her own into the string of the message.
3. **Inserting.** Eve may insert a false message. Eve may send a false message to Alice or Bob.
4. **Intercepting.** Eve may intercept a message sent from Bob to Alice and drop it.
5. **Masquerading.** Eve may masquerade as Bob and send messages to Alice pretending she was Bob. This is in the realm of identity theft.
6. **Disconnecting.** Eve may completely sever the connection between Alice and Bob so that no further messages can be transmitted.

The question lends itself to *where* we can find Eve in Biology. There is an interesting aspect to the parallel of Eve, as a cryptographic entity, and its counter character in Biology. We can *find* Eve at every level of biological scrutiny, that is to say, both at sub-cellular, as well as cellular and multicellular levels within the hierarchy of life.

The next section explores a few encounters of Eve in nature, as the above cryptographic identity. They show the range and diversity of parallels that can be drawn between cryptography and biology.

3 The presence of Eve in nature

3.1 Eve at high levels of biological organization

Let us first explore cryptographic needs during the interaction of organisms such as animals or plants. Organisms typically communicate through visual, acoustic, or chemical signals. We think that many of Eve's actions as described in subsection 2.2 can be found in interfering with a variety of modes of communication.

Eavesdropping. Eavesdropping is pervasive among animal predators. A predator, such as a large cat, stalks its prey before bouncing on it. Stalking implies listening, or studying the prey's behavior, and also trying to conceal the presence of the predator. Analogously, Eve in cryptography listens to the communication channel between Alice and Bob and endeavors to hide her action.

Masquerading. Masquerading was described as the attempt of Eve to present herself as Bob. Eve sends messages to Alice pretending that she is Bob. Eve achieves this by forfeiting Bob's signature on a message, or more generally, Eve exhibits Bob's characteristics.

Many animal and plant species have evolved to take the visual aspect of another species or of an inanimate object, in order to gain some advantage over their predators or prey. Such phenomena of deceit are called *mimicry* in biology[15], and were first described over a century ago by Henry Walter Bates [1] as he studied butterflies in the Amazon forest.

The biological concept of mimicry is very large and encompasses different types of signaling and behavior. In *aggressive mimicry*, a predator aims to hide under the characteristics of a harmless species or object. It is the "wolf wearing a sheep's skin". Consider the North American *Photuris* firefly [15]. The *Photuris* female attracts males of another firefly genus, namely *Photinus*, by emitting light flash patterns that mimic those emitted by *Photinus* females. When a *Photinus* male mistakenly approaches a *Photuris* female, he is eaten. Thus, *Photuris* females mimic the *behavior* of another genus, by sending out wrong signals.

Defensive mimicry aims to protect a species against its predator, while masquerading as a dangerous or unpalatable species. There are many cases of defensive mimicry. For example, there is a snake species called the false cobra, *Malpolon moilensis*. Its venom is mild compared to the Indian cobra, *Naja naja*. Nevertheless, the false cobra has a similar hood to threaten with. The duped enemy usually backs off at the false threat.

Inserting a false message. In this case Eve sends a false message to Alice. A strategy used by some birds, fish, or insects is *brood parasitism*. In this behavior individuals of a host species are manipulated to raise the young of another, called the brood parasite. For instance, females of the North American brown-headed cowbird *Molothrus ater* lay their eggs into the nests of a large number of other species. The parasitic young compete with their foster siblings for parental care. By begging for food more intensely and loudly [6], the parasites have an advantage in attracting the attention of the parents.

3.2 Eve at the molecular level

When we think of Eve in terms of cryptography, Eve has the full characteristics of a person. She has her own will, has intentional actions, is intelligent, cunning and shrewd.

Accidental damage of the communication channel between Alice and Bob is possible and has to be considered. Nevertheless, the treatment of accidental failures of the communication is rather a problem of technical reliability in a possibly adverse physical environment, not so much a problem of security. It is rather the "human" characteristics of Eve that bring us into the realm of cryptography. The enemy is an enemy that thinks and acts based on her will. Also, Eve understands the unencrypted content of a message as a human would do ... "Eve knows English".

When we deal with biological entities, especially at the suborganismal and subcellular levels, it is rather a stretching of the mind to consider an enemy with the proper characteristics of evil intentions and intelligence. To be able to keep the same setting as we are accustomed to in cryptography, the person of Eve has to be understood in a larger context. Eve would become a dummy person, responsible of any destructive action on the integrity and life of a cell. In this context, such attacks would not have a real intention behind them, but may be defined as attacks with a physical or chemical, or even biological cause. Such attacks can be repetitive, forcing the cell to develop methods to protect itself from them.

The most important message and information carrying molecules found in cells is DNA, which constitute the genome of an organism. Chemically, DNA is a linear polymer consisting of a phosphate-sugar backbone. To each sugar one of four nitrogenous bases, adenine, thymine, cytosine, and guanine, is attached. The linear sequence of bases within a DNA molecule allows for an enormous number of possible combinations. The sequence itself encodes the information carried by the molecule, and constitutes the genetic message. Within a cell, DNA is present as a double helix, consisting of two complementary strands, which means that if we know the sequence of one strand, we can deduce the sequence of the other. The information stored in DNA is used by cells in two ways. First, through a process called *transcription*, selected short stretches of sequence are copied into RNA, another type of information carrying molecule. RNA transcripts then go on to support all cellular functions. Second, before a cell can divide, each DNA molecule undergoes *replication*, by which two identical copies are created. The two copies are distributed to the two daughter cells during cell division.

Damage to DNA may be physical or chemical, i.e. lesions produced by endogenous and exogenous agents. Endogenous agents have the source inside the cell or organism, whereas exogenous agents originate in the environment. Even if the damage is physical or chemical in nature, it has biological consequences [5]. It affects the health of the cell and consequently of the organism and may fully inhibit the replication process of the DNA molecule. Some common physical agents are ultraviolet light and ionizing radiation (e.g. X-rays and gamma rays). Chemical agents that affect DNA sequence integrity are called mutagens. They have diverse modes of action, and many cause direct damage to DNA through specific chemical reactions. Others, however, interfere with replication, causing errors during copying of the sequence into newly synthesized DNA.

A remarkable class of agents that affect DNA integrity act by altering the sequence itself, without producing structural damage to the DNA molecule. Such agents reside at the boundary between living and nonliving matter, as they exhibit some, but not all characteristics of living organisms.

One such type of agent is represented by transposons, also called transposable elements or "jumping genes" [12]. Transposons are relatively short sequence segments found in the DNA of many species. They are mobile, in the sense that they can insert themselves into a new location within the DNA molecule through a mechanism of either "cut and paste" or "copy and paste". When a transposon moves, it alters the sequence of the DNA at the old and/or new location, which possibly has consequences on biological function and is a threat to genome integrity.

Certain viruses are sequence-altering agents as well. Viruses are biological entities that have a defined structural organization, have their own DNA or RNA, and are able to reproduce. However they are not composed of cells, and do not support their own metabolism, and thus lack some key characteristics of living things. Some types of viruses, such as the bacteriophage λ [4], or the HIV virus [11], insert their own sequence into the DNA of host cells. Through this

process they are able to alter the function of the host cell, hijacking the cellular machinery for their own purpose.

4 Electronic and Molecular Information Safety

The present section is dedicated to the actions that Alice and Bob may take to ensure or at least improve the reliability of their communication. Problems of protecting the message contents, correcting errors and mistakes, recovering as much as possible from the initial message are inherent to communication processes. The following is a non-exhaustive study of reliability or safety issues. The purpose is to parallel the two information holders: electronic and biological, in view of the fact that the problems faced by them are similar.

4.1 Protected Public Information

The first action that would be considered here is to *protect* the information from attempted change. Protecting information is rarely discussed explicitly in cryptography, but is nevertheless presupposed in several instances, such as public key cryptosystems.

Most commercially successful cryptosystems rely on public key cryptosystems, such as the Rivest-Shamir-Adleman (RSA) protocol [13]. Acceptable security levels are reached using “one-way” functions, functions that are easy to compute but difficult to invert. Such a system needs two keys: a public key and a private key. Bob encrypts a message with Alice’s public key, and then sends the message to Alice. Alice decrypts the message using her private key. Note that, the public key, as the name suggests, is visible and known to everyone, including Eve. Nevertheless, the message is unintelligible to anyone unless it is decrypted with the private key, which is known only to Alice. In order for the protocol to work, the public key is guaranteed to be protected, unchangeable. There is a consensus about the public key value. Eve is not allowed to change the public key value, or else Bob may not correctly encrypt the message. It is a *strong* requirement in public key cryptosystems, that public information *can* be protected from interference. This may not seem theoretically so obvious but works acceptably in practice. For example, if many copies of Alice’s public key exist, such as on the internet, and in several other public multi-media of a large audience, it can be assumed that Eve cannot control *all* public channels of communication.

Similar mechanisms of protecting information can be found at the molecular level, concerning the DNA molecule. The structure of the molecule itself has properties that offer protection in a possibly adverse chemical environment. In the DNA double helix, the phosphate sugar backbone faces the outside, while the information-carrying bases are hidden inside. Thus the bases themselves are *protected* from chemical attack [3].

Considering the idea that public information can be protected by keeping it in many copies, cells and organisms have several methods to provide the same kind of redundancy. Bacterial cells often contain several replicas of their

DNA. In organisms where DNA is organized into multiple chromosomes and packaged into a nucleus, chromosomes typically exist in pairs in each cell, such that there are two copies of each DNA sequence available at all times. If one of the copies is altered and loses a certain gene function, the other copy can often times compensate. Tissues such as skin or bone contain large numbers of cells, thousands to millions, and thus the same number of copies of the full DNA complement of a single cell. If the information in any single cell is damaged, the cell is destroyed by the immune system and replaced through cell division from a normal cell.

A way of protecting information is to keep it in different formats, or on different hardware supports. Backups are very usual for humanly manipulated information. An interesting analogy is represented by the two strands that make up the DNA double helix. They contain the same information, but in two complementary formats. At any particular location along the DNA, only one of the strands is biologically functional. The other serves as the "backup". If either of the strands is damaged, it can be reconstructed from the complementary strand. DNA repair mechanisms exist for various situations [10].

Another strategy for protecting information is to identify and "quarantine" changes to the original message. Cells have mechanisms by which they can inactivate defined regions of their DNA. Inserted foreign sequences, such as transposons or viral DNA, can be identified and silenced. One such mechanism found in animals and plants involves a chemical modification, i.e. methylation, of cytosine in the DNA molecule. Transposon sequences are specifically identified and methylated, thus preventing them to jump and insert themselves into new locations [9].

4.2 Error Correcting or Repair Mechanisms

If binary information is transmitted over an unreliable channel, the message may reach its destination in a corrupted form. If the message has been partially altered by faulty transmission, the correct message needs to be reconstructed. The field of error correcting codes [7] aims to develop encoding techniques that allow for errors to happen during transmission, while preserving the full content of the message. Shannon proved [14] that at a rate below the capacity of the communication channel, the message can be sent with arbitrarily high accuracy. Probably the best known error code is the Huffman code. Error codes protect the message in that some n bits of information are encoded into m ($m > n$) bits to be sent across the channel.

Similarly, DNA information needs to be corrected when damaged. Molecular mechanisms that deal with reconstructing the DNA molecule after a damage are called repair mechanisms [10]. Some repair mechanisms deal with specific damage types, others work more generally.

For example, ultraviolet light of type UV-C and UV-B produces a specific damage on DNA chains for which a specific repair mechanism exists [8]. The range of UV-C and UV-B radiation is from 180 to 320 nm, and includes the DNA absorption maximum at 260 nm. When two adjacent thymine bases absorb

a photon, they bond covalently forming a dimer. This results in a structural distortion of the DNA double helix that physically prevents replication. Through a process called photoreactivation, the covalent bond in the thymine dimer is reversed, and the DNA strand is directly repaired. Photoreactivation occurs in most organisms (but not in humans), and requires the action of the enzyme photolyase, which depends on the presence of light in the range of 313-475 nm.

Another example of a DNA repair mechanism is nucleotide excision repair [2]. Faults in the pairing of complementary strands are detected by the presence of distortions in the DNA structure. A stretch of one strand around a distorted area is cut off from the double helix. Subsequently, the enzyme DNA polymerase fills out the missing part according to its complementary strand. In the end, DNA ligase seals the nicks and thus completes the sugar-phosphate backbone of the repaired strand.

5 Conclusion

We have shown that some rules that apply to secure communication among humans are directly translatable to rules in biology, at different hierarchical levels of organization. As computer science and biology have developed separately, meaning in their own scientific community, we observe that the language that describes similar concepts, naturally differs from the computer science community to the biology community.

This paper explores basic cryptographic needs for molecular biology, such as protecting information and correcting errors. At the higher level of interacting organisms, the eavesdropper has a higher level personality, including the capability of masquerading.

With the advent of quantum explanations to biological processes, we may well look forward to find mechanisms of quantum cryptography imbued in the processes of life.

References

1. Henry Walter Bates. Contributions to an insect fauna of the Amazon valley. lepidoptera heliconidae. *The Transactions of the Linnean Society of London*, 23:495-566, 1862.
2. Dawn P Batty and Richard D Wood. Damage recognition in nucleotide excision repair of dna. *Gene*, (241):193-204, 2000.
3. Stephen R. Bolsover, Jeremy S Hyams, Elizabeth A Shephard, Hugh A. White, and Claudia G. Wiedemann. *Cell Biology: A Short Course*. John Wiley & Sons, Hoboken, New Jersey, 2004.
4. Allan Campbell. Phage integration and chromosome structure. a personal history. *Annu Rev Genet*, (41):1-11, 2007.
5. Nicholas E. Geacintov and Suse Broyde. *The Chemical Biology of DNA Damage*. Hiley-VCN Verlag, Weinheim, 2010.
6. M E Hauber. Lower begging responsiveness of host versus parasitic brown-headed cowbird (*molothrus ater*) nestlings is related to species identity but not to early social experience. *J Comp Psychol*, (117):24-30, 2003.

7. W. Carry Huffman and Vera Pless. *Fundamentals of Error Correcting Codes*. Cambridge University Press, Cambridge, 2003.
8. John Jagger. Photoreactivation. *Photobiological Sciences Online (KC Smith, ed.) American Society for Photobiology*, page <http://www.photobiology.info/>, 2008.
9. Julie A Law and Steven E Jacobsen. Establishing, maintaining, and modifying dna methylation patterns in plants and animals. *Nat Rev Genet*, (11):204–220, 2010.
10. Benjamin Lewin. *Genes VIII*. Pearson Prentice Hall, Upper Saddle River, 2004.
11. Mary K. Lewinski and Frederic D. Bushman. Retroviral DNA integration—mechanism and consequences. *Advances in Genetics*, 55:147–181, 2005.
12. Hitoshi Nakayashiki. The trickster in the genome: contribution and control of transposable elements. *Genes Cells*, (16):827–841, 2011.
13. Ronald L. Rivest, Adi Shamir, and Len M. Adleman. A method of obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978.
14. C. Shannon. A mathematical theory of communication. *Journal of Bell System Technology*, (27):379–423 and 623–656, 1948.
15. Wolfgang Wickler. *Mimicry in plants and Animals*. MacGraw-Hill, New York, 1968.

Comparison of Sequence Similarity Measures for Distant Evolutionary Relationships

Abhishek Majumdar, Peter Z. Revesz

Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE. USA

Abstract—Sequence similarity algorithms are used to reconstruct increasing large evolutionary trees involving increasingly distant evolutionary relationships. This paper proposes two sequence similarity algorithms, called the Greedy Tiling and the Random Tiling algorithms, that are both based on the idea of tiling one sequence by parts of another sequence. Experimental comparisons show that the new algorithms are better at detecting distant evolutionary relationships than the Needleman-Wunsch sequence similarity algorithm.

Keywords: bioinformatics; Needleman-Wunsch; protein; sequence; similarity.

1. Introduction

Sequence similarity in genetics is often used to identify homologous genes, that is, genes which have evolved from a common ancestry. Similarly, sequence similarity of proteins allows identification of homologous proteins whose encoding genes evolved from a common ancestor. Therefore, similarly to the case of genes, biologists can describe an evolutionary hierarchy of proteins.

There are several ways of measuring the similarity between pairs of proteins [1]. Most protein similarity algorithms are based on the alignment of the sequences of the amino sequences. Such sequence similarity algorithms include Needleman-Wunsch [4], Smith-Waterman [9], and its extension by Gotoh [2]. Other protein similarity measures consider the 3-D structure of the proteins, especially the binding sites of the proteins, to determine their similarity [5,8]. In this paper we are only interested in sequence similarities because while sequence information is commonly available in databases because the 3-D structure of most proteins is still unknown [10].

Although sequence similarity plays a major role in genetics, there is little information about the relative reliability of various similarity measures, which is a general problem in data integration [7]. This project proposes two novel sequence similarity algorithms, called the Greedy Tiling and the Random Tiling algorithms, and compares their effectiveness with older similarity measures in recreating the evolutionary hierarchy of related proteins. Both of the tiling algorithms implement the tiling similarity measure that was non-algorithmically defined by Revesz [6] based on the idea of tiling one sequence by parts of another sequence.

This paper is organized as follows. Section 2 describes two new algorithms for finding the tiling similarity of two sequences. Section 3 describes experimental results. Section 4 analyses the results. Finally, Section 5 concludes the paper.

2. Implementations of Tiling Similarity

Revesz [6] introduced the tiling similarity measure, which is based on the idea of tiling one sequence with parts of the other sequence. The tiling similarity value depends on finding the optimal tiling and is an intractable problem for large sequences. Nevertheless, we give below two algorithms that in many cases give a good approximation of the optimal tiling. Our approximation algorithms, called Greedy Tiling and Random Tiling, both run efficiently even on large sequences.

2.1 Greedy tiling

Given as input two protein sequences X and Y with $\text{length}(Y) \leq \text{length}(X)$, *GreedyTiling* tries to reconstruct Y using segments, called *tiles*, from X . This is done using the algorithm with the following pseudocode:

Greedy Tiling *GreedyTiling*(X, Y, Tiling)

1. $X' = X$
2. $Y' = Y$
3. $\text{Tiling} = \emptyset$
4. $i = 0$
5. **while** Y' is not empty **do**
6. $i = i + 1$
7. Smith-Waterman(X', Y', x_i, y_i)
8. $X' = X' - x_i$
9. $Y' = Y' - y_i$
10. **if** x_i and y_i are subsequences of X and Y **then**
11. $\text{Tiling} = \text{Tiling} \cup \{(x_i, y_i)\}$
12. **else** split x_i and y_i into proper subsequences
13. $x_i = x'_i \mid x''_i \mid \dots$
14. $y_i = y'_i \mid y''_i \mid \dots$
15. $\text{Tiling} = \text{Tiling} \cup \{(x'_i, y'_i), (x''_i, y''_i), \dots\}$
16. **end-if**
17. **end-while**

The above algorithm assumes that we have the function Smith-Waterman(X, Y, x, y) that finds the best locally matched segments x in X and y in Y , when given as

input the sequences X and Y . LCS^* repeatedly calls the Smith-Waterman algorithm to find the longest common subsequences between X' , the remaining X , and Y' , the remaining Y . In each iteration, the pair of longest common subsequences x_i of X' and y_i of Y' are added to the set of tiles and deleted from X' and Y' . Each segment x_i and y_i is inspected whether it is a proper subsequence of X and Y , respectively, or a concatenation of two or more parts of X and Y . Accordingly x_i and y_i are broken up into its constituent components as necessary and added to the tiles as a set of pairs. This process is repeated iteratively until Y' is empty.

