



# SHORT ABSTRACTS

---

# ISBRA 2011

7<sup>TH</sup> INTERNATIONAL SYMPOSIUM ON  
BIOINFORMATICS RESEARCH  
AND APPLICATIONS

May 27-29, 2011  
Central South University  
Changsha, China

<http://www.cs.gsu.edu/isbra11>

# ISBRA 2011 Short Abstracts

## Contents

Zhi John Luin. cRNA: Integration of High-throughput Data .....	1
Yuan Li and John Adeyeye. Some remarks on the limit cycle of monotone functions with symmetric connection graph.....	2
Amrendar Kumar, Abhilasha Singh and Tripti Singh. Use of natural compounds as AchE inhibitors for the treatment of early stage Alzheimer’s disease – an insilico approach.....	5
Ekaterina Ermakova, Mikhail Gelfand and Dmitry Rodionov. CodY regulon in Bacillaceae.....	10
Bo Hou and Hang He. Close Upper Bound for rSPR distance between two rooted binary phylogenetic trees.....	11
Chanchala Kaddi, Chang Quo and May Wang. Analytical Comparison of Information Criteria for Systems Biology Model Selection.....	15
Wen-Lin Huang, Chyn Liaw and Shinn-Ying Ho. Predicting non-classical secretory proteins by using Gene Ontology terms and physicochemical properties.....	19
Jiang Zhang and Lihui Bai. A Network-based Approach for Hospital Capacity Management in a Pandemic.....	23
Jiahai Shi and Jianxing Song. Dynamical Inactivation/Enhancement of the Catalytic Machinery of the SARS 3C-like Protease by its Evolutionarily Acquired Extra-domain.....	24
Guicheng Zhang, Jack Goldblatt and Peter Lesouëf. Genome-wide association studies with a mother/father-child paired design hold the key to dissecting the aetiology of complex traits.....	26
Qiang Jiang. Predicting multiplex subcellular location of proteins using protein-protein network: a comparative study.....	30
Hao Jiang and Wai-Ki Ching. Support Vector Machine for Classification of DNA repair genes.....	34
Katerina Korenblat, Zeev Volkovich and Alexander Bolshoy. Novel Phylogenomic Method for Prokaryotes.....	38
Leo Liberti, Carlile Lavor and Antonio Mucherino. An exponential algorithm for the Discretizable Molecular Distance Geometry Problem is polynomial on proteins .....	40
Zhiyong Wang and Jinbo Xu. A Conditional Random Fields Method for RNA Sequence-Structure Relationship Modeling and Conformation Sampling.....	44
Andrea Szabóová, Ondřej Kuželka, Filip Zelezny and Jakub Tolar. Searching for Important Amino Acids in DNA-binding Proteins for Histogram Methods.....	48
Jin Zhang and Yufeng Wu Finding Deletions with Exact Break Points from Noisy Low Coverage Paired-end Short Sequence Reads.....	52
Hong Cai, Jianying Gu, Zhan Zhou and Yufeng Wang. Genome and Systems Evolution in Chlamydiae.....	54
Lu Fan, Staffan Kjelleberg and Torsten Thomas. Accurate Reconstruction of Microbial Community From Environmental Shotgun Sequences Avoiding Primer Bias.....	58