Example 1: Suppose we have the following two sequences: $X = WARICDFLRE$ and $Y = FIREICEWAR$.

In the first iteration, the Smith-Waterman algorithm finds between $X' = X$ and $Y' = Y$ the best local alignment to be $x_1 = WAR$ and $y_1 = WAR$, which are proper subsequences of X' and Y' . Hence x_1 and y_1 are added as a pair to $Tiling$ and deleted from X' and Y' to yield $X' = ICDFLRE$ and $Y' = FIREICE$, respectively.

In the second iteration, the best matching segments are $x_2 = FLRE$ and $y_2 = FIRE$, which are also proper subsequences of X' and Y' . Deleting those yields $X' = ICD$ and $Y' = ICE$.

In the third iteration, the best matching segments are $x_3 = ICD$ and $y_3 = ICE$, which are also proper subsequences of X' and Y' . Deleting y_3 from Y' will make it empty. Hence the algorithm terminates. Hence in this case, LCS^* will return the following:

$$Tiling = \{(WAR, WAR), (FLRE, FIRE), (ICD, ICE)\}$$

As a measure of the similarity between X and Y , we use the *tiling similarity*, or *TS*, measure of Revesz [6], which is defined as follows:

$$TS(X, Y) = \frac{\sum_{i=1}^{i=n} s_i}{n}$$

where n is the number of segments used for reconstruction and s_i is the similarity score between tiles x_i and y_i . For example, if we use the BLOSUM62 similarity matrix, then:

$$s_1 = sim_{BLOSUM62}(WAR, WAR) = 20$$

$$s_2 = sim_{BLOSUM62}(FLRE, FIRE) = 18$$

$$s_3 = sim_{BLOSUM62}(ICD, ICE) = 15$$

Hence the tiling similarity will be:

$$TS(X, Y) = \frac{20+18+15}{3} = \frac{53}{3} = 17.66$$

2.2 Random tiling

The second algorithm uses a randomized approach to find the different segments/tiles required for the reconstruction of sequence Y . It randomly breaks up sequence X into tiles of different lengths. Then filters out a select few using a constraint for a valid range of tile-length. Finally it uses this selected set of tiles (say $x_1, x_2, x_3, \dots, x_n$) to match the different portions of Y . Since this approach is randomized the entire process needs to be iterated an arbitrary number of times, each time with a set of randomly generated tiles, and the tiling with the highest tiling similarity score selected. Below we give only the pseudocode of the basic algorithm that needs to be repeated.

Random Tiling *RandomTiling*($X, Y, Tiling$)

1. Split X into a random set of tiles $T(X)$
2. $Y_U = Y$
3. $Tiling = \emptyset$
4. **while** Y_U is not empty and longer than the shortest tile **do**
5. $BestScore = -100$
6. **for each** tile $x_i \in T(X)$ **do**
7. $y_m = \text{prefix of } Y_U \text{ with } length(x_i)$
8. **if** $BestScore < sim(x_i, y_m)$ **then**
9. $BestScore = sim(x_i, y_m)$
10. $BestPair = (x_i, y_m)$
11. **end-if**
12. $Tiling = Tiling \cup BestPair$
13. **end-while**

In each iteration we begin the tile-matching from the leftmost end of Y . Let Y_U denote unmatched section of Y . Clearly, initially $Y_U = Y$. For each tile x_i from the tile set we match it with left most segment y_m of Y_U which is of same length as x_i . We always select the tile which gives the highest matching score. In the next iteration we update Y_U by deleting from it the initial segment y_m . Then we continue the tile-matching process with the updated Y_U . This iteration is carried out from left to right until Y is fully matched. In the last iteration, if there is a case that the length of the current Y_U is less than the length of smallest tile then that remaining Y_U is matched with gaps.

Example 2: Consider the following two sequences:

$X = ABCDEFGHIJKLOIYITB$ and

$Y = WUFGDJVMBKUG$.

We will reconstruct Y using tiles from X . Let the tiles obtained from X be $x_1 = BCDE$, $x_2 = IJKL$, $x_3 = DEFGHI$, $x_4 = LOIYITB$, and $x_5 = AB$. Initially $Y_U = WUFGDJVMBKUG$. We start by matching each tile x_i with left-most portions of Y_U which is of same length as x_i . That is we match x_1 with WUFG, x_2 with WUFG, x_3 with WUFGDJ, x_4 with WUFGDJV and x_5 with WU. Say x_1 gives the best

matching. So now Y_U becomes DJVMBKUG. The above process is repeated again. That is x_1 matched with DJVM, x_2 with DJVM, x_3 with DJVMBK, x_4 with DJVMBKU and x_5 with DJ. Let the best tile be x_4 . So now $Y_U = G$. This is matched with a gap as its length is less than that of x_5 . So the reconstructed Y looks like:

$Y = WUFG | DJVMBKU | G$
 $X = BCDE | LOIYITB | -$

We can repeat the above process an arbitrary number of times and select the tiling which gives the highest tiling similarity score. Obviously, the more the basic algorithm is repeated, the higher tiling similarity is found. However, there is a trade-off between repetitions and increased tiling similarity values. There is a point where the increase in execution time may not be worth the diminishing chance of an increase in the tiling similarity score.

3. Experimental Results

In the experiments we focused on the Type III Pyridoxal 5-phosphate(PLP) dependent enzymes subfamily. This is important and well-studied subfamily is composed mainly of proteobacterial alanine racemases that help in the inter-conversion between L- and D-alanine, which is an essential component of the peptidoglycan layer of bacterial cell walls. Figure 1 shows a small portion of this subfamily hierarchy as described in the National Center for Biotechnology Information (NCBI) Conserved Domain Database [3].

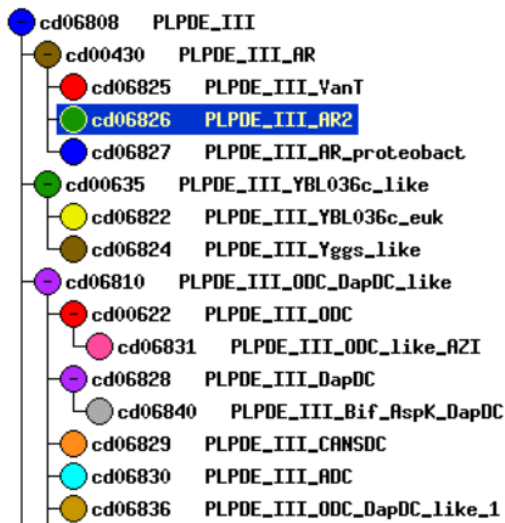


Fig. 1: Hierarchy tree.

Each node in Figure 1 is a also cluster of subsequences. That is, each node is composed of closely related bacterial genome sequences, which have a hierarchical relation among themselves as well. For instance, the node cd06825 is actually composed of nine sequences shown in Figure 2.

Figure 2 shows the gi version numbers (164602518, 44805037, etc.) which uniquely identify each sequence. The

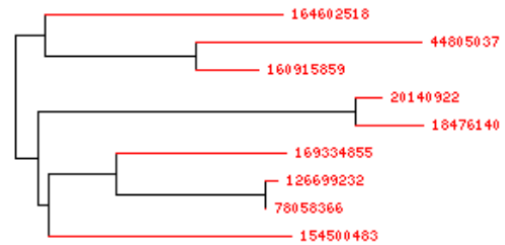


Fig. 2: cd06825 cluster

FASTA sequence description was also obtained from the NCBI website and used as input to our similarity algorithms and to the Needleman-Wunsch algorithm.

Both our algorithms take all possible combinations of two subsequences from the clusters to measure their tiling similarity scores (TSs). For example, the cd06825 cluster contains 36 pairs of sequences and yields as many similarity scores. Because of the large set of data, our experiments focused on the following five randomly selected clusters: cd06815, cd06817, cd06822, cd06825 and cd06826. The size of each cluster (in terms of number of constituent sequences) and the number of associated tiling similarity TS score combinations obtained is shown below in Table 1.

Table 1: Cluster details.

Cluster	Size	Score Combinations
cd06825	9	36
cd06826	11	55
cd06817	15	105
cd06822	36	630
cd06815	39	741

We define below the following relationship terminologies that are used in comparison of the similarity measures:

Siblings are sequences that are separated by only one evolutionary branching from a common ancestor in the evolutionary family tree. For instance, in Figure 2 sequences 44805037 and 160915859 are separated by a single evolutionary step from their common ancestor, hence they are siblings.

First cousins are sequences that are separated by at most two branching from a common ancestor in the evolutionary family tree. For instance, in Figure 2 sequences 164602518 and 44805037 are both at most two evolutionary steps distant from the common ancestor, which makes them first cousins.

Second cousins are sequences that are separated by at most three branching from a common ancestor in the evolutionary family tree. Again in Figure 2 sequences 164602518 and

20140922 are second cousins.

i^{th} **cousins** are sequences that are separated by at most $(i+1)$ branching from the closest common ancestor.

In our experiments, we compared the Needleman-Wunsch algorithm [4], the greedy tiling, and the random tiling algorithms. We ran for each of the five above listed clusters each of the three algorithms. We calculate the similarity scores between siblings, first cousins and second cousins. For the calculation of the similarity scores, we used the common PAM250 substitution matrix [1] and a constant value of -8 as the gap-penalty. In addition, for the random tiling algorithm, the larger sequence X is always divided into 70 segments randomly to generate the available tiles $T(X)$. In the case of the random tiling algorithm, we ran the basic algorithm 1000 iterations before selecting the tiling that gave the highest score. The scores Needleman-Wunsch, the greedy tiling, and the random tiling are shown in Tables 2, 3, and 4.

Table 2: Needleman-Wunsch similarity scores.

Sequence	Siblings	First Cousin	Second Cousin
cd06815	860.20	704.88	653.56
cd06817	672.00	333.17	76.96
cd06822	496.50	115.25	143.17
cd06825	2050.33	-199.14	-1593.89
cd06826	987.75	985.43	815.75

Table 3: Greedy Tiling similarity scores.

Sequence	Siblings	First Cousin	Second Cousin
cd06815	472.87	353.09	321.17
cd06817	391.52	313.01	190.65
cd06822	357.96	236.28	175.46
cd06825	1499.11	327.42	128.24
cd06826	325.48	387.07	289.09

Table 4: Random Tiling similarity scores.

Sequence	Siblings	First Cousin	Second Cousin
cd06815	37.07	23.90	20.93
cd06817	56.23	44.63	42.90
cd06822	22.88	25.82	27.36
cd06825	68.57	49.08	83.88
cd06826	39.55	33.65	23.44

We also show the same results as a set of graphs in Figures 3, 4, and 5.

4. Discussion of the Results

The essential difference between the Needleman-Wunsch and the tiling similarity measures is that the Needleman-Wunsch method is good for random mutations, insertions

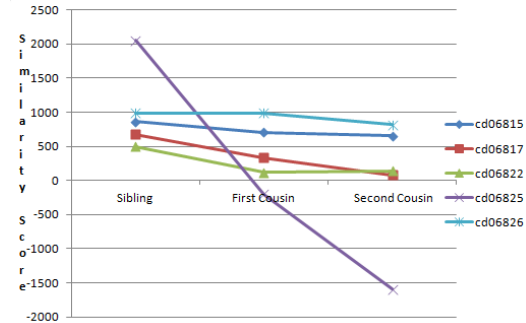


Fig. 3: Needleman-Wunsch similarity scores.

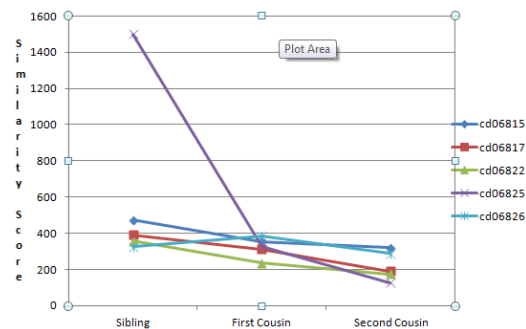


Fig. 4: Greedy Tiling similarity scores.

and deletions but is not good for reordering of parts of the sequences. In contrast, the tiling similarity measures are designed to be able to detect similarities in case of reordering. For example, recall that for the sequences $X = \text{WARICDFLRE}$ and $Y = \text{FIREICEWAR}$ a high tiling similarity was found in Example 1. In this case, it is possible to imagine a common ancestor $A = \text{FLREICDWAR}$ that branches and develops first as

$$\text{FLREICDWAR} \rightarrow_{\text{mutate LI, D/E}} \text{FIREICEWAR}$$

and second as

$$\text{FLREICDWAR} \rightarrow_{\text{switch WAR/FLRE}} \text{WARICDFLRE}$$

yielding, therefore, Y and X , respectively.

Transpositions of parts of the genome are known to occur and would be reflected also in the amino acid sequences of the corresponding proteins. While mutations are expected to be much more frequent than such transpositions, they may not be enough to explain very distant evolutionary relationships because over large evolutionary distances some transpositions may also occur. The proteins we studied are considered ancient proteins because they help build the bacterial cell wall, which is an essential part of bacteria. Hence some transpositions may have occurred in various branches of this ancient evolutionary tree.

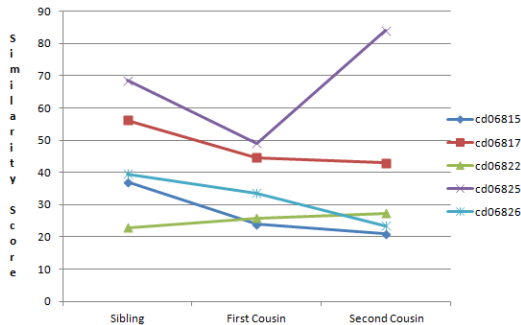


Fig. 5: Random Tiling similarity scores.

The natural expectation for all the similarity measures was the following:

1. All the similarity values were positive.
2. The average similarity among siblings was higher than among first cousins which was higher than among second cousins.

The **Needleman-Wunsch algorithm**, as shown in Figure 3, did not fulfil these expectations in three instances. In two instances, it gave a negative similarity value, namely for first and second cousins for cluster cd06825. In addition, for cluster cd06822 the similarity for first cousins was significantly less (115.25) than the similarity for second cousins (143.17).

The **greedy tiling method** gave only positive scores. The average scores for first cousins were always larger than the average scores for second cousins. The only anomaly was in cluster cd06826 where the average sibling similarity was slightly less (325.48) than the average first cousin similarity (387.07).

The **random tiling method** also gave only positive scores. The average scores for first cousins were less than the average scores for second cousins in the case of two clusters, namely, cd06822 and cd06825. In the case of cd06822, the average sibling similarity was also slightly less than the average first cousin similarity. Hence the random tiling method did not fulfil the expectations in three instances.

Therefore, our experiments suggest that the greedy tiling method is the most robust method, especially comparing larger evolutionary distances (first cousins versus second cousins). The random tiling method seems intermediate in performance. Probably it can be improved to be as good as the greedy tiling method by increasing the number of times its basic algorithm is repeated. Finally, the Needleman-

Wunsch algorithm was good in comparing shorter evolutionary distances (siblings versus first cousins) but deteriorated considerably in comparing longer evolutionary distances (first cousins versus second cousins).

The experimental results suggest that the tiling similarity measure is better than the Needleman-Wunsch measure for distant evolutionary relationships. Intuitively, the reason seems to be that the tiling similarity allows transpositions of a subsequence on the genome. These transpositions may be only relatively rare evolutionary changes compared to random mutations, Nevertheless, if a significant number of transpositions accumulate in at least one branch of a large evolutionary tree, then the Needleman-Wunsch algorithm may be unable to detect them and give a low (even negative) similarity score for distantly related sequences. Based on the experimental results, we suspect that cluster cd06825 may contain some transpositions because the Needleman-Wunsch algorithm gave negative similarity scores for first cousins and second cousins, but both of the Greedy Tiling and the Random Tiling algorithms gave positive scores. Further, in the Random Tiling method the average similarity increased from first cousins to second cousins for the same cluster.

The above type of anomaly may be explained in an example. Suppose that in an evolutionary tree branch A has some transpositions that are not shared with its first cousin branch B and also not shared by A and B's second cousin branch C. In this case, the similarity between A and B, which is a first cousin similarity, could be lower than the similarity between B and C, which is a second cousin similarity. Hence if the evolutionary tree is extremely simple and has no other first cousin pairs and no other second cousin pairs beside A and C and B and C, then the average similarity among first cousins could be less than the average similarity among second cousins. The larger the evolutionary tree, the less likely such anomalies could occur. It is important to note that the cd06825 cluster is the smallest in size as shown in Table 1.

5. Conclusion and Future Work

We need to investigate further the reasons why the tiling similarity measure is better than the Needleman-Wunsch similarity measure for distant evolutionary relationships. In particular, it would be interesting to find actual examples of transpositions of subsequences within any of the clusters.

Another direction for further experiments would be to consider even larger evolutionary trees where we have enough data for third and fourth cousins. Experiments on such a larger data could show clearer the differences among the similarity measures. We suspect that the Needleman-Wunsch algorithm will perform even poorer on higher cousins but the tiling similarity algorithms will keep detecting well the more distant evolutionary relationships.

References

- [1] R. Durbin and S. R. Eddy and A. Krogh and G. J. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.
- [2] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, vol. 162, no. 3, pp. 705–708, 1982.
- [3] (March 20, 2012) The National Center for Biotechnology Information. [Online]. Available: <http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=143500>
- [4] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.
- [5] R. Powers and J. Copeland and K. Germer and K. Mercier and V. Ramanathan and P. Z. Revesz, "Comparison of Protein Active-Site Structures for Functional Annotation of Proteins and Drug Design," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 124–135, 2006.
- [6] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, Springer-Verlag, 2010.
- [7] P. Z. Revesz and T. Triplet, "Classification Integration and Reclassification using Constraint Databases," *Artificial Intelligence in Medicine*, vol. 49, no. 2, pp. 79–91, 2010.
- [8] M. Shortridge and T. Triplet and P. Z. Revesz and M. Griep and R. Powers, "Bacterial Protein Structures Reveal Phylum Dependent Divergence," *Computational Biology and Chemistry*, vol. 35, no. 1, pp. 24–33, 2011.
- [9] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.
- [10] T. Triplet and M. Shortridge and M. Griep and J. Stark and R. Powers and P. Revesz, "PROFESS: a PROtein Function, Evolution, Structure and Sequence database," *Database – The Journal of Biological Databases and Curation*, doi no. 10.1093/baq011, 2010.

Bioinformatics: an overview for cancer research

M. Al-Rajab¹, J. Lu¹

¹School of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom

Abstract - *Bioinformatics is a new science that is glowing out in the recent years. It is a multidisciplinary science that is made out of different kinds of other scientific fields like biology, computer science, chemistry, statistics, mathematics and others. It was a big challenge for researchers to describe this new field in a systematic scientific way and bring out the attention of its applications and services; one of these important services that Bioinformatics can be applied in, is the cancer studies, research and therapies for many beneficial reasons. This paper will give a clear glance overview of bioinformatics, its definition, aims, applications, technologies, the large amount of data produced in the biological field and how bioinformatics can organize, analyze and store them, discuss some algorithms that can be implemented over bioinformatics data, and how to apply bioinformatics to discover and diagnose diseases like cancer.*

Keywords: Bioinformatics, Applications, Technologies, Data, Algorithms, Cancer.

1 Introduction

Bioinformatics is a new multidisciplinary field that comes out from the combination of other sciences and fields like biology, computer science, statistics, chemistry, mathematics and even more [3, 6, 8, 9, 14, 15, 16, 17]. In recent years new sciences have risen up due to the demand in understanding more the world around us like Bioinformatics, Biotechnology, Computational Biology, Biochemistry and others. It was a big challenge for researchers and scientists to give an adequate definition for each of these newly emerged sciences [5, 9, 18]. One of these sciences that have a huge influence in the medical field is Bioinformatics but also can play a key role in other fields like agriculture, livestock and even space explorations [1, 19]. Bioinformatics which attracts people in the academic field in addition an interest to those in the medical industry [4, 15, 20, 21].

There were many contributions to define and explain Bioinformatics in scientific ways, but all researchers agree that it is a combination of Biology, Computer Science, Statistics and Mathematics. Each one of these disciplines is playing an important role for collecting, organizing, analyzing and digitizing the biological data and even classifying and storing it in an efficient manner [1, 3, 12, 16, 19].

The main purpose of this paper is explore and explain Bioinformatics in a more scientific way, the paper will try to define Bioinformatics scientifically and try highlight applications of bioinformatics in the medical sector specially, and in the diagnosis of critical diseases like cancer. The race of bioinformatics research is now passing long rounds in many areas in the Biological life, so; the goal of this paper is to provide an overview summary of bioinformatics definition from different articles written in this field, what are the main implementations and aims under the skin of this science, how to understand the data and what are the most important databases used, give a snapshot over the most common algorithms implemented in the field and how important to apply bioinformatics in the cancer research and study.

This paper will target four categories of readership who are interested in the field. (1) Students who are interested in studying this new field. (2) Instructors who would like to prepare a fundamental course to teach in bioinformatics. (3) Researchers who would like to understand more about Bioinformatics and the relationship with cancer. (4) Experts in the medical field who are interested in implementing the understanding of this field in the medical life.

The remainder of this paper will be structured as follows: Section 2 will discuss the background in methodologies applied in this paper; while Section 3 will focus on Bioinformatics definition, on the other hand section 4 will figure out the aim of studying the field. Moreover in section 5 data, data types and databases will be presented in Bioinformatics. On the other hand, section 6 will discuss the most common Algorithms implemented in Bioinformatics. Section 7 will discuss the role of Bioinformatics in cancer research and how important to be implemented in that field. In section 8 current problems in Bioinformatics are represented, and finally section 9 will conclude this paper.

2 Background in methodologies

As well as sufficient number of papers, articles, websites, and books are talking about Bioinformatics. It was clear to us that all have no unified definition for Bioinformatics as a science or a new born field emerging in the life of biology and technology, add to that there were rare papers systematically constructing and directing the road for all Bioinformatics basic knowledge. From this point an effort was implemented to conduct a deep search to collect as many papers and articles discussing the historical and

fundamentals of Bioinformatics in order to establish a unified basis form understanding the basics of Bioinformatics and links that with importance of applying the field in the cancer study, research and therapy. More than seventy papers, articles, websites and books that are talking about introduction in bioinformatics were collected. A profound reading took place to classify the papers. To write about the basics, we put out all the keywords (bioinformatics, database, algorithms, technologies, cancer, applications), then we started classifying the papers related to the collected data as in Table 1.

Table 1: Summary of Papers Number Read

Topics	No. of Papers
Bioinformatics Definition	49
Databases	12
Algorithms	6
Technologies and Tools	12
Applications	12
Cancer	12

To remark the numbers in the table, 49 references were introducing a definition to Bioinformatics, 12 of them talked about the databases in bioinformatics, 6 discussed the most important algorithms used in Bioinformatics, 12 mentioned out the most important technologies and tools used in the field, the same number discussed where Bioinformatics is applied, and also the same number introduced the relationship of the field with cancer. After that grouped out the data that are relevant together from the different resources and put them together for the literature review and the findings. It was noticed that the different resources collected were not focusing on a basic knowledge of Bioinformatics, they started by defining the field then highlighting one part of the field like databases, tools, applications, algorithms, etc...

Our contribution in this paper is to gather all the distributed fundamental information about Bioinformatics and summarize them in a systematic fundamental way. Jawdat [1] discussed that the storage and analysis of biological data using certain algorithms and computer software is called Bioinformatics, so it was defined as the design, construction and use of software tools to generate, store, annotate, access and analyze data and information related to molecular biology. The authors in [2] said that bioinformatics is basically a study to model, to organize, to understand and to

discover interesting information associated with the large scale molecular biological databases. The term Bio (Molecular Biology) informatics (Information Technology) which encompasses tools and methods used to manage, analyze, and manipulate large set of biological data. In [3] the authors claimed that the use of bioinformatics to organize, manage, and analyze genomic data which is the genetic material of an organism, this new IT discipline fuses computing, mathematics, and biology to meet the many computational challenges in modern molecular biology and medical research. Chavan in [4] argued that biological data include extensive information regarding genomic sequences of different species, changes due to evolution, and changes in their protein sequences. Such a massive data cannot be handled with ease. This requires systematic sieving of data to categorize and catalogue them. Based on this need arose the field of Bioinformatics. So Bioinformatics can be defined as the discipline, which encompasses branches like biology, computer science, IT and mathematics. It is a science of managing and analyzing vast biological data using advanced computing techniques. On the other hand, in [5] the authors commented that defining the terms bioinformatics and computational biology in addition is not an easy task. They are both multidisciplinary fields, involving researchers from different areas of specialty, including (but in no means limited to) statistics, computer science, physics, biochemistry, genetics, molecular biology and mathematics. In [6] Zadeh defines bioinformatics as a new discipline that has emerged from the areas of biology, biochemistry, and computer science. Bioinformatics is an interdisciplinary and rapidly evolving field that has emerged from the fields of biology, chemistry and computer science. Add to that Kasabov in [7] said that bioinformatics is concerned with the application and the development of the methods of information sciences for the analysis, modeling and knowledge discovery of biological processes in living organisms. Furthermore in [8] the authors illustrate Bioinformatics as the combination of biology and information technology which focuses on cellular and molecular levels for application in modern biotechnology. So as a result they said that Bioinformatics is the combination of biology and information technology. It is the branch of science that deals with computer based analysis of large biological data sets. Fenstermacher in [9] is defining Bioinformatics as a multifaceted discipline combining many scientific fields including computational biology, statistics, mathematics, molecular biology, and genetics. So Bioinformatics is conceptualizing biology in terms of macromolecules and then applying "informatics" techniques to understand and organize the information associated with these molecules, on a large scale. Moreover, Nair in [10], explained Bioinformatics to be the application of computer sciences and allied technologies to answer the questions of Biologists, about the mysteries of life. In addition the authors in [11] discussed that bioinformatics is a new and rapidly evolving discipline that has emerged from the fields of

experimental molecular biology and biochemistry, and from the artificial intelligence (AI), database, pattern recognition, and algorithms disciplines of computer science. Finally, in [12] the authors summarized the definition of bioinformatics as the application of computer technology to the management of biological information.

3 Bioinformatics Definition

The origin of bioinformatics goes back to Mendel's discovery of genetic inheritance in 1865. Since the 1953, big revolution achievements took place by James Watson and Francis Crick as they determined the structure of DNA [13]. Later in 1960s, the hard work of bioinformatics research started, symbolized by Dayhoff's atlas of protein sequences and the early modeling analysis of protein and RNA structures [12]. After a while, the term Bioinformatics came to sense and use in around 1990s and was described by the management and analysis of DNA, RNA, and protein sequence data. Later in 2000 a big achievement took place which is the announcement of the initial draft of the Human Genome Sequence. Later after 13 years of research and work from 1990 up to spring 2003, in which the official announcement of the Human Genome Sequence Project took place. In this project around 20,000 – 25,000 of human genes were discovered, so the access to this huge amount of gene data and its information was not an easy task for the biologists and for this it opened the doors for a new era in modern biology with an assistant to new computerized technology or in other words the marriage between Biology and Computer Science to bear a new baby known as Bioinformatics which will play a significant role in gathering, analyzing, classifying and storing genetic data collected from the human project or at biological points in a more efficient or powerful way. From here raised the question, what is the importance of Computers in Biology? The accurate answer of this question will be resulted out from the following formula: $\text{Biology} + \text{Computer Science} = \text{Bioinformatics}$. So what is Bioinformatics? What are the main problems that this field can help in?

As a result of the literature review, Bioinformatics can be defined from different perspectives, first from the English Oxford Dictionary, and then from the summary of all researchers' definitions.

Bioinformatics: (According to the Oxford English Dictionary) (Molecular) bio – informatics: bioinformatics is conceptualizing biology in terms of molecules (in the sense of Physical chemistry) and applying “informatics techniques” (derived from disciplines such as applied math, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale.

In short, bioinformatics is a management information system for molecular biology and has many practical applications. So, Bioinformatics can be defined as a new hybrid emerging field of science in which biology, computer science, mathematics, statistics and Information Technology merge and interact together to form a whole new discipline field. It is a science used to manage, analyze, organize, and classify the huge amount of biological data by using well developed algorithms, computational and statistical techniques, designing and construction of software tools and theories to solve different problems arising from biological data and help in generating, storing, accessing and analyzing data and information that are related to molecular biology. Noting that the suffix “informatics” is from European origin; “informatique” means and indicates computer science in French and Bio means Biology [13]. Figure 1 below illustrates all the sciences that make up the Bioinformatics field.

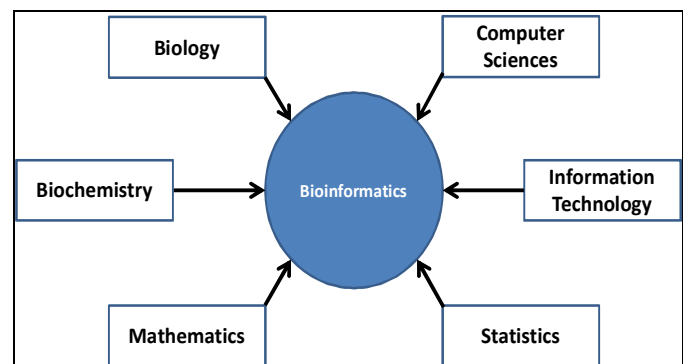


Figure 1: Bioinformatics multidisciplinary sciences

Bioinformatics has four main components: Databases, Computational Tools, Algorithms and Software. Biologists and other related people must be aware of the difference between Bioinformatics and Computational Biology and this is not an easy task, the latter is not a “field” like bioinformatics but it is “an approach” involved in using computers to study biology [9]. So, bioinformatics is concerned with information while computation biology is concerned with hypothesis [14].

4 Bioinformatics Aims

There are five main aims of Bioinformatics [12]:

1. To organize the biological data in an easy manner that helps biologists and researchers to store and access exiting information.
2. To develop and design software tools that help in the analysis and management of data.

3. To use these biological data in the analysis and interpretation of the results in a biological meaningful manner.

4. To assist researchers in the pharmaceutical industry to understand the protein structures that lead and help in the drugs industry development.

5. To help and assist physicians in the medical fields to understand gene structures that will help in detecting and diagnosing disease like cancer.

5 Biological Data, Data types and Databases

Biological Data is often characterized by huge size. There are four important data generated and collected at biological points [10]: DNA, RNA, Protein Sequences, and Micro Array images. The first 3 of them are text data and the last one is a digital image. As the different biological data generated, it can be noticed that these data is represented with different types. There are four types of the data structures [13]: String to represent DNA, RNA, and protein sequences; Trees to represent protein structures; Graphs to represent metabolic and signaling pathways; and Strings (like words and phrases) are also used to express comments that reflect meanings to researchers. Moreover, researchers and biologists are also interested in substrings, subtrees and subgraphs.

The large, huge and complex amount of biological data needed to be stored, accessed and manipulated in an efficient and powerful manner. So it was the need to build Bioinformatics databases which are classified into sequence databases, microarray databases, genome databases, protein structures databases and many more [2].

The sequence databases represent sequence information of all the organisms. GeneBank at the National Center for Biotechnology information, EMBL (European Molecular Biology Laboratory) DNA database, Bethesda and DNA Data Bank Japan (DDNJ), and Protein databases at SWISS-PORT (Protein sequence database at Swiss Institute of Bioinformatics, Geneva) all of them are the largest databanks of the sequence databases. Micro array databases include micro array gene expression under different biological conditions. Example databases of this category are Array Express, and Gene Expression omnibus. Genome databases collect organisms' gene (DNA) sequences. Example of this category databases are Xenbase, Corn, SEED, and RGD. There is another example of Bioinformatics databases that comes from the integration with cheminformatics which is the DrugBank database

(<http://redpoll.pharmacy.ualberta.ca/drugbank>), this database contains 4300 drug entries for and more than 6000 protein sequences which are linked to these drug entries [1].

6 Common Bioinformatics Algorithms [12-13]

This section sheds the light on algorithms that are of interest to bioinformaticians and researchers. The following are some of the most important algorithmic trends in bioinformatics:

1. Finding similarities among strings (such as proteins of different organisms).
2. Detecting certain patterns within strings (such as genes).
3. Finding similarities among parts of spatial structures (such as motifs).
4. Constructing trees (called phylogenetic trees expressing the evolution of organisms whose DNA or proteins are currently known).
5. Classifying new data according to previously clustered sets of annotated data.
6. Reasoning about microarray data and the corresponding behavior of pathways.

7 Bioinformatics Applications in Cancer Research

Cancer is classified as a genetic disease in which the cells cannot follow the sequential phases of the cell cycle and divide in a normal manner. That is cells will lose the control in the cell cycle and starts to divide uncontrollably and the chromosomes of the cancer cells will be arranged incorrectly, or have large pieces missing.

Due to large and fast steps in the medical field research, a lot of efforts are extended in order to find a way to detect, diagnose and treat such hazardous disease. Also the raise of the Human Genome project discovery in 2003 had put more pressure on Bioinformatics to be applied in the cancer therapy. Bioinformatics is now being applied in the cancer research and therapy [21], and it is clear that experts and researchers have implemented rapid and expanded amount of research on the tools of bioinformatics that are considered necessary during the cancer therapies. One of these applications is to use the computerized models that represent biological data and information to know about the quantity of cancer cells in the body or about the biological state of the

patient [22]. Such way has a positive result after the cancer therapy in which experts are now being able to monitor the tumor growth that was not possible earlier during the absence of bioinformatics. In addition, many studies have indicated that gene expression of cancer cells is imperative and this will ensure efficient results after the treatment [9, 23]. Also bioinformatics can be applied to cancer by using the database among the cancer cells' expression and to study the drug response and tumor response also [23]. Until now bioinformatics studies show that it had succeeded in the cases of breast and ovarian cancer and future will insure the effectiveness of bioinformatics in the therapies of other cancer types [24]. Moreover, bioinformatics has made it possible for therapists to analyze immune responses that allow an understanding of the differences between controlled and uncontrolled tumors for better treatment of cancer patients. In other words bioinformatics succeeded in explaining out the effects of the chemotherapy and the radiation therapy with the help of the mathematical models that are part of the bioinformatics discipline. It was noticed that experts and physicians try to use the multiple databases available and the different search engines like Google in order to look for biological data and apply bioinformatics in cancer research and treatment, that due to some organizations and experts limit their work and information and do not allow other experts to benefit from the same work and information. In other words, integration of bioinformatics databases data types, and structures are an important factor to decide the future of Bioinformatics application the medical field science and especially in the cancer treatment and therapies.

The Human Genome Project has enriched the human research community with massive amount of huge biological data and information by the year 2003 [1]. In this case Bioinformatics has found its applications in many areas, and below is a list of some of the important problems where applications in Bioinformatics can be applied in[4, 10]:

- Analyzing DNA sequence data to locate genes.
- Analyzing RNA sequence data to predict their structures.
- Analyzing protein sequence data to predict their location inside the cell.
- Analyzing gene expression images.
- Understanding genetic diseases like cancer, cystic fibrosis, and sickle cell anemia.
- For gene therapy in general.
- In designing drugs for better treatment, and avoid drugs side effects and develop better drug delivery system.

Moreover, NASA's experts are using Bioinformatics in their operations to explore the space and study the universe. So, NASA is also interested in Bioinformatics in their researches and discoveries.

8 Conclusions

The paper tried to give an overview of this multidisciplinary field, by forming a unique clear definition that is introduced by the reaction of Biology and Computer Science in addition to some assessment factors like statistics and mathematics to result into the newly born field "Bioinformatics" after this strong reaction. At the end the paper highlighted the importance of applying bioinformatics in cancer research which will open the horizons for experts and researchers to continue in this specialized field. The future of Bioinformatics will be bright in many biological and life areas, but one of the important issues that must be worked in for this; is the integration of the wide and huge amount of data sources and databases to unify them for better life and for a huge revolution in the biological life as will reaching the moon.

9 References

- [1] Jawdat, D.; , "The Era of Bioinformatics," Information and Communication Technologies, 2006. ICTTA '06. 2nd , vol.1, no., pp.1860-1865, 0-0 0
- [2] Raut, S.A.; Sathe, S.R.; Raut, A.; , "Bioinformatics: Trends in gene expression analysis," Bioinformatics and Biomedical Technology (ICBBT), 2010 International Conference on , vol., no., pp.97-100, 16-18 April 2010
- [3] See-Kiong Ng; Limsoon Wong; , "Accomplishments and challenges in bioinformatics," IT Professional , vol.6, no.1, pp. 44- 50, Jan.-Feb. 2004
- [4] Dr.(Mrs.) Padma R. Chavan; , "Application of Bioinformatics in the Field of Cancer Research", 11th Workshop on Medical Informatics & CME on Biomedical Communication, vol., no., 20-22 November 2008.
- [5] Ackovska, N.; Madevska-Bogdanova, A.; , "Teaching Bioinformatics to Computer Science Students," Computer as a Tool, 2005. EUROCON 2005.The International Conference on Computers as a Tool, vol.1, no., pp.811-814, 21-24 Nov. 2005
- [6] Zadeh, J.; , "An undergraduate program in bioinformatics," Potentials, IEEE , vol.25, no.3, pp.43-46, July-Aug. 2006
- [7] Kasabov, N.; , "Bioinformatics: a knowledge engineering approach," Intelligent Systems, 2004.