## ISBRA 2011 Short Abstracts

Li-Ping Tian, Hengyao Lu and Fang-Xiang Wu. Nonlinear Model-Based Clustering for Periodically Expressed Gene Profiles.....	62
Yixiang Shi and Yuan-Yuan Li Study of the Xylose Isomerase Reveals Certain Fingerprint of Its Evolution.....	67
Serghei Mangul, Irina Astrovskaya, Bassam Tork, Ion Mandoiu and Alex Zelikovsky. Viral Quasispecies Reconstruction Based on Unassembled Frequency Estimation.....	71
Wen-Chi Chang and Ying-Chi Wen. Comprehensive analysis of promoter features related to tissue-specific genes in rice.....	76
Damiano Piovesan, Pier Luigi Martelli, Piero Fariselli, Andrea Zauli, Ivan Rossi and Rita Casadi. Extended and robust protein sequence annotation over conservative non hierarchical clusters .....	80
Akshaye Dhawan and Alison Nolan. Designing Reusable User-Interfaces for Querying a Collection of Neuroscience Ontologies.....	84
Kwangsoo Kim, Chol Shin and Hong Seo Ryoo. A new approach to the analysis of GWAS data.....	88
Mikolaj Rybinski, Michal Lula, Slawomir Lasota and Anna Gambin. Tav4SB: grid environment for analysis of kinetic models of biological systems.....	92
Xueyi Wang. A Normalized Weighted RMSD Measure for Protein Structure Superposition.....	96
Vadim Mozhayskiy and Ilias Tagkopoulos. Accelerated Microbial Evolution in Complex Dynamic Environments through Step-wise Adaptation.....	100
Newton Miyoshi, Daniel Pinheiro, Wilson Silva Junior and Joaquim Felipe. Computational Framework to Support Data Storage and Analysis in Translational Medicine.....	104
Lingling Zheng, Minhua Chen, Joseph Lucas and Lawrence Carin. Bayesian Elastic Net for Multi-Class Classification and Survival Analysis.....	108
Min-Seok Kwon, Kyunga Kim, Sungyoung Lee, Wonil Chung, Sung-Gon Yi, Junghyun Namkung and Taesung Park. GWAS-GMDR: a program package for genome-wide scan of gene-gene interactions with covariate adjustment based on multifactor dimensionality reduction.....	109
Jaehoon Lee, Soyeon Ahn, Sohee Oh and Taesung Park. SNP-PRAGE: SNP-based Parametric Robust Analysis of Gene set Enrichment.....	113
Yuanyuan Huang, Stephen Bonett and Zhijun Wu. P.R.E.S.S. -- R-package for Exploring Residual-Level Protein Structural Statistics.....	117
Pavel Skums, Zoya Dimitrova, David Campo, Gilberto Vaughan, Livia Rossi, Jonny Yokosawa, Alex Zelikovsky, Yury Khudyakov and Joseph C Forbi. Efficient Error Correction for Deep Sequencing of Viral Amplicons.....	133
Ming Fang, Weiling Li and Raj Sunderraman. An Efficient Constraint Planning Algorithm for Distributed Bio-Ontologies.....	138
Ajit Pandey, Sonia Avasthi and Pranjali Pandey. Screening of natural compounds for the treatment of Diabetes Mellitus Type 2 -An insilico approach.....	142

# incRNA: Integration of High-throughput Data

*Zhi John Lu*

*China Tsinghua University-Yale University, China*

*urluzhi@gmail.com*

I will present an integrative, machine-learning method, incRNA, for whole-genome identification of non-coding RNAs (ncRNAs) in *C. elegans*. It combines a large amount of expression data from the modENCODE consortium, RNA secondary-structure stability, and evolutionary conservation at the protein and nucleic-acid level. Using this model, we were able to separate known ncRNAs from coding sequences and other genomic elements with high accuracy (97% AUC on an independent validation set), and find more than 7,000 novel ncRNA candidates, among which more than 1,000 are located in intergenic regions. We estimate based on the validation set that 91% of the ~7K predicted ncRNAs are true positives. We then analyzed fifteen of them by RT-PCR and detected the expression of fourteen. In addition, we characterized the novel ncRNA candidates and found that they have distinct expression patterns across developmental stages, tend to use novel RNA structural families, and are targeted by specific transcription factors (~59% of intergenic ncRNAs). Overall, our study identifies many new potential ncRNAs in *C. elegans*. Furthermore, I also applied incRNA to human using ENCODE data. I added the chromatin features in human ncRNA prediction. In addition to ncRNA finding, I will also introduce my other work in modENCODE project, such as analysis of transcription factor (TF) binding sites and integration of miRNA-TF network.

Keywords: High-throughput data, noncoding RNA, RNA secondary structure prediction, RNA-seq, ChIP-seq

**SOME REMARKS ON THE LIMIT CYCLE OF MONOTONE FUNCTIONS  
WITH SYMMETRIC CONNECTION GRAPH**

YUAN LI<sup>1\*</sup> AND JOHN O. ADEYEYE<sup>2\*</sup>.

ABSTRACT. In this short notes, we give a low bound to the maximum size of the limit cycle of a monotone function constructed in [1] by providing an antichain. We conjecture that this antichain has the maximum size. We provide a conjecture about any infinite walking along a tree. Based on this conjecture, one can show that the length of the limit cycle of monotone function with connection graph equal to a tree can not be a multiple of four. These combinatorial conjectures may be of independent interests.