- Proceedings. 2004 2nd International IEEE Conference , vol.1, no., pp. 19- 24 Vol.1, 22-24 June 2004
- [8] Fulekar, M.H. and J. Sharma. 2008. "Bioinformatics Applied in Bioremediation". Innovative Romanian Food Biotechnology. Vol. 2 No. 2. pp 28-36.
- [9] David Fenstermacher, Introduction to bioinformatics: Research Articles, Journal of the American Society for Information Science and Technology, v.56 n.5, p.440-446, March 2005
- [10] Achuthsankar S Nair, ; "Computational Biology & Bioinformatics – A gentle Overview", Communications of Computer Society of India, January 2007.
- [11] Doom, T.; Raymer, M.; Krane, D.; Garcia, O.; , "Crossing the interdisciplinary barrier: a baccalaureate computer science option in bioinformatics," Education, IEEE Transactions on , vol.46, no.3, pp. 387- 393, Aug. 2003
- [12] Jana, R., Aqel, M., Srivastava, P., and Mahanti, P. K., Soft Computing Methodologies in Bioinformatics, European Journal of Scientific Research, Vol. 26, No. 2, pp. 189-203, 2009.
- [13] Jacques Cohen, Computer science and bioinformatics, Communications of the ACM, v.48 n.3, p.72-78, March 2005
- [14] DOMOKOS, A.. BIOINFORMATICS AND COMPUTATIONAL BIOLOGY. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture, North America, 6527, p. 571 – 574, 09 2008.
- [15] Poe, D.; Venkatraman, N.; Hansen, C.; Singh, G.; , "Component-Based Approach for Educating Students in Bioinformatics," Education, IEEE Transactions on , vol.52, no.1, pp.1-9, Feb. 2009
- [16] Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002;324:1018–1022.
- [17] National Center for Biotechnology, "Bioinformatics Factsheet," <http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html>, last accessed June 13, 2012.
- [18] Gavin J. Gordon, Bioinformatics in Cancer and Cancer Therapy (Cancer Drug Discovery and Development) [Kindle Edition] , ISBN: 978-1-58829-753-2 e-ISBN: 978-1-59745-576-3, Library of Congress Control Number: 2008931368
- [19] Jacques Cohen, Computer science and bioinformatics, Communications of the ACM, v.48 n.3, p.72-78, March 2005
- [20] Umarji, M.; Seaman, C.; Koru, A.G.; Hongfang Liu; , "Software Engineering Education for Bioinformatics," Software Engineering Education and Training, 2009. CSEET '09. 22nd Conference on , vol., no., pp.216-223, 17-20 Feb. 2009
- [21] Simon R. Bioinformatics in cancer therapeutics-hype or hope?. Nat Clin Pract Oncol. 2005;2:223
- [22] Goldin, L.; , "Bioinformatics Integration for Cancer Research-Goal Question analysis," Information Technology: Research and Education, 2006. ITRE '06. International Conference on , vol., no., pp.248-252, 16-19 Oct. 2006
- [23] Kihara D, Yang YD, Hawkins T. Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. Cancer Inform. 2007;2:25-35.
- [24] Ardekani AM, Aslani F, Lakpour N. Application of genomics and proteomics technologies to early diagnosis of reproductive organ cancers. J Reprod Infertil. 2007;8(3):259-278.

***In Silico* Screening of the Library of
Pyrimidine Derivatives as Antitubercular agent**

Gopinathan. N, K. Chitra.

Faculty of Pharmacy, Sri Ramachandra University, porur, Chennai, Tamilnadu. pincode-600116

Email I.D. gopipharmacist@rediffmail.com

The paper is submitted to Biocomp2012

Abstract

In the present study, a novel series of 3,4 dihydro pyrimione derivatives were docked against the mycobacterium tuberculosis protein. Docking study was performed to rationalize the possible interactions between test compounds and active site of protein 1DQZ. The SAR study reveals the importance of presence of electronegative group for better activity. The selected residues for docking are LEU540, MET625, PHE650, ILE662, and ASP668. Library of the molecules was constructed based upon structural modifications of pyrimidine nucleus. Structural modifications were performed for the series of pyrimidines. Thus a library of pyrimidine derivatives was constructed based upon the feasibility of synthesis and *in silico* screened to prioritize the molecules and to obtain potential lead molecules as inhibitors. The three Dimension structure of the protein is retrieved from the PDB and its active sites are predicted from Qsite Finder. All the 48 structures of the ligand were drawn using Chemsketch12 and they are converted to PDB format. The docking was carried out using auto dock software 4. In these docking studies thio analogues are showing good binding energy. Fourteen compounds exhibiting good binding energy. The compounds can be synthesised in future and *in vivo* activities can be carried out. Further structural analysis of docking studies on our compound suggests attractive starting point to find new lead compounds with potential improvements.

Keywords: *InSilico* Screening, Docking, anti tubercular, Pyrimidines.

INTRODUCTION

Tuberculosis (TB) is a chronic infectious disease caused by mycobacteria of the “tuberculosis complex”, including primarily *Mycobacterium tuberculosis*, but also *Mycobacterium bovis* and *Mycobacterium africanum*. It was estimated that nearly 1 billion more people will be infected with TB in the next 20 years. About 15% of that group will exhibit symptoms of the disease, and about 3.6% (36 million) will die from TB if new disease prevention and treatment measures are not developed. The identification of novel target sites will also be needed to circumvent the problems associated with the increasing occurrence of multi-drug resistant strains. To do this, biochemical pathways specific to the mycobacteria and related organisms’ disease cycle must be better understood. Many unique metabolic processes occur during the biosynthesis of mycobacterial cell wall components. One of these attractive targets for the rational design of new antitubercular agents are the mycolic acids, the major components of the cell wall of *M. tuberculosis*[1]. Mycolic acids are high molecular weight C74eC90 α -alkyl, β -hydroxy fatty acids covalently linked to arabinogalactan. *In silico* screening methods such as docking have a great advantage as compared to 2D similarity and 3D pharmacophore search methods as it utilizes the 3D receptor structure a quantitative way..[4]

Docking is often used to predict the binding orientation of ligand to their protein targets in order to predict the affinity and activity of the small molecule. Hence docking plays a vital position in the rational design of drugs. *M. tuberculosis* a small aerobic non motile bacillus is the primary cause of tuberculosis. High lipid content of this pathogen accounts for many of its unique clinical characteristics. Each protein possesses a mycolyl transferase activity required for the biogenesis of trehalose dimycolate, a dominant structure necessary for maintaining cell wall integrity[3]. The docked complex of the designed compounds were found to display good binding affinity to the receptor. Molecular docking studies help to determine possible interaction of ligand with the enzyme[2].

METHODOLOGY

- All chemical structures were drawn using chem sketch software and all the files were converted to PDB file format.
- Protein target was downloaded from the PDB(Protein Data Bank). PDB ID 1DQZ
- Active site in protein target were determined from the online software qsite finder.
- Docking of protein with pyrimidine derivatives were carried using Autodock 4.0.
- The binding energy and the hydrogen bonds were observed as docking parameters.

RESULTS AND DISCUSSION

5-(1*H*-benzimidazol-2-yl)-4-(4-phenyl)-6-methyl-3,4-dihydropyrimidin-2(1*H*)-one has two hydrogen bond. The hydrogen 1- NH of 3,4 pyrimidine dione interact with the carbonyl oxygen of tryptophan TRP765 amino acid residue of protein to form hydrogen bond and its bond length is 2.176 Å. The carbonyl oxygen of 3,4 pyrimidine dione interact with the amine part of asparagine ASN 721 and its bond length is 2.083 Å. 5-(1*H*-benzimidazol-2-yl)-4-(2-chlorophenyl)-6-methyl-3,4-dihydropyrimidin-2(1*H*)-one forms two hydrogen bond. The

carbonyl oxygen of 3,4 pyrimidine dione interact with the amine of serine SER 624 forms hydrogen bond with length of 2.108 Å. The 3-N of benzimidazole of ligand interact with NH in indole nucleus of tryptophan TRP762 and its bond length is 2.017 Å. 5-(1*H*-benzimidazol-2-yl)-4-(4-chlorophenyl)-6-methyl-3,4-dihydropyrimidin-2(1*H*)-one hydrochloride forms one hydrogen bond. The 3-N of benzimidazole of ligand interact with 2 amino in side chain of tryptophan TRP762 and its bond length is 2.485 Å. 5-(1*H*-benzimidazol-2-yl)-4-(4-hydroxyphenyl)-6-methyl-3,4-dihydropyrimidin-2(1*H*)-one hydrochloride forms three hydrogen bond. The 4' OH of the ligand interact with carbonyl oxygen of glutamic acid GLU 655 and its bond length is 2.152 Å. The 4' OH of the ligand interact with NH of in indole nucleus of tryptophan TRP762 and its bond length is 2.061 Å. The 2nd position oxygen of 3,4 pyrimidine dione interact with the hydrogen of hydroxyl of threonine THR760 amino acid residue of protein to form hydrogen bond and its bond length is 1.747 Å. 5-(1*H*-benzimidazol-2-yl)-4-(2-hydroxyphenyl)-6-methyl-3,4-dihydropyrimidin-2(1*H*)-one has two hydrogen bond. The 2nd position oxygen of 3,4 pyrimidine dione interact with with NH in indole nucleus of tryptophan TRP762 to form hydrogen bond and its bond length is 1.747 Å. The 3-N of benzimidazole of ligand interact with NH of imidazole ring in histidine and its bond length is 2.24 Å. 5-(1*H*-benzimidazol-2-yl)-4-(furan-2-yl)-6-methyl-3,4-dihydropyrimidine-2(1*H*)-thione forms one hydrogen bond. The hydroxyl group of serine SER 624 interact with the 3-N of benzimidazole of ligand and its bond length is 1.923 Å.

Ethyl 4-(4-chlorophenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate forms two hydrogen bond. The free amino group of asparagine ASN 552 interact with carbonyl oxygen of the ester moiety of ligand and its bond length is 2.142 Å. The 2nd position oxygen of 3,4 pyrimidine dione interact with free amino group of leucine LEU 540 and its bond length is 2.009 Å. ethyl 4-(3-hydroxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate forms two hydrogen bond. The 2nd position oxygen of 3,4 pyrimidine dione interact with free amino group of leucine LEU 540 and its bond length is 1.023 Å. The carbonyl oxygen of ester moiety interacts with hydrogen of carboxylic acid group of glycine GLY 518. ethyl 4-(4-hydroxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate formed two hydrogen bond. The 2nd position oxygen of 3,4 pyrimidine dione interact with free amino group of leucine LEU 540 and its bond length is 2.138 Å and also interact with the hydroxyl group of serine SER 624 and its bond length is 1.325 Å. ethyl 4-(2-hydroxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate had three hydrogen bonds. The hydroxyl oxygen interact with the guanidine nitrogen of arginine and its bond length is 1.992 Å. The carbonyl oxygen of ester moiety interacts with hydrogen of hydroxyl group of serine SER 124 and its bond length is 1.968 Å and also interact with amino group of leucine LEU 40 and its bond length is 2.159 Å. ethyl 4-(4-aminophenyl)-6-methyl-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate forms two hydrogen bond. The 2nd position oxygen of 3,4 pyrimidine dione interact with NH of in indole nucleus of tryptophan TRP262 and its bond length is 2.25 Å. The 3rd position NH of 3,4 pyrimidine dione interact with carboxylic hydrogen of histidine HIS 250 and its bond length is 2.06 Å.

Ethyl 4-(4-hydroxyphenyl)-6-methyl-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate had two hydrogen bonds. The carbonyl oxygen of ester moiety interact with amino group of leucine LEU 40 and its bond length is 2.108. The oxygen of ester moiety interact with the hydrogen of hydroxyl group of serine SER 124 to form hydrogen bond with length of 2.048A. ethyl 4-(2-hydroxyphenyl)-6-methyl-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate had three hydrogen bond . The carbonyl oxygen of ester moiety interact with amino group of leucine LEU 41 and its bond length is 2.235A. The carbonyl oxygen of ester moiety interact with hydroxyl group of serine SER124 and its bond length is 1.047A. The 2' hydroxyl oxygen of ligand interact with the hydrogen of amino group of arginine ARG41 and its bond length is 2.004. ethyl 4-(4-hydroxy-3-methoxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate had three hydrogen bond. The hydroxyl oxygen of vaniline interact with hydroxyl hydrogen of serine SER 624 and its bond length is 2.223A. The 3-NH of 3,4 pyrimidine dione interact with oxygen of carboxylic acid of tryptophan TRP765 to form hydrogen bond with length of 1.583A. The 2nd position oxygen of 3,4 pyrimidine dione interact with free amino group of ASN 721 and forms hydrogen bond with length of 2.115A. The carbonyl oxygen of ester moiety interact with NH of in indole nucleus of tryptophan TRP762 with hydrogen bond length of 1.003A.

Conclusion:

The benzimidazole substitution at 5 the position increases the binding energy. unsubstituted 3,4 dihydro pyrimidones have good binding energy and more number of hydrogen bond. In pyrimidine derivative first position NH, 2nd position carbonyl, 3rd of benzimidazole is needed for the activity electro negativity substitution increases the activity. The predicted active sites from docking were TRP765, ASN721, SER624, GLU 655, HIS 760 etc..

Reference

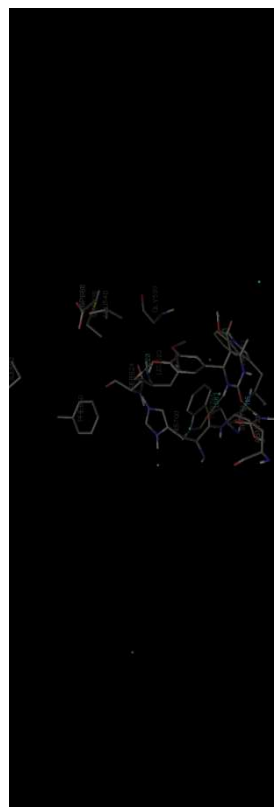
1. Suvarna G. Kini a,* , Anilchandra R. Bhat b, Byron Bryant c, John S. Williamson c, Franck E. Dayan d Synthesis, antitubercular activity and docking study of novel cyclicazole substituted diphenyl ether derivatives European Journal of Medicinal Chemistry xx (2008) 1e9.
2. T.S. Chitre*, K.G. Bothara, S.M. Patil, K.D. Asgaonkar, S. Nagappa and M.K. Kathiravan Design, Synthesis, Docking and Anti-mycobacterial activity of some novel thioracil derivatives as thymidine monophosphate kinase (TMPKmt) inhibitors Vol. 2 (2) Apr – Jun 2011 www.ijrpsonline.com pg.no 616.
3. Elucidating Drug-Enzyme Interactions and Their Structural Basis for Improving the Affinity and Potency of Isoniazid and Its Derivatives Based on Computer Modeling Approaches Auradee Punkvang 1, Patchreenart Saparpakorn 2, Supa Hannongbua 2, Peter Wolschann 3 and Pornpan Pungpo 1,* *Molecules* 2010, 15, 2791-2813; doi:10.3390/molecules15042791.
4. A. G. NERKAR*, S. A. GHONE and A. K. THAKER *In Silico* Screening of the Library of Pyrimidine Derivatives as Thymidylate Synthase Inhibitors for Anticancer Activity *E-Journal of Chemistry* <http://www.e-journals.net> 2009, 6(3), 665-672
5. Sylvie Pochet,[b] Laurence Dugue[□], [b] Gilles Labesse,[c] Muriel Delepierre,[d] and

Helene Munier-Lehmann*[a] Comparative Study of Purine and
Pyrimidine Nucleoside Analogues Acting on the Thymidylate Kinases of
Mycobacterium tuberculosis and of Humans DOI: 10.1002/cbic.200300608
ChemBioChem 2003, 4, 742 ± 747

Table of Binding Energies						
S.NO	Compounds	Binding Energy (kcal/mol)	Hydrogen bonds	Residues	Bond length (Å)	
1.	5-(1 <i>H</i> -benzimidazol-2-yl)-4-(4-phenyl)-6-methyl-3,4-dihydropyrimidin-2(1 <i>H</i>)-one	-9.06	2	ASN721 TRP765	2.083 2.176	
2.	5-(1 <i>H</i> -benzimidazol-2-yl)-4-(2-chlorophenyl)-6-methyl-3,4-dihydropyrimidin-2(1 <i>H</i>)-one	-9.29	2	SER624 TRP762	2.108 2.017	
3.	5-(1 <i>H</i> -benzimidazol-2-yl)-4-(4-chlorophenyl)-6-methyl-3,4-dihydropyrimidin-2(1 <i>H</i>)-one hydrochloride	-9.39	1	TRP762	2.485	
4.	5-(1 <i>H</i> -benzimidazol-2-yl)-4-(4-hydroxyphenyl)-6-methyl-3,4-dihydropyrimidin-2(1 <i>H</i>)-one	-9.05	3	TRP762 THR760 GLU655	2.061 1.747 2.152	
5.	5-(1 <i>H</i> -benzimidazol-2-yl)-4-(2-hydroxyphenyl)-6-methyl-3,4-dihydropyrimidin-2(1 <i>H</i>)-one	-11.05	2	TRP762 HIS760	1.042 2.24	
6.	5-(1 <i>H</i> -benzimidazol-2-yl)-4-(furan-2-yl)-6-methyl-3,4-dihydropyrimidine-2(1 <i>H</i>)-thione	-9.04	1	SER629	1.923	
7.	ethyl 4-(3-hydroxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-8.04	2	LEU540 GLY518	1.023 1.030	
8.	ethyl 4-(4-chlorophenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-8.84	2	LEU540 ASN552	2.009 2.142	
9.	ethyl 4-(4-hydroxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-8.19	2	LEU540 SER624	2.138 1.325	
10.	ethyl 4-(2-hydroxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-8.92	3	ARG41 SER124 LEU40	1.992 1.968 2.159	
11.	ethyl 4-(4-aminophenyl)-6-methyl-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-8.21	2	TRP262 HIS250	2.25 2.06	
12.	ethyl 4-(4-hydroxyphenyl)-6-methyl-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-8.42	2	LEU40 SER124	2.108 2.048	
13.	ethyl 4-(2-hydroxyphenyl)-6-methyl-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-7.57	3	LEU41 ARG41 SER124	2.004 2.235 1.047	
14.	ethyl 4-(4-hydroxy-3-methoxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate	-7.37	3	SER624 TRP765 ASN721	2.223 1.583 2.115	



Figure: 5-(1*H*-benzimidazol-2-yl)-4-(4-hydroxyphenyl)-6-methyl-3,4-dihydropyrimidin-2(1*H*)-one



Ethyl 4-(4-hydroxy-3-methoxyphenyl)-6-methyl-2-oxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate

Comparison of Statistical Tools for Microarray Data Analysis

O. Kaissi¹, A. Moussa¹, A. Ghacham¹, B. Vannier²

¹LTI Laboratory,ENSA , University Abdelmalek Essaadi , Tangier, Morocco

²IPBC,University of Poitiers,France

Contact :amoussa@uae.ac.ma

Abstract - *The present paper proposes a comparative study of two statistical tools integrated in R-Bioconductor Project, Expander, and Bioinformatics ToolBox of Mathworks, for gene selection in microarray data analysis. The main objective is to show the impact of results on selected genes when using statistical algorithms under different environments. This study compares results related to two data sets, the first one is the well known Latin Square Affymetrix data, and the second one is provided from a public data base.*

Keywords: Gene Selection, Statistical Algorithm, Soft Tools Comparison

1 Introduction

The technology of DNA microarrays currently experiencing an exceptional growth and has attracted tremendous interest in the scientific community. This interest lies in its efficiency; speed of obtaining results; and in its ability to study the expression of thousands of genes simultaneously [1].

The use of microarray in various fields including biology and health, allows development of several technologies grafting and in situ [2, 3]. Therefore several computational and statistical tools were developed to store, analyze and organize data [4].

A DNA chip consists of a DNA fragment immobilized on a solid support according to an ordered arrangement. The principle is based on the chip hybridization using a probe carrying the radioactive labeling [5]. Intensity of the signal generated is measured using a scanner. Image obtained, is analyzed to quantify the level of gene expression. Given the volume of data generated by this technology, several statistical methods based on the statistical t-test [6] were developed under some soft- tools for analyzing and selecting genes. But, the literature remains very poor in comparative studies showing the impact of the used algorithm and used materials in gene selection procedure. For this, the study proposed in this paper comes to show the performance of the statistical algorithm when using different soft tools.

This paper is organized as follows: an overview on Affymetrix technology and description of the three soft tools and statistical methods used in gene selection are given in section 2. In section 3, we present our comparative study of the data sets with some explanatory plots. We concluded this paper by discussing the results of this study.

2. Technologies and Tools

Affymetrix Gene Chip represents a very reliable and standardized technology for genome-wide gene expression screening [7]. In this technology; probe sets of 11–20 pairs with 25-mer oligonucleotides are used to detect a single transcript. Each oligonucleotide pair consists of a probe with perfect match to the target (PM probe) and another probe with a single base mismatch in the 13th position (MM probe) [8].

In the absolute analysis the goal is to answer the question: if the transcript of a particular gene is present or absent? The advantage to answer this question is that we can easily evaluate the expression and interpretation of results, by comparing the p-values expression levels off all genes to threshold α_1 and α_2 . Affymetrix technology offers two levels by default of α_1 and α_2 significances ($\alpha_1=0,04$ and $\alpha_2=0,06$). Genes with expression p-values under α_1 are called Present, genes with expression p-values higher then α_2 are called Absent, the genes with p values between α_1 and α_2 are called Marginal (Fig.1).

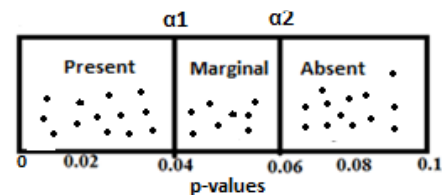


Fig.1: Significance levels in absolute analysis study

When the experiments concerned comparison of two conditions (treated # baseline) the objective of the comparative analysis is to answer the question: does the expression of a transcript on a chip (treated) change significantly with respect to the other chip (baseline)? In this context, five possible distinct answers are: Increase, Decrease, Marginal Decrease, Marginal Increase and No Change. These detections calls are giving by comparing change p-values of each gene the four thresholds chosen by the analysis for Affymetrix technology. Those thresholds are given in the Fig.2 [9].

Based on absolute and comparative analysis results, several methods have been developed to select the genes of interest. Many of these methods would be quite appropriate if genes would be analyzed one at a time. Some methods like T-test, ANOVA and F-test can easily be carried out for many genes simultaneously [10].

In the case of a lot of experiments, statistical test for selection is difficult to apply and multiple corrections need to be made. The most common multiple comparisons correction is the Bonferoni correction [11]: Rather than adjusting p-values for individual genes, he suggests to control the False-Discovery Rate (FDR) which is the fraction of false positives among the genes that are called, changed [12].

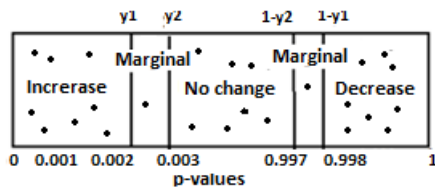


Fig.2: Significance levels in analyzer comparative study

In the comparison study of this work, we have chosen two well used methods for gene selection:

The SAM statistical algorithm [13]

The FDR controlling algorithm [11]

These algorithms, integrated in three software tools, are used as gene selection tools. Before presenting results, we recall in the two followed subsections the used data and software tools.

We used two data sets available on the public databases (NCBI and EBI) [14,15].

The first data set [16] includes 14 samples each of three replicated microarray oligonucleotides, in which multiple RNAs were added to the growing concentrations a common RNA preparation. Genes that should show variations in intensity are known (spikes genes), for this these data are generally used as references to validate development algorithms and software.

The second data we used provide from the article [17]. In this study we have to compare healthy and affected individuals, where this last have a dysfunction of lymphocytes. Different samples were taken for each dysfunction: 10 samples with Waldenstrom Macroglobulinemia (WM), 12 with Multiple Myeloma (MM), 11 with Chronic Lymphocytic Leukemia (CLL), with normal cases, 8 of B Lymphocytes (NBL), and 5 Plasma Cells (NPC). The differentially expressed genes explain relationship between the various syndromes or dysfunction [17].

Several software's has been developed to facilitate the analysis of microarray data. In this context, the most used free softwares is Bioconductor. However, Bioinformatics ToolBox of Mathworks and Expander offer a convivial interface to analyze data provided from microarray.

Standardization of the chips is applied on all chips and assumes that the distributions of intensities must be homogeneous. Several studies have focused on the performance of different normalization methods. In this study we use the Robust Multichip Analysis algorithm (RMA). This last provides accurate estimation of inter-array variability through a robust background correction and quantile normalization computed over the whole dataset [18]. The first used software is Bioconductor that is a collaborative project

using the statistical programming language R [19]. It allows statistical analysis on the use of different packages grouped under the name "biocLite". Bioconductor develops between other free applications especially designed for the analysis of biological data including microarray. For the analysis of Affymetrix chips with Bioconductor, we must first ensure that the Affymetrix libraries are installed [20]. The selection of differentially expressed genes realized by the "limma" package integrated in Bioconductor.

To assess the significance of genes, it is interesting to compare the value of 'fold change' which gives the direction of the stimulation of the gene, with the significance that quantified the importance of this direction. The volcano plot (Fig.3) arranges the genes along two axis that represent statistical significance and biological significance.

Bioinformatics ToolsBox of Mathworks offers biologists an open systems environment and stretch in which to explore ideas, prototype share new algorithms, and build applications for the analysis and simulation of biological systems [21]. It also offers interactive tools for designing and editing graphics (Fig.4).

Expander (Analyzer and Expression Displayer) is integrated software for the analysis of gene expression data. It was originally designed as a classification tool [22]. Today it has evolved to support all stages of data analysis chips, from the normalization of raw data to the inference of regulatory networks transcriptional [23].

3. Results and Discussions

We analyzed the performance of statistical tests integrated in Soft Tools cited below using Latin square and Leukemia data. Results are evaluated on the with the percentages of True Detection Rate (TDR=number of Spike detected / number of modulated genes reported). In leukemia data we consider the 69 genes cited in the work of [17] as spikes.

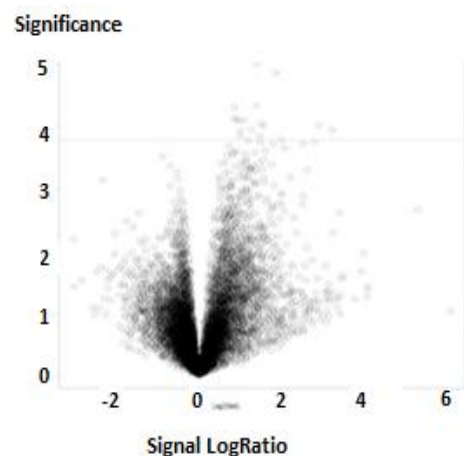


Fig.3: Volcano plot of leukemia data using Bioconductor

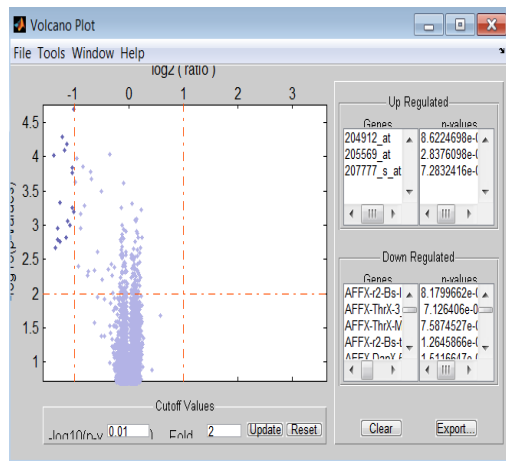


Fig.4: Volcano plot Latin Square data using Bioinformatics ToolBox of Matworks

For Both SAM and FDR controlling algorithm, we used two cutoff of pvalue for gene selection. Results are summarized in Figures 5 and 6 that represent the distributions of genes selected according to each software and each statistical algorithm.

Our comparative study allows us to define and determine that p-values 0.001 is more significant than p-values of 0.01 for both SAM and FDR, and the Expander allows to select a maximum of TDR and Spike. In addition we show that this analysis confirms that selected genes depend both on the used algorithm and the used Soft Tools. This analysis gives some list of new interest genes.

Finally, we remind that this work focus the problem of used algorithm and tools in gene selection problem. In this context we have used two p-values with screening tests: FDR and SAM. To highlight the difference between these two selection methods we tested their effectiveness on three environmental developments chips Bioconductor, Bioinformatics tool box and Expander, using Latin square data and leukemia public data. We conclude that in microarray data analysis, the best way is to work with different approaches for statistical analysis at the same time for a better validation of results.

References

- [1]: V. Gomases, S. Tagore and K.V. Kale, 'Microarray: an approach for current Drug targets' *Current Drug Metabolism*, vol. 9, pp. 221-31, 2008.
- [2]: Y. F. Leung and D. Cavalieri, 'Fundamentals of cDNA microarray data Analysis', *Trends Genet*, vol.19, pp. 649-659,2003
- [3]:.D. J Lockhart, H. Dong, M. C. Byrne, and al. 'Expression monitoring by hybridization to high-density oligonucleotid arrays' *Nat. Biotechnol*, vol.14,pp.1675-1680, 1996.

- [4]: D. J. Duggan, M. Bittner, Y. Chen,P. Meltzer and J. M. Trent, 'Expression profiling using cDNAMicroarrays'. *Nat.Genet*, vol.21,pp.10-14, 1999.
- [5]: V.Frouin, and X.Gidrol,'Analyse des données d'expression issues des puces à ADN' *Biofutur*, vol.252, pp. 22 – 26,2005.
- [6]:F. Chu and L. Wang, 'Applications of support vector machines to cancer classification with microarray data', *International Journal of Neural Systems*, vol. 15, pp475-484,2005
- [7]: D. Lockhart, H. Dong, M. Byrne,M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton and al 'High density synthetic oligonucleotide arrays'. *Nat.Biotechnol*, vol. 14, pp. 1675-1680, 1996.
- 8]: D. Choaglin, F.Mosteller, J.W. Tukey, 'Understanding Robust and Exploratory Data Analysis' *Wiley*, vol.79,pp.7-32, 2000.
- [9]: W. mLiu, R. Mei, X. Di, T. Ryder, E. Hubbel, S. Dee, T. Webster, C.Harrington, M. Ho, J. Baid and S. Smeekens 'Analysis of High Density Expression Microarray with Signed-Rank Calls Algorithmes'. *Bioinformatics*, vol.12, pp.1593-1599, 2002.
- [10]: D. Faller, H. U. Voss, J. Timmer and U. Hobohm. 'Normalization of DNA-microarray data by nonlinear correlation maximization' *J. Comput. Biol.*, vol.10, pp. 751-762. 2003.
- [11]: S.Dudoit, Y.H. Yang, M.J. Callow and T.P. Speed,'Statistical methods for identifying differentially expressed genes in replicated DNA microarray experiments'. *Statist.Sinica*, vol.12,pp. 111–139, 2002.
- [12]: Y. Benjamani, Y. Hochberg. 'Controlling the false discovery rate :a practical and powerful approach to multiple testing'. *Journal of the Roy.Soc.* vol. 57,pp. 289-300, 1995.
- [13] : V.G. Tusher, R. Tibshirani, G. Chu, 'Significance analysis of microarrays applied to the ionizing radiation response', *Proc. Nat. Acad. Sci. USA*, 2001, Vol. 98, pp. 5116-5121.
- [14]: www.ncbi.nlm.nih.gov/
- [15]: www.ncbi.nlm.nih.gov/
- [16]:www.affymetrix.com
- [17]: NC. Gutiérrez and All 'Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglobulinemia: comparison with expression patterns of the same cell counterpartsfrom chronic lymphocytic leukemia, multiple myeloma and normal individuals', *Leukemia*, vol. 21(3), pp. 541-550, 2007.
- [18]: B. Bolstad, R Irizarry., MAstrand., and T .Speed, A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics*,vol.19, pp.185-193,2003.
- [19]: www.r-project.org
- [20]: G. K. Smyth 'Linear Models and Empirical Bayes Methods for Assessing Differential Expression in MicroarrayExperiments. Statistical Applications' in *Genetics and Molecular Biology*, vol.3 pp. 2004.
- [21]:www.mathworks.com/products/bioinfo

[22]: R. Sharan, A. Maron-Katz, and R Shamir, 'CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics* vol.19,pp.,1787–1799, 2003.

[23]: R. Shamir and al., 'EXPANDER: an integrative program suite for microarray data analysis. *BMC Bioinformatics* vol 6 pp, 232, 2005.

Table I: Results of Latin Square Dataset

Pvalues	0,01				0,001			
	T-Test (SAM)		FDR		T-Test (SAM)		FDR	
	TDR	Spike	TDR	Spike	TDR	Spike	TDR	Spike
Bioconductor	40,22 %	83,33 %	53,33 %	76,19 %	54,76 %	54,76 %	55,55 %	35,71 %
Bioinformatics Tools Mathworks	35,95 %	76,19 %	42,64 %	69,04 %	49,75 %	50% %	52,94 %	21,42 %
Expander	57,07 %	95,23 %	60,34 %	83,33 %	64,1% %	59,52 %	65,62 %	50% %

Fig.5: Number of genes selected and grouped according to the used statistical tool.

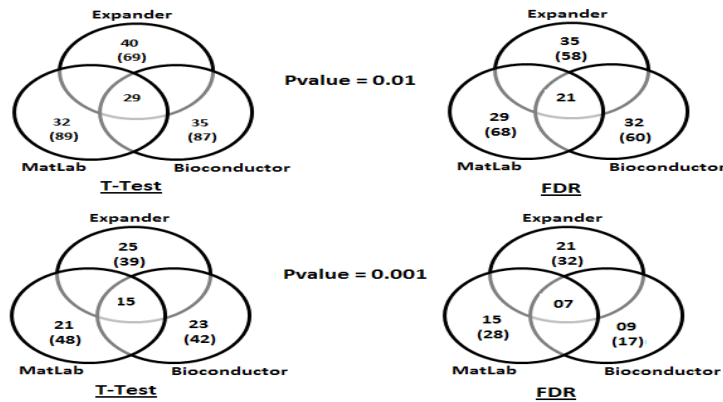


Table II: Results of Leukemia Dataset

Pvalues	0,01				0,001			
	T-Test		FDR		T-Test		FDR	
	TDR	Spike	TDR	Spike	TDR	Spike	TDR	Spike
Bioconductor	36,5%	86,95%	51,85%	79,71%	93,84%	56,52%	55,6%	46,37%
Bioinformatics Tools Mathworks	33,82%	79,71%	45,39%	75,36%	71,13%	50,72%	98,57%	42,02%
Expander	55,64%	89,85%	70,4%	58,5%	66,45%	65,21%	65,34%	56,52%

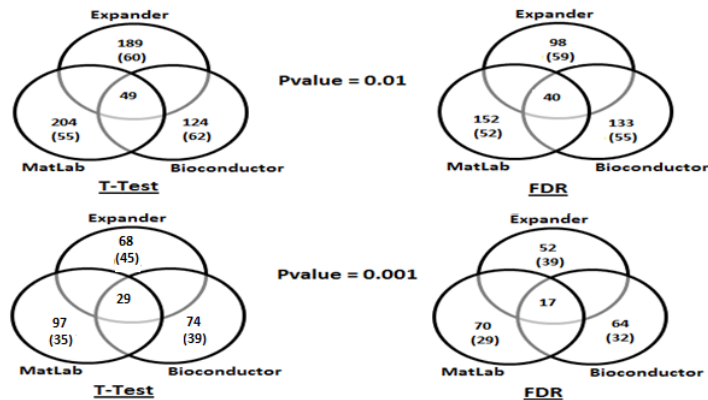


Fig.6: Number of genes selected and grouped according to the used statistical tool.

Assessment of variability of the mouse myocardial fiber structure via principal component analysis

S. Merchant¹, Y. Jiang², S. Joshi³ and E. W. Hsu¹

¹Department of Bioengineering, University of Utah, Salt Lake City, UT, United States, ²Center for In Vivo Microscopy, Duke University, Durham, NC, United States, ³School of Computing, University of Utah, Salt Lake City, UT, United States

Abstract - *In this study, we applied principal component analysis (PCA), a well-established data reduction technique, to helix angle maps generated from high-resolution 3D DTI datasets acquired on a group of 11 normal and 2 hypertrophied mouse heart specimens. Results of the study show that ventricular myocardial fiber orientations among the hearts have three main modes of variation which include variation in constant offsets, transmural slope and the longitudinal slope of the associated helix angles. Moreover leave-one-out experiments indicate that in general a subset of as few as 10 hearts is adequate in capturing the variability in the most prominent patterns of myocardial fiber structure as well distinguishing between normal and diseased hearts. These results strengthen the fact that structural variability is spatially dependent and suggest that whole-heart based approach can be more advantageous over voxel-based statistics for analyzing and comparing the myocardial fiber structure.*

Keywords: diffusion tensor imaging; myocardial fiber structure; mouse ; principal component analysis

1 Introduction

The structure of the heart is highly organized and is a key determinant in its electrophysiological and mechanical properties. A well-known hallmark of the left ventricular myocardium is that its fiber orientation undergoes a counter-clockwise rotation from the epicardium to the endocardium [1]. Magnetic resonance diffusion tensor imaging (MR-DTI, or DTI for short) [2] has emerged as a promising alternative to conventional histology for characterizing myocardial structures [3]. The non-destructive, inherently 3D and high-resolution nature of DTI allows the myocardium to be examined at unprecedented level of detail.

To date, DTI has been used to characterize structures of normal hearts across a variety of species, ranging from mouse [4] to humans [5]. DTI has also been utilized to characterize heart diseases, including infarct [6-9], heart failure [10,11] and hypertrophic cardiomyopathy [12]. The data analysis employed in most studies has relied on ROI-based or voxel-by-voxel statistics. The main limitation in this approach is that it requires standardization of the coordinates among the hearts, which is performed mostly by visual inspection. Additionally, the statistics often assumes that different voxels or regions of the heart behave independently, which is not

necessarily true given the highly organized structure-function relationship of the organ.

Recently, advances in computational anatomy have made possible atlas-based approaches to analyze biomedical images, including cardiac DTI datasets [13]. Besides more objective coordinate standardization for group analysis, atlas-based analyses allowed the computation of the group average (or atlas) of heart structures. However, a mathematical average is not necessarily the representative average for the group. For the latter, the average needs to be accompanied by some estimate of the variability. Given the conspicuous patterns that exist, an analysis of variability can likely identify ways in which myocardial structures vary and facilitate the comparisons of both normal and diseased hearts.

The aims of the current study are two-fold. First, principal component analysis (PCA), a well-established data reduction technique, is used to determine whether they exist and to identify the main modes of fiber structural variability among a group of similar hearts. Second, in a preliminary demonstration, the PCA variability analysis is applied to diseased hearts to determine whether the approach can be used for detecting structural remodeling.

2 Methods

Hearts were isolated from normal 8-month-old male 129/ola mice (n = 11) and additional animals with cardiac hypertrophy (n = 1 for each mild and severe case) induced via aortic banding. DTI datasets were obtained as described previously [3] on a 9.4 T MRI instrument. Each DTI dataset consisted of a fully encoded 128 x 128 x 128 (readout x phase x slice) matrix-size b₀ (i.e., b ~ 0) and 12 reduced-encoded (128 x 64 x 64) spin-echo diffusion-weighted (b = 1130 s/mm²) images sensitized in each of an optimized set of 12 directions [14]. The diffusion-weighted scans were then reconstructed using reduced encoding imaging via generalized series reconstruction (RIGR) [15] to full matrix size. Diffusion tensors were computed on a voxel-by-voxel basis via nonlinear least squares fitting and diagonalized.