1. INTRODUCTION AND NOTATIONS

Finite dynamical systems, that is, discrete dynamical system with a finite state space, have been used extensively in systems biology to model a variety of biochemical networks, such as metabolic networks, gene regulatory networks and signal transduction networks. To determine the limit cycles or fixed points for different networks is of great importance. Monotone functions (positive functions) have been well studied for their complexity and application in circuits [4, 5, 6, 7]. In [3], Robert and Tchunte established some relationships between the circuits of connection-graph and the circuits of the iteration-graph of a monotone discrete dynamical system. Julio et al considered the discrete networks with multi state monotone functions in [1]. In this short notes, we give an antichain. By using the results of the integer solutions of a linear equations, we obtain its length. Hence, a lower bound of the maximum size of the limit cycle of a monotone function constructed in [1] is obtained. We conjecture our antichains are the only antichains with the maximum size. We also provide another conjecture about the walking in a tree. This conjecture will be useful to determine the length of the limit cycle of a monotone system with symmetric connection graph equal to a tree.

In the following, we will give some definition.

Let  $S = \{(x_1, x_2, \dots, x_n) | 0 \leq x_i \leq m - 1, i = 1, 2, \dots, n\}$ . Where  $m \geq 2$  and  $n$  are given positive integers. Let  $F = (f_1, f_2, \dots, f_n) : S \rightarrow S$  denote a function of  $n$  variables and  $m$  states.  $F$  is Boolean if  $m = 2$ .

Let  $X = (x_1, \dots, x_n)$  and  $Y = (y_1, \dots, y_n) \in S$ , we define  $X \leq Y$  iff  $x_i \leq y_i$  for any  $i = 1, \dots, n$ .

We say  $F$  is monotone if  $F(X) \leq F(Y)$  given  $X \leq Y$ .

Let  $A \subset S$ , we say  $A$  is an antichain of  $S$  if for any  $\{X, Y\} \subset A$ , we have neither  $X \leq Y$  nor  $Y \leq X$ .

We call a sequence of vectors  $[X^0, X^1, \dots, X^{p-1}, X^0]$  a limit cycle of  $F$  of length  $p$  if  $X^{j+1} = F(X^j)$  for  $j = 0, \dots, p - 2$  and  $X^0 = F(X^{p-1})$ . It can be easily shown that  $\{X^0, X^1, \dots, X^{p-1}\}$  is an antichain (see [1]) given that  $F$  is monotone.

The Connection graph (dependency graph) of  $F$ , denoted by  $G_C(F) = (V_C, E_C)$ , is a directed graph where  $V_C = \{v_1, \dots, v_n\}$  is the set of nodes and arc  $(v_i, v_j)$  is in  $E_C$  iff the function  $f_j$  depends on  $x_i$ . In this notes, we assume  $E_C$  is symmetric, i.e.,  $(v_i, v_j) \in E_C$  iff  $(v_j, v_i) \in E_C$ . Let  $\Gamma(V_i) = \{v_j | (v_i, v_j) \in E_C\}$ .

---

<sup>0\*</sup>: Supported by an award from the USA DoD # W911NF-11-10166.

*Key words and phrases.* Monotone function, Discrete network, Graph, antichain, Systems biology.

2. SOME REMARKS

In Proposition 1 of [1], a monotone function was constructed with limit cycle of maximum possible length which is equal to the size of an antichain. We have

**Remark 2.1.**  $A_{k,n,m} = \{(x_1, x_2, \dots, x_n) | x_1 + x_2 + \dots + x_n = k\}$  is an antichain of  $S$ . Where  $0 \leq x_i \leq m - 1$ ,  $k = 0, \dots, n(m - 1)$ .

The proof of this remark is evident.

Regards to the size, we have

**Lemma 2.2.** [8] *The number of integer solutions of the linear equation  $x_1 + x_2 + \dots + x_n = k$ , with the restrictions  $s_i \leq x_i \leq m_i$ ,  $i = 1, 2, \dots, n$ , with  $s \leq k \leq m$ ,  $s = s_1 + s_2 + \dots + s_n$ ,  $m = m_1 + m_2 + \dots + m_n$ , and for  $u_i = m_i - s_i \geq 0, i = 1, 2, \dots, n$ , is given by*

$$N_{n,k}(u_1, u_2, \dots, u_n) = \binom{n+k-s-1}{n-1} + \sum_{r=1}^n (-1)^r \sum \binom{n+k-s-u_{i_1}-u_{i_2}-\dots-u_{i_r}-r-1}{n-1},$$

where in the inner sum, the summation is extended over all  $r$ -combinations  $\{i_1, i_2, \dots, i_r\}$  of the  $n$  indices  $\{1, 2, \dots, n\}$ .