All post processing was conducted via custom codes written in Matlab (Mathworks, R2011b). The eigenvector of the largest diffusion tensor eigenvalue was taken as the local myocardial fiber orientation, which was then projected onto a cylinder coaxial with the cardiac long axis to obtain its helix angle [3]. The 3D helix angle maps of the normal hearts were

transformed onto a common template using unbiased large deformation diffeomorphic metric mapping (LDDMM) registration [16-18]. To reduce the data dimensionality for subsequent analysis, each 3D volume was downsampled by a factor of 4, cropped to span the cylindrical portion of each left ventricle (~75% of the chamber), vectorized and inserted row-wise into the data matrix.

PCA was then performed on the data matrix to yield the PCA coefficients and the data scores, which respectively represent the transformation coefficients that define the principal components forming the PCA space and the coordinates of the location of the original data points in this space. To determine which of the PCA components or modes of variation were significant, Parallel Analysis [17] was performed. Briefly, the correlation coefficient matrix (CCM) of the data matrix was calculated, diagonalized, and its eigenvalues compared to those similarly obtained from CCM of a normally distributed random data matrix (with the same size as that of the combined matrix) and averaged 100 times. The eigenvalues of the data matrix that surpassed their noise-derived counterparts were deemed as the significant modes of variation.

To visualize the modes of variation, the normal heart datasets were projected onto each of the significant principal component. Subsequently, the projections were reshaped back onto 3D space and averaged. Transmural profiles of the mean and standard deviation, for each principal component, were taken and plotted at the left ventricular hemispherical slice and center of the free wall between the papillary muscles at the center of the free wall between the papillary muscles on 4 equally spaced short-axis slices.

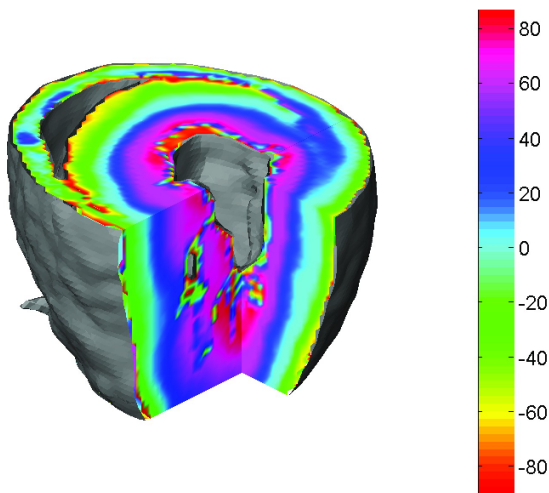


Figure 1: Helix angle of the primary eigenvector of 3D LV volume represented in the local cylindrical coordinate system and shown in false-color (in degrees). The region shown was used for analysis.

Lastly, to evaluate the extent of the variability captured by PCA and demonstrate whether the analysis is suitable for detecting, for example, remodeled (i.e., non-normal) cardiac

structure, leave-one-out trials [18] were conducted by excluding one heart at a time and forming combinatorial 10-heart subsets ($n = 11$) of the original normal heart group. PCA was performed on each subset and the normal heart excluded and the hypertrophy hearts were projected onto the principal component space spanned by the subset of normal hearts. Based on the Gaussian statistics where 95% of the distribution is contained within 2 standard deviations from the mean, hearts that lie outside the ellipsoid spanned by 2 standard deviations along each significant principal component axis were treated as being significantly different from the subset.

3 Results

Figure 1 shows a cut-open view of the 3D myocardial fiber helix angle map of a representative normal heart, demonstrating the distinctive counter-clockwise helical pattern of the fiber structure [3]. The results from the Parallel analysis is shown in Fig. 2, which indicates that only first three principal components or modes of variation exceed the average noise level and are thus deemed significant. Figure 2 also indicates that, together, first 3 principal components capture approximately 50% of the total variation observed in the datasets.

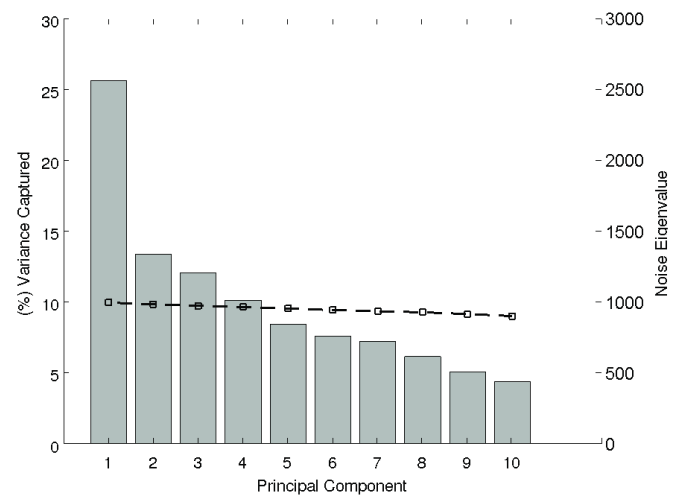


Figure 2: Bar graphs of percent variance captured by each principal component of 11 normal hearts. The dashed line represents the average eigenvalues calculated from random (noise) data. Only the first 3 principal components of the heart data are higher than noise average.

Figure 3 shows transmural profiles obtained from the LV lateral wall that provide a visual representation of the mean and one standard deviation variability along the 3 significant principal components. The profiles at different short-axis slices suggest that the first and second principal components capture mostly the constant offsets and slopes of in the helix angles, respectively. On the other hand, the third component captures largely the longitudinal variation in helix angle transmural slope. All principal components capture the longitudinal (base-to-apex) variation in the transmural slope of the helix angle.

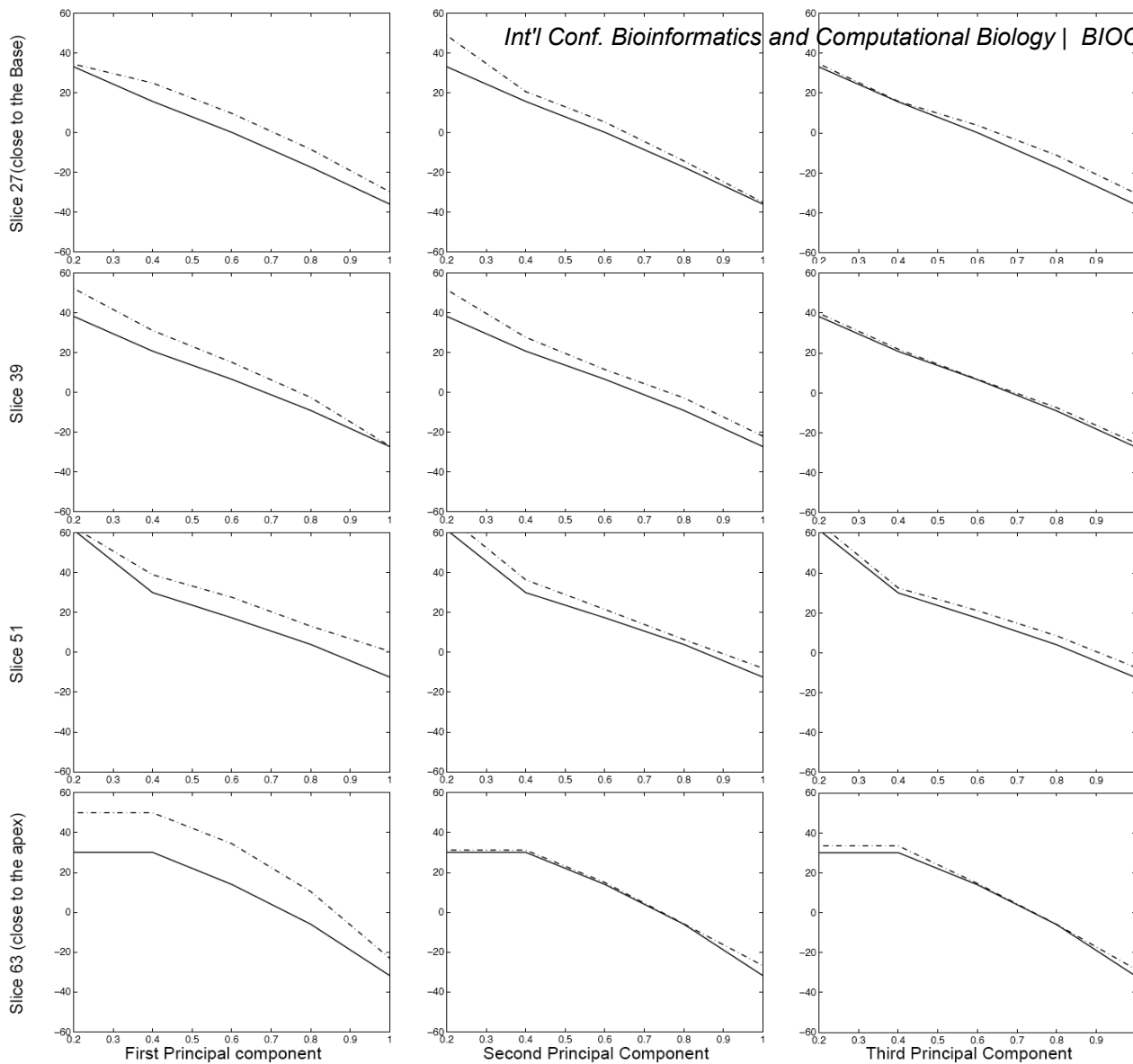


Figure 3: Profiles from the mean of the projected data using the first three significant principal components. Left column: first principal component, middle column: second principal component, and right column: third principal component. Profiles were selected from the lateral wall in 4 slices evenly spaced from base to apex (shown in rows). Solid lines represent the mean whereas the dashed dot lines represent one positive standard deviation around the mean. For each image the x-axis represents the transmural distance (in mm) and the y-axis represents the helix angle (in degrees). From base to apex, the first and second components capture the offset variation while the third component captures the slope of the helix angle (from epi- to endocardium). All component capture some longitudinal variation.

Figure 4 shows the PCA space plotted for a representative leave-one-out trial. Among the trials tested, the excluded heart was projected inside the 2-standard deviation ellipsoid in 10 out of the 11 cases. One normal heart was consistently projected outside of the ellipsoid, suggesting that the heart may be an outlier of the normal group (due to several possible causes such as mislabeling or inconsistency in specimen preparation, etc). Aside from the outlier, the inclusiveness of the left-out heart in all trials suggests that each of the 10-heart PCA space sufficiently span the variability space of the fiber structure. For the case shown in Fig. 4, the two hypertrophic hearts lie outside the 2-standard deviation ellipsoid, suggesting that their myocardial fiber structure as represented by the helix angle, in terms its offset and slope, has higher

variability than the normal hearts. Among all the trials tested, the severely hypertrophic heart lies outside the ellipsoid in 10 out of the 11 tests, whereas the same is true for the mildly hypertrophy heart in only 2 out of 11 cases. Although it is unclear whether fiber structural remodeling has taken place in the mildly hypertrophic heart, the successful exclusion of the heart with severe hypertrophy is promising for the PCA approach as a means to detect cardiac structural remodeling.

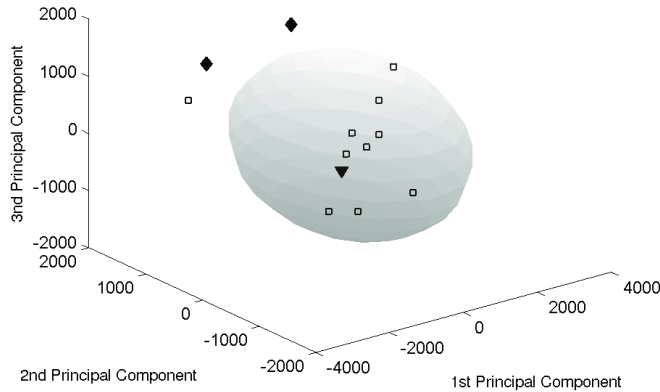


Figure 4: Representative PCA space plot of a leave-one-out test. The axes of the PCA comprise of the three principal components. The ellipsoid represents the 95% confidence interval with the randomly 10 normal hearts datasets (open squares), the excluded normal heart (filled triangle), and the two hypertrophic hearts (filled diamonds).

4 Discussion

Results of the current study, combined, indicate that the left ventricular myocardial fiber orientations among the hearts examined varied in only specific manners, including the constant offsets, transmural slope and the longitudinal slope change of the associated helix angles. Although the finding is not surprising given the known and distinctive organization of the left ventricular fiber structure, to the authors' knowledge, the study is the first time these spatial patterns of variability among similar hearts have been identified and quantified. Because the structural variability is spatially dependent, an immediate impact is that myocardial structures across hearts cannot be simply compared on a voxel-by-voxel or ROI-by-ROI basis.

The leave-one-out experiment, although preliminary, is also a first of its kind in evaluating the robustness of the PCA variability space, and to apply the identified patterns of variability to detect structural remodeling in diseased hearts. Despite that the "accidental outlier" in the normal group and the remodeling of the mildly hypertrophic heart need to be further verified, the results in general suggest that a subset of as few as 10 hearts is adequate in capturing the variability in the most prominent patterns of myocardial fiber structure. The number required is an order of magnitude smaller than the 100s of datasets typically involved in the construction of common structural atlases of the brain.

One potential limitation of the current study is, due to logistical constraints, its small dataset size. Ideally a larger size is preferable, which would, for example, allow independent subsets of the hearts to be generated to investigate the robustness of the PCA variability. Although a larger dataset size can possibly identify more significant minor PCA components and increase the detection sensitivity of

diseased hearts, it is unlikely to change the few major patterns of structural variability. Another limitation is that the current study focuses on a specific species and strain, and the results may or may not be extended to hearts of other species or other strains of the same species. However, the methodology taken by the study is expected to be applicable for the latter studies.

In summary, PCA was performed on mouse cardiac DTI datasets to investigate the intra-species variability of myocardial fiber structure. Results indicate that the fiber structures vary among only specific spatial patterns, and that the variability can be captured by relatively few datasets. Taken together, these findings are supportive of the construction of representative atlases or parametric analytical models of the myocardial structure.

5 Conclusion

In summary, PCA was performed on mouse cardiac DTI datasets to investigate the intra-species variability of myocardial fiber structure. Results indicate that the fiber structures vary among only specific spatial patterns, and that the variability can be captured by relatively few datasets. Taken together, these findings are supportive of the construction of representative atlases or parametric analytical models of the myocardial structure.

6 References

- [1] Streeter DDJ, Spotnitz HM, Patel DP, Ross J, Sonnenblick E. *Circ Res.* 1969;24:339-347
- [2] Bassar PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal.* 1994; 66:259-67.
- [3] Jiang Y, Pandya K, Smithies O, Hsu EW. Three-dimensional diffusion tensor microscopy of fixed mouse hearts. *Magnetic resonance in medicine.* 2004; 52:453-60.
- [4] Quantitative comparison of myocardial fiber structure between mice, rabbit, and sheep using diffusion tensor cardiovascular magnetic resonance. Healy LJ, Jiang Y, Hsu EW. *J Cardiovasc Magn Reson.* 2011 Nov 25;13:74.
- [5] Eggen MD, Swingen CM, Iaizzo PA. Analysis of fiber orientation in normal and failing human hearts using diffusion tensor MRI. *IEEE.* 2009;:642-645.
- [6] Chen J, Song S-K, Liu W, McLean M, Allen JS, Tan J, Wickline S a, Yu X. Remodeling of cardiac fiber structure after infarction in rats quantified with diffusion tensor MRI. *American journal of physiology: Heart and circulatory physiology.* 2003; 285:H946-54.
- [7] Wu EX, Wu Y, Nicholls JM, Wang J, Liao S, Zhu S, Lau C-P, Tse H-F. MR diffusion tensor imaging study of postinfarct myocardium structural remodeling in a porcine model. *Magnetic Resonance in Medicine.* 2007; 58:687-95.

- [8] Wu M-ting, Tseng W-YI, Su M-yuan M, Liu C-peng, Chiou K-R, Wedeen VJ, Reese TG, Yang C-F. Diffusion tensor magnetic resonance imaging mapping the fiber architecture remodeling in human myocardium after infarction: correlation with viability and wall motion. *Circulation*. 2006; 114:1036-45.
- [9] Wu M-T, Su M-YM, Huang Y-L, Chiou K-R, Yang P, Pan H-B, Reese TG, Wedeen VJ, Tseng W-YI. Sequential changes of myocardial microstructure in patients postmyocardial infarction by diffusion-tensor cardiac MR: correlation with left ventricular structure and function. *Circulation. Cardiovascular Imaging*. 2009; 2:32-40.
- [10] Helm P a, Younes L, Beg MF, Ennis DB, Leclercq C, Faris OP, McVeigh E, Kass D, Miller MI, Winslow RL. Evidence of structural remodeling in the dyssynchronous failing heart. *Circulation Research*. 2006; 98:125-32.
- [11] Eggen MD, Swingen CM, Iaizzo PA. Analysis of fiber orientation in normal and failing human hearts using diffusion tensor MRI. *IEEE*. 2009;:642-645.
- [12] Tseng W-YI, Dou J, Reese TG, Wedeen VJ. Imaging myocardial fiber disarray and intramural strain hypokinesis in hypertrophic cardiomyopathy with MRI. *Journal of magnetic resonance imaging*. 2006; 23:1-8.
- [13] Peyrat, J.M. et al., "Towards a statistical atlas of cardiac fiber structure: A computational framework for the statistical analysis of cardiac diffusion tensors: application to a small database of canine hearts," *IEEE Trans Med Imaging* 26(11), 1500-14 (2007)
- [14] Papadakis NG, Xing D, Huang CL, Hall LD, Carpenter TA. A comparative study of acquisition schemes for diffusion tensor imaging using MRI. *J Magn Resonance* 99;137:67– 82
- [15] Liang ZP, Lauterbur PC. An efficient method for dynamic magnetic resonance imaging. *IEEE Trans Med Imag* 1994;13:677– 686.
- [16] Lorenzen, P.J., Davis, B.C., Joshi, S.: Unbiased atlas formation via large deformations metric mapping. In: Duncan, J.S., Gerig, G. (eds.) *MICCAI 2005*. LNCS, vol. 3750, pp. 411–418. Springer, Heidelberg (2005)
- [17] Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage* 23 (2004)
- [18] C. Yan, M. I. Miller, R. L. Winslow, and L. Younes, "Large deformation diffeomorphic metric mapping of vector fields," *tmi*, vol. 24, no. 9, pp. 1216–1230, 2005.
- [19] Franklin, Scott B.; Gibson, David J.; Robertson, Philip A.; Pohlmann, John T.; and Fralish, James S., "Parallel Analysis: a Method for Determining Significant Principal Components" (1995). Publications. Paper 9.
- [20] Vik T, Heitz F, Armspach JP. Lecture Notes in Computer Science 2003;2879:838- 845.

Deciphering Splicing Codes of Spliceosomal Introns

Degen Zhuo¹, Wenhong Cao², Shoukang Zhu³, Chunming Dong³, and A.D. M. Glass⁴

¹SplicingCodes.com, BioTailor Inc, 8800 SW153 Terrace, Miami, FL, 33157, USA;

²2212 McGavran Greenberg Hall, UNC, 135 Dauer Drive, Chapel Hill NC 27599USA;

³812 BRB, 1501 NW 10th Ave Miller School of Medicine, University of Miami, Miami, FL, 33136;

⁴Dept of Botany, UBC, 6270 University Blvd, Vancouver, BC, Canada, V6T 1Z4;

Abstract *Splicing codes, defined as cis-regulatory sequences conferring pre-mRNA splicing specificities, have been poorly understood. Here we reported that in humans, as well as other vertebrates, there are 2.5-5 fold more cases of 6 nt identical length between 5' exonic (E5) and 3' intronic (I3) sequences than between 5' intronic and 3' exonic ones. This disparity in conservation extends well beyond 5' exonic CAG and raises the provocative possibility that 5' exonic sequences are similar to the intronic-binding sites (IBS) of group II ribozymes. Based on this finding, a web-base software system has been developed to predict alternative splicing. The mouse insulin receptor gene is predicted to encode complex splice sites, 58.3% of which have been verified. This work not only provides accurate computation systems to predict alternative splicing, but also has laid foundations to develop simple, accurate and personalized diagnosis and therapy methods for complex diseases.*

Keywords: splicing code, spliceosomal introns, splice junctions, cis-elements, splicing specificity, spliceovariants

1. Introduction

Most genes of eukaryotic organisms have spliceosomal introns, whose lengths and sequences are highly variable and range from 20 bp to 800 kb[1]. 5' splice sites of introns have the conserved motif of an exonic AG followed by an intronic GTRAGT (R: purine). 3' splice sites are composed of the branch point (YNYURAC, Y: pyrimidine), the polypyrimidine tract and the conserved splice site YAG[1]. The majority of mammalian spliceosomal introns undergo extensive alternative splicing[2, 3], which has been suggested to be responsible for the "missing" protein-coding genes and proteomic diversity in mammals. Aberrations in pre-mRNA splicing have played an essential role in almost every known disease with genetic aetiology, disease susceptibility and severity[4] and in development, differentiation, aging and cancer[5].

Spliceosomal introns are removed from nuclear pre-mRNAs via two consecutive *trans*-esterification reactions before mature mRNAs are exported into the cytoplasm for translation into proteins. Intron removal from pre-mRNAs is mediated by spliceosomes which are known to be comprised of several hundred proteins and five small U snRNAs packaged as ribonucleoprotein particles (RNPs)[6]. These U

snRNAs have highly conserved secondary structures among eukaryotic organisms and are similar to domains of group II introns[7, 8], which are believed to be ancestors of spliceosomal introns[1].

Many methods have been developed to predict alternative pre-mRNA splicing with limited success. Traditionally, alternative spliceovariants were identified by aligning different cDNAs/ESTs to the different regions of the same genomic sequences. Next-generation sequencing (NGS) provides more tools to identify novel splice variants. Using paired-end RNA sequencing and RNA-seq, surprisingly >23,000 introns have been identified in *D. melanogaster*[9]. More recently, Pickrell et al. has used RNA-seq technology to sequence cDNA libraries constructed from the mRNAs of human cell lines and have identified approximately 150,000 previously unannotated splice sites out of 306,606 splice junctions[10]. While 50% of all observed junctions are not present in gene models, these account for only 1.7% of all junction-spanning sequencing reads[10].

Splicing codes have been proposed to explain the conundrum of the diversity and specificity of pre-mRNA splicing, *trans*-splicing and alternative splicing[11]. Exonic and intronic splicing enhancers and silencers have been suggested to be potential candidates of splicing codes[11]. Recently, Barash et al. have assembled "the splicing code" representing several hundreds of RNA features and predicted mouse tissue-dependent changes in alternative splicing for thousands of exons[12].

However, the approaches described fail to explain the universality (*cis*- vs *trans*-splicing) and diversities (fungi vs mammals) as does specificity of pre-mRNA splicing and splice site choices in alternative splicing [13]. In an attempt to identify and characterize the sequences that confer pre-mRNA specificity, here we have analyzed spliceosomal intron datasets from invertebrates and vertebrates, proposed that 5' exonic and 3' intronic sequences constitute splicing codes of spliceosomal introns and discussed our findings.

2. Results

2.1 Asymmetric conservation of 5' exonic (E5) and 3' intronic (I3) sequences.

We previously had shown that recently-acquired human spliceosomal introns had signatures of similar 5' and 3'

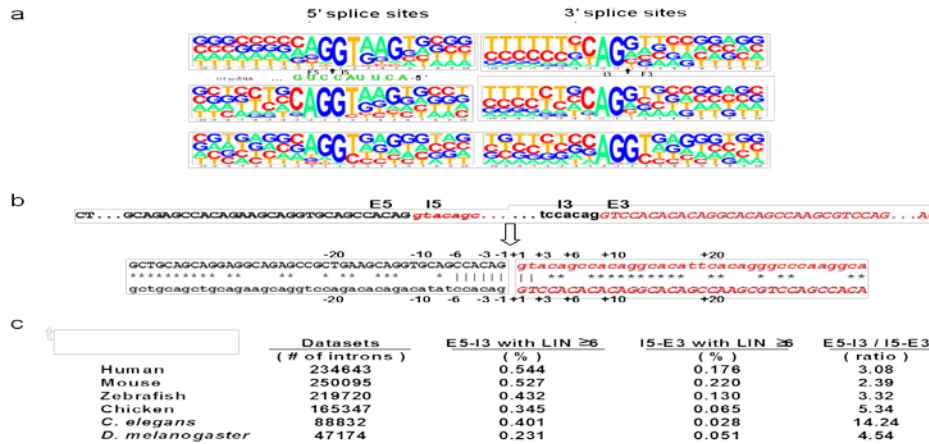


Fig.1 Splice junction features. a) Consensus sequences of 5' and 3' splice sites from the total human intron dataset (top panel), for E5-I3 with $LIN \geq 6$ (middle panel) and I5-E3 with $LIN \geq 6$ (bottom panel). The graphics are generated by Pictogram (<http://genes.mit.edu/pictogram.html>). The splice junctions are shown by arrows. The blue sequence below E5-I5 shows the 5' splice site recognition motif of U1 snRNA. b) Example of E5-I3 and I5-E3 alignments for intron 8 (168 bp) of the human *ciz1* gene. The black and gray italic uppercase letters represent the 5' and 3' exonic sequences at splice sites, respectively, and the gray italic and black lowercase letters indicate the 5' and 3' intronic sequences. The vertical lines indicate uninterrupted identical nucleotides extending from the splice junctions for the E5-I3 and I5-E3 alignments, and are designated as LIN (length of identical nucleotides). Asterisks represent identical nucleotides outside of this region. c) Sizes of animal intron datasets, and proportion with $LIN \geq 6$, also expressed as the ratio between E5-I3 and I5-E3. All observed differences between E5-I3 and I5-E3 are statistically significant ($p < 0.001$).

splice sites[14]. The signatures of such introns were 5' and 3' intron boundaries that were very similar to each other (Fig. 1a, middle and bottom panels) and that did not conform well to the typical splice site consensus sequences (Fig.1a, top panel). To assess if the degree of similarity was uniform along the splice junctions, we divided each splice junction into its exonic and intronic portions (designated as E5 and I5 for the 5' splice site and I3 and E3 for the 3' splice site) and starting from the splice junction, we scored the length of identical nucleotides (LIN) in an uninterrupted stretch independently for the E5-I3 and I5-E3 alignments. Figure 1b gave a specific example showing the sequences flanking intron 8 of the human *ciz1* gene which encoded *Cip1*-interacting zinc finger protein 1[15]. This was done for human introns as well as those for other vertebrates (mouse, zebrafish and chicken) and the invertebrates *C. elegans* and *D. melanogaster* (Fig.1c, Figure 2). Notably, as shown in Figure 1c, the percentage of E5-I3 alignments with $LIN \geq 6$ is significantly higher ($p < 0.001$) than for I5-E3 in humans (by 3-fold), in other vertebrates (by 2.4 to 5.3 fold)

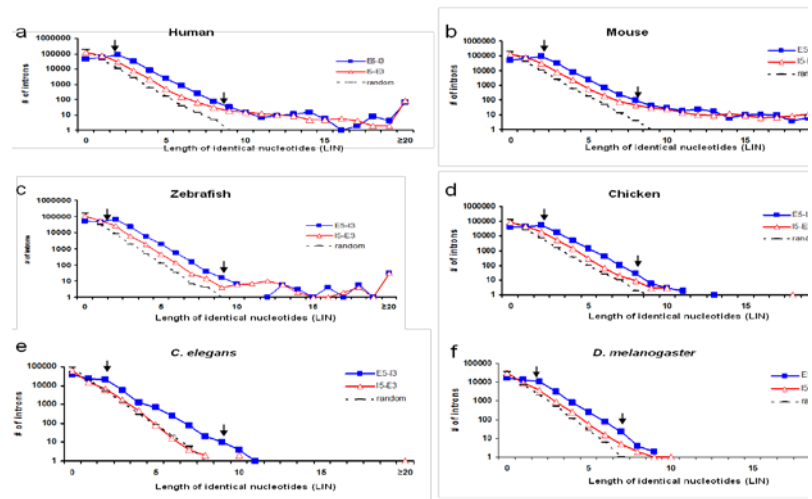


Fig.2. Comparison of LIN (length of identical nucleotides) distributions for E5-I3 and I5-E3 alignments from various animals. a) human, b) mouse, c) zebrafish, d) chicken, e). *C. elegans* and f) *D. melanogaster*. The solid black squares and gray triangles represent E5-I3 and I5-E3 alignments, respectively. The dashed lines show the random sequence controls. The black arrows delimit the windows for which the frequencies of E5-I3 were statistically significantly higher than those of I5-E3 ($p < 0.001$), with the exception of one case in each of zebrafish and human ($p < 0.05$).

and in the invertebrate *D. melanogaster* (by 4.5 fold). Interestingly, in *C. elegans* whose genome was believed to contain relatively few recently-gained introns[16], there was a 14-fold excess of E5-I3, driven in part by a low frequency of I5-E3 with $LIN \geq 6$ (compared to vertebrates).

To examine the distributions of E5-I3 and I5-E3 alignments for the full range of LIN from 0 to 20, we plotted their frequencies for human, mouse, zebrafish, chicken, *C. elegans* and *D. melanogaster* (Fig. 2a-f and Supplementary Tables 1-6). The black arrows delimited the window for which there was a significantly higher value observed for E5-I3 (black) than for I5-E3 (gray), as judged by *U*-test with $p < 0.001$, and the values for all vertebrates were significantly higher than for random sequences (Fig. 2a-d, black). For large $LIN \geq 8$ for human and *C. elegans*, ≥ 9 for the others), no significant differences were seen between E5-I3 and I5-E3 at distances that were more

than 10 nt away from the splice junctions, suggesting that the asymmetry was restricted to the vicinity near splice sites. For LIN between 2 and 9 in *C. elegans*, there was a very marked excess of E5-I3 relative to I5-E3 (Fig.2e), whereas *D. melanogaster* showed a more vertebrate-like profile (Fig.2f). For both *D. melanogaster* and *C. elegans*, the virtually complete absence of introns with long LINs was consistent with few recent intron gains[16, 17]. Moreover, the observed bias was not the result of multiple linked evolutionary events in a few genes for any of the six organisms (data not shown). The E5-I3 and I5-E3 alignments were also compared to scramble (mix-and-match) data produced by randomly aligning E5 with I3, and I5 with E3, from a non-redundant intron dataset, and again, statistically significant differences were seen in all cases (Supplementary Tables 7 and 8 for human and *C. elegans*).

Because it was known that U1 snRNA, in addition to base-pairing with sequences at the 5' end of the intron, also imposed a strong constraint on the terminal exonic cAG (within E5)[18], as does the binding of U2AF35 to the cAG region at the 3' end of the intron (within I3)[19], we repeated the analysis omitting the sequences located at positions -3 to +3 of both the 5' and 3' splice sites. As shown in Supplementary Table 9, the frequencies of LINs with values ≥ 1 and ≥ 5 for human E5-I3 and I5-E3 were significantly higher ($p < 0.05$) than those of the corresponding scrambles. The same held for the *C. elegans* E5-I3 alignments and their scrambles, whereas no statistical differences were observed for I5-E3 (Supplementary Table 10), consistent with the profiles shown in Fig. 2e. The E5-I3 values for the $LIN \geq 3$ (and therefore comparable to $LIN \geq 6$ in Fig. 1c) were still significantly higher ($p < 0.001$) than those for I5-E3 by about 0.3-fold for human introns and about 2.7-fold higher in the case of *C. elegans*. Taken together, our analyses indicated that the known preference of AG at the 5' (E5) and 3' (I3) splice sites, although strong, was not entirely responsible for the observed E5-I3 bias. Thus, the excess of introns with high LIN values for the E5-I3 alignment (compared to I5-E3) appeared not to be due simply to the conservation of sequences that are part of the splicing consensus motifs.

2.2 Asymmetric E5-I3 conservation supported by hexamer distributions of splice sites

To explore the specific nature of the splice junction sequences, we plotted the distributions of hexamers which are adjacent to the 5' and 3' splice sites (from positions -1 to -6 and from +1 to +6) for each of E5, I5, I3 and E3 for the total human intron set (Supplementary Fig. 1a-d), and for the $LIN \geq 6$ subset for E5 (Supplementary Fig.1e-f) and E3 (Supplementary Fig. 1g-h). For the total set, the distribution of exon hexamers (E3) located immediately downstream of introns, was much broader than for those upstream (E5) (Supplementary Fig. 1b vs. 1a). This uneven distribution of E5 for the I5-E3 with $LIN \geq 6$ dataset was supported by an E5 variance (σ^2) which was 50% larger than E3 variance for the E5-I3 $LIN \geq 6$ set (F -test, $p < 0.00001$) (Supplementary Fig. 3g vs. 3f grey), suggesting that E5 hexamers were not randomly distributed and more constrained than E3 ones.

These non-random distributions were also seen for the subset of E5 hexamers which end with CAG (Supplementary Fig. 1a-c). In the case of the I5 hexamer plots, the two sharp peaks (Fig. 3d, green), namely GTGAGT and GTAAGT were consistent with a role in U1 snRNA base-pairing (see Fig. 1a), as are those seen in the $LIN \geq 6$ dataset, namely GTAAGA and GTAAGT (Fig. 3h). For the profiles in Supplementary Fig. 1e (blue) and Supplementary Fig. 1h (blue) (which overlaid perfectly in keeping with their $LIN \geq 6$ values), the highest peaks represent the hexamer, CTGCAG, which incidentally was present in *Alu* repeats. Furthermore, the E5 hexamers (from positions -4 and -9) of E5-I3 with $LIN \geq 3$ are much clustered and more unevenly distributed than the E3 hexamers (from positions +4 and +9) of E5-I3 with $LIN \geq 3$ (Supplementary Fig.3 e vs f), while the corresponding I5 hexamers and I3 pyrimidine-rich hexamers were more restricted (Supplementary Fig. 3g&f). Therefore, the evolutionary constraints of E5 sequences upstream of 5' exonic CAG are much stronger than corresponding E3 sequences and much weaker than I5 and I3 sequences.

2.3 Use of splicing code to predict novel alternative splice sites.

The conservation beyond exonic cAG and evolutionary divergence between E5 and I5 hexamers raised the possibility that conservation of the 5' exonic sequences upstream of spliceosomal introns were similar to intronic-binding sites (IBS1 and IBS2) recognized by exonic-binding sites (EBS1 and EBS2) of their ancestors. If spliceosomal introns had inherited this splicing feature, one would expect that 5' exonic sequences immediately upstream of spliceosomal introns and 3' intronic sequences constituted the splicing code of spliceosomal introns and could be used to predict alternative splicing. The numbers of predicted splice sites were functions of intron sequence lengths and the size of splicing code. Based on our findings, a web-based software system (<http://splicingcodes.com/>) had been developed to accurately predict alternative splice sites of sequences from human, mice, *D. melanogaster* and *C. elegans* and more species would be added in the future.

For example, a mouse gene (*insr*) encoding an insulin receptor (IR), which was 128,255 bp in length and interrupted by 20 introns and well studied, was chosen to predict alternative splicing because mouse tissues could be obtained more conveniently. We first had used 9 bp E5 sequence plus the first intronic dinucleotides and the last 9 bp intronic nucleotide sequence to construct a 20 bp mouse splicing code table, which contained 290,000 unique sequences. A probability to find one of 20 bp identical sequences in the mouse genome is 9×10^{-13} . Assuming intron sizes of 150 bp to 50,000 bp, the mouse *insr* gene was predicted to encode large numbers of 4,631 putative novel splice sites (PPASSs), which was 3.4 times larger than the expected number of 1358 PPASSs ($p < 0.001$) (Fig.3a and Supplementary Table 11).

To further understand the dynamics of predicting mammalian PPASSs predicted by the software, we further used variable lengths of E5 sequences plus the first intronic

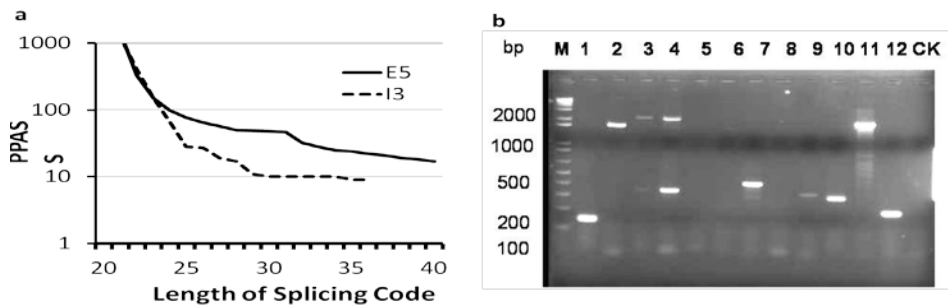


Fig.3 Verification of splicing code model. a) Relationship between PPASSs and lengths of splicing code. Black line indicates that numbers of PPASSs are predicted when the numbers of I3 sequences are fixed at 9 bp and variable lengths of E5 sequences plus the first intronic dinucleotide while dashed line represents numbers of PPASSs using fixed 9 bp of E5 sequences plus the first intronic dinucleotide and variable lengths of I3 sequences. b). RT-PCR verification of PPASSs. M is DNA markers. CK is negative control.

dinucleotide and fixed 9 bp of I3 sequences to look up their respective splicing code tables to predict PPASSs. The solid line in Fig. 3a showed that the numbers of PPASSs displayed two distinct phases as the total numbers of nucleotides were increased: from 20 to 24 bp, from which numbers of PPASSs were dramatically decreased and from 25 to 40 bp, from which numbers of PPASSs declined slowly and became almost flat. The predicted numbers of PPASSs were statistically significantly larger than what had been expected by chance ($p < 0.001$) (Supplementary Table 11). Similarly, we used fixed numbers of nine bp E5 sequences plus the first intronic dinucleotide and variable lengths of I3 sequences to search their corresponding mouse splicing code tables, respectively (Fig.3a, dashed line). The mouse *insr* I3 sequences showed two distinct phases similar to that of E5 sequences. The difference was that the numbers of PPASSs from 30 bp to 36 bp were 2.4 to 4.8 folds smaller than those observed by corresponding E5.