Let  $s_i = 0$  and  $m_i = m - 1$  for all  $i$  in the above sum, then  $u_i = m - 1$ . Let  $[n] = \{1, 2, \dots, n\}$ . The sum can be simplified as

$$\begin{aligned} & \binom{n+k-1}{n-1} + \sum_{r=1}^n (-1)^r \sum_{\{i_1, \dots, i_r\} \subset [n]} \binom{n+k-r(m-1)-r-1}{n-1} = \\ & = \binom{n+k-1}{n-1} + \sum_{r=1}^n (-1)^r \binom{n}{r} \binom{n+k-1-rm}{n-1} = \sum_{r=0}^n (-1)^r \binom{n}{r} \binom{n+k-1-rm}{n-1}. \end{aligned}$$

Hence, when  $0 \leq x_i \leq m - 1$ , we get the number of solutions of

$$x_1 + x_2 + \dots + x_n = k \tag{2.1}$$

$$\text{is } |A_{k,n,m}| = \sum_{r=0}^n (-1)^r \binom{n}{r} \binom{n+k-1-rm}{n-1}.$$

We have

**Lemma 2.3.**  $|A_{k,n,m}| = |A_{n(m-1)-k,n,m}|$ .

*Proof.* Let  $y_i = m - 1 - x_i$ ,  $i = 1, 2, \dots, n$ , then  $y_1 + y_2 + \dots + y_n = n(m - 1) - (x_1 + x_2 + \dots + x_n)$ . Hence,  $y_1 + y_2 + \dots + y_n = n(m - 1) - k$  iff  $(x_1 + x_2 + \dots + x_n) = k$ .  $\square$

When  $m = 2$ , with a direct evaluation, we know the number of solutions of equation 2.1 should be  $\binom{n}{k}$ . Hence we obtain the following combinatorial identity

$$\binom{n}{k} = \sum_{r=0}^n (-1)^r \binom{n}{r} \binom{n+k-1-2r}{n-1}. \tag{2.2}$$

It is well known that  $\binom{n}{k-1} \leq \binom{n}{k}$  given  $k \leq \lfloor \frac{n}{2} \rfloor$ . Hence, we have the following

**Conjecture 2.4.**  $|A_{k-1,n,m}| \leq |A_{k,n,m}|$  given  $k \leq \lfloor \frac{n(m-1)}{2} \rfloor$

Because of the famous Sperner's theorem [9], the greatest size of any antichain of  $S$  is  $\binom{n}{\lfloor \frac{n}{2} \rfloor}$  when  $m = 2$ . We naturally have

**Conjecture 2.5.** *The greatest size of all the antichains of  $S$  is  $|A_{\lfloor \frac{n(m-1)}{2} \rfloor, n, m}|$ .  $A_{\lfloor \frac{n(m-1)}{2} \rfloor, n, m}$  is the only antichain attain this cardinality when  $n(m-1)$  is even. When  $n(m-1)$  is odd, only  $A_{\frac{n(m-1)-1}{2}, n, m}$  and  $A_{\frac{n(m-1)+1}{2}, n, m}$  attain this cardinality.*

Numerical results of small values of  $n$  and  $m$  shows these two conjectures are true.

Hence, we guess the limit cycle of monotone function  $F$  over  $S$  has the greatest size

$$\sum_{r=0}^n (-1)^r \binom{n}{r} \binom{n + \lfloor \frac{n(m-1)}{2} \rfloor - 1 - rm}{n-1}.$$

In [1], the authors proposed the following

**Conjecture 2.6.** [1] *The maximum length of the limit cycles of a monotone function with more than two states and symmetric connection graph equal to arbitrary tree is at most two.*

We guess the following extension of Property 2 in [1] is also true, we have

**Conjecture 2.7.** *Let  $G = (V, E)$  be a tree and let  $V_\infty = (v_{i_k})_{k \in N}$  a sequence of nodes of  $V$  such that  $v_{i_k} \in \Gamma(v_{i_{k-1}})$  for every  $k \in N$ . Then for any positive even integer  $2a$ , there exist  $r \in N$  such that  $v_{i_r} = v_{i_{r+2a}}$ .*

Julio et al [1] have proved the above conjecture is true for  $a = 1, 2$ .

Based on Conjecture 2.7, one can use the same way in [1], to prove that the length of the limit cycle of monotone functions in conjecture 2.6 can not be multiple of four.

In Conjecture 2.7, without lost the generality, we can assume that every node of  $v$  will be visited infinite times. Otherwise, we delete those nodes which was visited only finite times and consider  $V_\infty = (v_{i_k})_{k \in N}$  when  $k$  is great enough.

Conjecture 2.7 is obviously true when the tree is a simple star. But it seems not easy to prove this conjecture even if the tree is a chain.

### 3. CONCLUSION

We gave some remarks about the limit cycles of monotone functions discussed in [1]. We proposed some conjectures which may be of independent combinatorial interests.

*Acknowledgment:*

The authors are supported by an award from the USA DoD # W911NF-11-10166.

### REFERENCES

- [1] Julio Aracena, Jacques Demongeot and Eric Goles, "On the limit cycle of monotone functions with symmetric connection graph", *Theoretical Computer Science* 322 (2004), pp. 237-244.
- [2] E. Goles and L. Salinas, "Comparison between parallel and serial dynamics of Boolean networks", *Theoretical Computer Science* 396 (2008), pp. 247-253 .
- [3] Y. Robert, M. Tchuente, "Connection-graph and iteration-graph for monotone Boolean function", *Discrete Applied Mathematics* 11 (1985), pp. 245-253 .
- [4] J. C. Bioch, T. Ibaraki, "Complexity of identification and dualization of positive Boolean functions", *Inform. and Comput.* 123 (1995), pp. 50-63.
- [5] M. Fredman, L. Khachiyan, "On the complexity of dualization of monotone disjunctive normal forms", *J. Algorithms* 21 (1996), pp. 618-628 .
- [6] K. Makino, T. Ibaraki, "The maximum latency and identification of positive Boolean functions", *SIAM J. Comput.* 26(5) (1997), pp. 1363-1383 .
- [7] N. N. Nurmeev, "On the complexity of the circuit realization of almost all monotone Boolean functions", *Izv. Vyssh. Uchebn. Zaved. Mat* 5 (1985), pp. 64-70 .
- [8] Charalambos A. Charalambides, "Enumerative Combinatorics", *Chapman & Hall/CRC*, 2002.
- [9] L. Comtet, "Analyse Combinatoire", *Presses Universitaires de France, Paris*, 1970.

<sup>1</sup>MATHEMATICS DEPARTMENT, WINSTON SALEM STATE UNIVERSITY, NC 27110,USA, EMAIL: LIYU@WSSU.EDU <sup>2</sup>MATHEMATICS DEPARTMENT, WINSTON SALEM STATE UNIVERSITY, NC 27110,USA,, EMAIL: ADEYEEJ@WSSU.EDU,

**USE OF NATURAL COMPOUNDS AS AchE INHIBITORS FOR THE  
TREATMENT OF EARLY STAGE ALZHEIMER'S DISEASE-AN *INSILICO*  
APPROACH**

Amrendar kumar<sup>1\*</sup>, Abhilasha Singh<sup>1</sup>, Biplab Bhattacharjee<sup>2</sup>

1. Amity Institute of Biotechnology, Lucknow campus, UP, India
2. Institute of Computational Biology (IOCB), Bangalore, India

\*Corresponding Author:  
amrendar2290@gmail.com  
Mobile: 9532395556



**Abstract**

Traditionally, drugs were discovered by testing compounds manufactured in time consuming multi-step processes against a battery of in vivo biological screens. Promising compounds were then further studied in development, where their pharmacokinetic properties, metabolism and potential toxicity were investigated. Here we present a study on herbal lead compounds and their potential binding affinity to the effectors molecules of major disease like Alzheimer's disease. Clinical studies demonstrate a positive correlation between the extent of Acetyl cholinesterase enzyme and Alzheimer's disease. Therefore, identification of effective, well-tolerated acetyl cholinesterase represents a rational chemo preventive strategy. This study has investigated the effects of naturally occurring nonprotein compounds polygala and bulbocapnine that inhibits acetylcholinesterase enzyme. The results reveal that these compounds use less energy to bind to acetylcholinesterase enzyme and inhibit its activity. Their high ligand binding affinity to acetylcholinesterase enzyme introduce the prospect for their use in chemopreventive applications in addition they are freely available natural compounds that can be safely used to prevent Alzheimer's Disease.

Keywords: Alzheimer's disease, acetyl cholinesterase, Docking, ADME, Acetyl cholinesterase inhibitor.

**Introduction**

Alzheimer's disease, most common form of dementia is incurable, degenerative, and terminal disease mostly diagnosed in people over 65 years of age. The disease advances with symptoms include confusion, irritability and aggression, mood swings, language breakdown, long-term memory loss, and the general withdrawal of the sufferer as their senses decline. Gradually, bodily functions are

lost, ultimately leading to death.