At 40 bp (or 29 bp of E5), the mouse *insr* gene was still predicted to encode 17 PPASSs, among which 7 were alternative splice sites of the existing exons (1, 9, 13, 14, 15, 16, and 17) and the remaining 10 were novel putative splice sites (Supplementary Fig.4). Based on intron-types, 12 of them were GT-AG, five were GC-AG introns and zero AT-AC introns. Using 36 bp (9bp of E5-(GT/GC/AT)-25 bp of I3 sequences), the mouse *insr* gene was predicted to encode the nine PPASSs, which showed characteristics similar to E5 (Supplementary Fig.5). Only, three out of 17 E5 PPASSs and two of nine E3 PPASSs had ≥ 6 bp of identical sequences between 5' and 3' splices had ruled out that long-stretch of DNA conservation between 5' and 3' splice sites in this study and in our previous study[14] was caused by template switching and missplicing[20].

2.4 Experimental verification of predicted putative alternative splice sites (PPASSs)

One of the natural questions was how many of 4,631 PPASSs predicted by the 20 bp splicing-code table of the 9 bp E5-(GT/GC/AT)-9 bp I3 sequences, were expressed. To verify our prediction, we had pseudo-randomly selected 12

PPASSs, which resulted in alternative splicing in the IR β tyrosine kinase region[5]. To perform isoform-specific PCR, a primer was designed to cross the putative splice junction of a PPASS while the other primer was located upstream or downstream of the mouse *insr* exonic sequences as shown in Supplementary Table 12. RT-PCR was performed on the pooled cDNAs from various mouse tissues as described in Materials and Methods and PCR products were separated on a 2.0% agarose gel. Nine out of 12 PCR reactions had products and were cloned and eight of them were shown to have inserts. Seven of the clones (7 out of 12) in Figure 3b had been verified by RT-PCR and sequencing of cloned RT-PCR products. Supplementary Figure 6 showed that these alternatively-spliced isoforms from these seven sequences resulted in truncated IR β subunits, which had functions different from full-length wild-type IR proteins[5]. Out of seven IR isoforms, three of these alternatively-spliced spliceovariants would produce almost identical truncated IR proteins and therefore had similar functions. These data indicated that the mouse insulin receptor gene encoded a far more complex system of alternatively-spliced isoforms than what had been discovered so far and were further confirmed by "splice-site walking" PCRs (data not shown). Many of these isoforms resulted in truncated proteins secreted into cellular matrix and blood, which were thought to be "protein-shedding" products[21]. Our computational and experimental data supported that mouse *insr* gene encoded large numbers of tissue-specific and low-level expressed alternatively-spliced isoforms, which enable mice to support their diverse functions[5].

3. Discussion

In this report, we have analyzed the spliceosomal intron datasets from invertebrate and vertebrates and have demonstrated that E5 sequences are more conserved than E3 ones. In *C. elegans*, there was a 14-fold excess of E5-I3 alignments with $LIN \geq 6$ than the corresponding I5 -E3 ones, which are much larger than invertebrate *D. melanogaster* whose E5-I3 alignments with $LIN \geq 6$ is 4.5 folds than those of I5-E3. Since both *D. melanogaster* and *C. elegans* are believed to contain relatively few recently-gained introns, one intriguing possibility is that this pronounced asymmetry might relate to *trans*-splicing which has played a significant role in gene expression in *C. elegans*[22] unlike in the other animals surveyed in this study.

In contrast to invertebrates, in mammalian spliceosomal introns, E5-I3 alignments are much longer, the majority of which are due to recently-gained introns. One of the questions is whether these longer E5-I3 alignments are

caused by 5' exonic CAG that is base-pairing with U1 snRNA. If this is the case, one would expect that the sequences upstream of the 5' CAG would be randomly distributed. The repeated analyses after omitting the sequences located at positions -3 to +3 of both the 5' and 3' splice sites have shown that the frequencies of LINs with values ≥ 1 and ≥ 5 for human and *C. elegans* E5-I3 and I5-E3 were significantly higher ($p < 0.05$) than those of the corresponding scrambles whereas much smaller differences were observed for I5-E3 (Supplementary Tables 9&10), consistent with E5 being more conserved than E3. In *C. elegans*, E5-I3 alignments of LIN values from 2 to 5 are statistically significantly higher than their corresponding scrambles (Supplementary Table 10). In contrast, there are no statistical differences between the I5-E3 alignments and their corresponding scrambles (Supplementary Table 10). These differences between E5-I3 and I5-E3 alignments are consistent with the notion that 5' exonic sequences beyond exonic CAG are conserved.

To confirm this, we performed hexamer plot analysis and have shown that the subset of E5 hexamers ending with CAG is not randomly-distributed (Supplementary Fig. 2.a-c). Since the 5' exonic CAG and conserved GTAAGT was recognized by the same U1 snRNA, one would expect that these sequences would share a common evolutionary trend. When the total E5 and I3 hexamers and the E5 and I3 hexamers with $LIN \geq 6$, which have been thought to be more recently-gained introns, are compared (Supplementary Fig.1a vs Fig.1e & Fig.1g and Fig.1b vs Fig.1f & Fig.1h), the total I5 hexamers have become more clustered together while total E5 hexamers are more evenly distributed. This opposite evolutionary trend of 5' splice sites is further confirmed by existence of much higher proportions of 5' exonic sequences ending without CAG in the invertebrate *D. melanogaster* and *C. elegans*, which have fewer recently-gained introns than the mammals, such as human and mice. Thus, the excess of introns with high LIN values for the E5-I3 alignment (compared to I5-E3) and conservation of 5' exonic sequences is caused by additional evolutionary forces.

Because 5' asymmetric conservation extends well beyond the 5' exonic CAG sequences at the splice junctions, it cannot simply reflect constraints imposed by the core known spliceosomal machinery, such as interactions with U1 snRNA and U2AF³⁵ and suggests that these sequences are constrained by yet-to-be characterized functions. This observation leads us to examine the features of the self-splicing group II ribozymes. It raises the possibility that conservation of 5' exonic sequences immediately upstream of

spliceosomal introns is similar to intronic-binding sites (IBS1 and IBS2) which are base-paired with exonic-binding sites (EBS1 and EBS2) of the self-splicing group II ribozymes. Therefore, we have proposed that the 5' exonic and 3' intronic sequences of the splice junctions constitute splicing codes of the spliceosomal introns, which are sequence-specifically decoded by as yet uncharacterized RNAs/proteins (Fig.4a&b)[23]. One can expect that these yet-to-be-characterized splicer RNAs/proteins can easily evolve from exonic-binding sequences (EBS1, EBS2 and EBS3)[24] while other conserved structural domains have evolved into spliceosomal U1 snRNAs. The mechanisms of deciphering splicing codes by splicer RNAs (or proteins) are similar to that of genetic codons decoded by tRNAs except that splicer RNAs/proteins hybridize to both E5 and I3 sequences, bring two exons together and guide spliceosome's removing intronic sequences. Both splicer RNA and protein models of pre-mRNA splicing can explain the conservation. However, evolutionary evidence favors the splicer RNA model over splicer-protein one. For example, *S. pombe* and *S. cerevisiae* have similar genome sizes (13.8 Mb vs 12.2 Mb) and encode similar protein-coding genes (4730 vs 5796)[25], which make it unlikely that *S. pombe* encodes superfamilies of RNA-binding proteins.

Since intronic-binding sites (IBS1 and IBS2) determine the

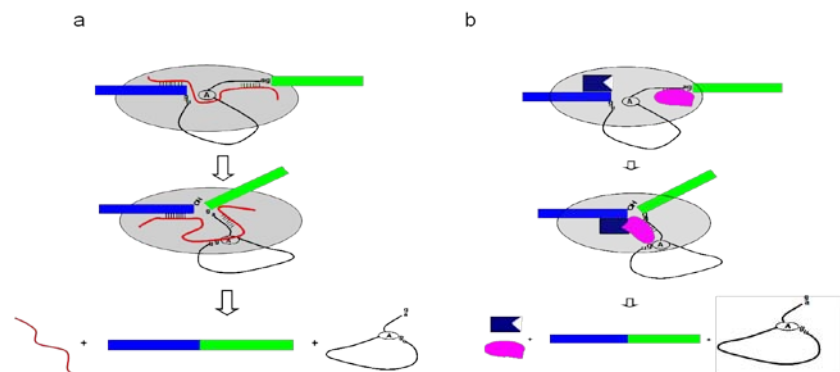


Fig.4 Schematic model of a nuclear pre-mRNA splicing pathway involving a splicer RNA (a) and proteins (b). The black and gray boxes represent the 5' and 3' exon sequences, respectively, and the shadowed oval represents a core spliceosome. The circled A is the branchpoint adenosine, and gu and ag represent the nucleotides typically present at the 5' and 3' ends of introns, respectively. a). Schematic model of E5 and I3 sequences that are recognized by splicer RNAs. The lines represent the intron and putative splicer RNA sequences, respectively. The vertical lines represent base-pairing between the putative splicer RNA and pre-mRNA (although these two *cis*-elements in a splicer RNA are not expected to be identical). The last nucleotide of the 5' exon and the last two nucleotides of the intron may lack perfect complementarities. For simplicity, a single splicer RNA has been shown, although the model is compatible with two RNAs (recognizing the 5' exon and 3' intron, respectively) in conjunction with other spliceosomal components. This model is conceptually similar to that first proposed by Holliday and Murray. b. Schematic model of E5 and I3 sequences that are recognized by as yet uncharacterized proteins. The E5 interacts in a sequence-specific manner with an as yet uncharacterized protein (square) and I3 is recognized by a different unknown protein (oval). These two proteins interact with each other to assist in bringing together the 5' and 3' splice sites.

splicing specificity of the self-splicing group II ribozymes, splicing code of spliceosomal introns can be used to predict alternative splicing accurately. Since splicing code and their deciphering splicer RNAs/proteins co-evolve like the base-pairing between IBSs and EBSs, the splicing code is species-specific. When total lengths of splicing code are determined, the numbers of expected splice sites are functions of length of the intronic sequence (or total gene sequence length) and the size of splicing code table. That is, the longer an intron sequence is, the more alternative splice sites are expected. If an organism has a larger size splicing code table, a gene is more likely to have more isoforms. Since human has the largest numbers of spliceosomal introns characterized so far and some of the longest introns, one would expect that human encodes large numbers of alternative splice sites. This may explain why mammalian genomes encode much smaller numbers of protein-coding genes than expected while the numbers of introns and intron lengths are significantly increased.

Based on the splicing code model, a web-based software system has been developed to predict alternative splicing. Now, it can predict alternative splice sites from four model organisms: human, mice, *D. melanogaster* and *C. elegans*. More species will be added in the future. Many factors may affect the software accuracy to predict alternative splice sites. Since we have limited information about evolutionary history of IBS3-EBS3 interaction and our 3' intronic data come from the computation analysis, we still need to test combinations of different lengths of 5' exonic and 3' intronic sequences to identify the best model to predict alternative splicing in mammals. Since the EBS1 and EBS2 are located in different parts of the self-splicing group II ribozymes, it is reasonable to believe that the 5' exonic sequences with splicer RNAs/proteins may be similar to those IBSs-EBSs, which may also have significant impacts on the prediction accuracy. For predicting alternative splice sites of mammalian genes, since we cannot distinguish conserved sequences of splicing code from repeated elements, which have played an important role in exonization and alternative splicing, additional characteristics of splicing code are required to be identified to differentiate them from recent-duplications with assistance of experimental data.

Both our computational and experimental data suggest that mammalian insulin receptor (*insr*) gene encodes much larger numbers of alternatively-spliced isoforms than what have been known previously, which are much larger than the numbers of single nucleotide polymorphisms (SNP) and many of which have very similar functions at the protein level (redundancy). Therefore, impacts on mammals by each of genetic mutations will be reduced to the minimum and their ability to adapt to their environments would be maximized. That means that a single gene possesses complex traits and the phenotypes controlled by a single gene behave like complex traits.

For example, the insulin resistance, which is a physiological condition where the natural hormone insulin becomes less effective at lowering blood sugars, is thought to be complex traits. The genetic and physiological studies have

shown that the insulin receptor gene is responsible for Leprechaunism (OMIM 246200), the most extreme form of the insulin resistance syndromes, Rabson-Mendenhall syndrome (OMIM 262190), severe forms of insulin resistance syndrome, and type A insulin resistance (OMIM 147670), milder forms of insulin resistance. However, it seems that there is no relationship between insulin resistance and insulin receptors in mammals. One of the reasons is that the mammalian *insr* gene encodes extremely complex and highly-redundant alternatively-spliced isoforms, which are supported by our verified >30 isoforms. The majority of low-level and tissue-specifically expressed PCR products encode truncated soluble insulin receptors that are secreted into blood and/or extracellular matrices and have been thought to be generated by a poorly-characterized "protein shedding" process[21]. Since soluble insulin receptors have the same affinities as insulin receptors on plasma membranes to bind insulin competitively, one can envision that elevated soluble receptors will form receptor-insulin complexes and reduce amounts of insulin to reach target cells and tissues and therefore cause insulin resistance. That insulin resistance is caused by elevated soluble insulin receptors has been directly supported by the fact that increasing blood soluble insulin receptors have been shown to be associated with type I and type II diabetes among Japanese patients[26]. However, the method used in that study can detect small fractions of insulin receptor isoforms. One would expect that these results will be incomplete and less dependable. If we can detect all forms of alternatively-spliced isoforms of insulin receptors, it is quite possible that we can establish clearer relationships between the alternatively-spliced isoforms and insulin resistances. It will help us to develop simple and accurate diagnosis and therapy methods for diabetes and metabolic syndrome.

4. Conclusions

We have demonstrated that 5' exonic sequences immediately upstream of spliceosomal introns are similar to the intronic-binding sites (IBS1 and IBS2) of the self-splicing group II ribozymes—ancestors of spliceosomal introns. Based on this observation, we have proposed that 5' exonic and 3' intronic sequences constitute splicing code of spliceosomal introns, which are deciphered by splicer RNAs/proteins originated from exonic-binding sites (EBS) of the group II ribozymes. Based on our findings, we have developed a web-based software system to predict pre-mRNA splicing. Computational and experimental data have demonstrated that the mouse insulin receptor gene encodes complex alternatively-spliced isoforms, many of which are secreted into cellular matrix and blood, where they are able to bind the insulin before it reaches target cells and tissues. This work has laid the foundation to develop simple, accurate and personalized diagnosis and therapy methods for complex diseases from diabetes to cancer.

5. Materials and Methods

See Supplementary data.

6. Acknowledgement

We thank Prof. Linda Bonen, Prof. Pascale Goldschmidt and Prof. Benoit Chabot for their supports and participations of this work, preparation and writing of the manuscript, and critical review of the manuscript. We thank R. Holliday for suggestions and comments on the manuscript. We are grateful to Drs. Xiongzhaoh Zhu and Aizhong Liu for their assistance in the statistical analyses.

References

- [1] F. Rodriguez-Trelles, R. Tarrío, and F. J. Ayala, "Origins and evolution of spliceosomal introns," *Annu Rev Genet*, vol. 40, pp. 47-76, 2006.
- [2] Q. Pan, O. Shai, L. J. Lee, et al., "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nat Genet*, vol. 40, pp. 1413-5, 2008.
- [3] H. Keren, G. Lev-Maor, and G. Ast, "Alternative splicing and evolution: diversification, exon definition and function," *Nat Rev Genet*, vol. 11, pp. 345-55.
- [4] T. A. Cooper, L. Wan, and G. Dreyfuss, "RNA and disease," *Cell*, vol. 136, pp. 777-93, 2009.
- [5] A. Belfiore, F. Frasca, G. Pandini, et al., "Insulin receptor isoforms and insulin receptor/insulin-like growth factor receptor hybrids in physiology and disease," *Endocr Rev*, vol. 30, pp. 586-623, 2009.
- [6] M. J. Moore, "From birth to death: the complex lives of eukaryotic mRNAs," *Science*, vol. 309, pp. 1514-8, 2005.
- [7] A. M. Pyle, O. Fedorova, and C. Waldsich, "Folding of group II introns: a model system for large, multidomain RNAs?," *Trends Biochem Sci*, vol. 32, pp. 138-45, 2007.
- [8] S. Valadkhan, "The spliceosome: a ribozyme at heart?," *Biol Chem*, vol. 388, pp. 693-7, 2007.
- [9] B. R. Graveley, A. N. Brooks, J. W. Carlson, et al., "The developmental transcriptome of *Drosophila melanogaster*," *Nature*, vol. 471, pp. 473-9.
- [10] J. K. Pickrell, A. A. Pai, Y. Gilad, et al., "Noisy splicing drives mRNA isoform diversity in human cells," *PLoS Genet*, vol. 6, pp. e1001236.
- [11] X. D. Fu, "Towards a splicing code," *Cell*, vol. 119, pp. 736-8, 2004.
- [12] Y. Barash, J. A. Calarco, W. Gao, et al., "Deciphering the splicing code," *Nature*, vol. 465, pp. 53-9.
- [13] M. Soller, "Pre-messenger RNA processing and its regulation: a genomic perspective," *Cell Mol Life Sci*, vol. 63, pp. 796-819, 2006.
- [14] D. Zhuo, R. Madden, S. A. Elela, et al., "Modern origin of numerous alternatively spliced human introns from tandem arrays," *Proc Natl Acad Sci U S A*, vol. 104, pp. 882-6, 2007.
- [15] P. den Hollander, S. K. Rayala, D. Coverley, et al., "Ciz1, a Novel DNA-binding coactivator of the estrogen receptor alpha, confers hypersensitivity to estrogen action," *Cancer Res*, vol. 66, pp. 11021-9, 2006.
- [16] A. Coghlan and K. H. Wolfe, "Origins of recently gained introns in *Caenorhabditis*," *Proc Natl Acad Sci U S A*, vol. 101, pp. 11362-7, 2004.
- [17] L. Banyai and L. Patthy, "Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues," *FEBS Lett*, vol. 565, pp. 127-32, 2004.
- [18] I. Carmel, S. Tal, I. Vig, et al., "Comparative analysis detects dependencies among the 5' splice-site positions," *Rna*, vol. 10, pp. 828-40, 2004.
- [19] S. Wu, C. M. Romfo, T. W. Nilsen, et al., "Functional recognition of the 3' splice site AG by the splicing factor U2AF35," *Nature*, vol. 402, pp. 832-5, 1999.
- [20] S. W. Roy and M. Irimia, "When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis," *Bioessays*, vol. 30, pp. 601-5, 2008.
- [21] K. Hayashida, A. H. Bartlett, Y. Chen, et al., "Molecular and cellular mechanisms of ectodomain shedding," *Anat Rec (Hoboken)*, vol. 293, pp. 925-37.
- [22] T. Blumenthal and K. S. Gleason, "Caenorhabditis elegans operons: form and function," *Nat Rev Genet*, vol. 4, pp. 112-20, 2003.
- [23] V. Murray and R. Holliday, "Mechanism for RNA splicing of gene transcripts," *FEBS Lett*, vol. 106, pp. 5-7, 1979.
- [24] A. Jacquier and F. Michel, "Multiple exon-binding sites in class II self-splicing introns," *Cell*, vol. 50, pp. 17-29, 1987.
- [25] V. Wood, R. Gwilliam, M. A. Rajandream, et al., "The genome sequence of *Schizosaccharomyces pombe*," *Nature*, vol. 415, pp. 871-80, 2002.
- [26] The Soluble Insulin Receptor Study Group, "Soluble insulin receptor ectodomain is elevated in the plasma of patients with diabetes," *Diabetes*, vol. 56, pp. 2028-35, 2007.