Acetylcholinesterase is also known as AChE. It degrades the neurotransmitter acetylcholine, producing choline and acetate group.

Acetylcholinesterase is mainly found at neuromuscular junctions and cholinergic synapses in the central nervous system, where its activity serves to terminate synaptic transmission.

An acetyl cholinesterase inhibitor (often abbreviated AChEI) or anti-cholinesterase is a chemical that inhibits the cholinesterase enzyme from breaking down acetylcholine, increasing both the level and duration of action of the neurotransmitter acetylcholine.

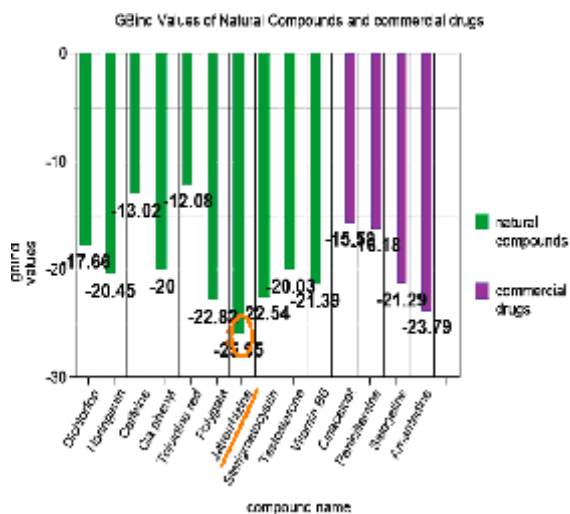
**Methodology**

Natural compounds displaying anti alzheimers activity was collected from various research articles. The investigation drug Penicillamine was used as a reference drug in the studies. These small molecules were screened on the basis of the Lipinski's Rule of 5. The screened compounds were put into enzyme – ligand interactions by docking with Quantum 3.3.0. The Target enzyme Acetylcholinesterase whose PDBID is 2W9I was taken for further analysis. The least energy conformers was obtained from MarvinSketch. Then docking is done between the protein Acetylcholinesterase and screened compounds which act as Acetylcholinesterase inhibitors with QUANTUM 3.3.0. The best three results obtained were analysed under HEX. Their IC50 value was also analysed using QUANTUM 3.3.0. Then graphs were plotted by analysing the values.

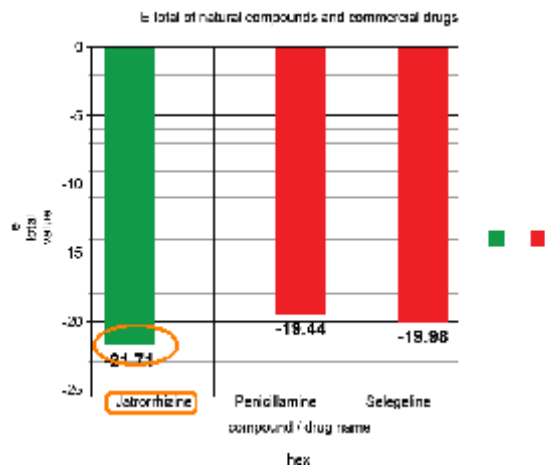
## ISBRA 2011 Short Abstracts

S.no	Name of compound	Gbind	Rms
1.	Dichlorfop	-17.66	69.85
2.	Naringenin	-20.45	72.55
3.	Caffeine	-13.02	80.61
4.	Cl <sub>a</sub>	-20.00	66.34
5.	Toluidine red	-12.08	69.48
6.	Polygala	-22.82	73.28
7.	Jatrorrhizine	-25.95	96.90
8.	Sterigmatocystin	-22.54	73.39
9.	Testosterone	-20.03	74.21
10.	Vitamin b6	-21.39	72.23

S.no	Name of drug	Gbind	Rms
1.	Cinacalcet	-15.58	95.83
2.	Penicillamine	-16.18	90.59
3.	Selegeline	-21.29	90.18
4.	Amantadine	-23.79	94.72



# ISBRA 2011 Short Abstracts



6. Massoulie J, Pezzementi L, Bon S, Krejci E, Valette F (1993). "Molecular and Cellular Biology of Cholinesterases.". *Prog. Brain Res.* 93 (1): 31–91. PMID 8321908.
30. Masserano JM, Weiner N (1983). "Tyrosine hydroxylase regulation in the central nervous system.". *Mol. Cell. Biochem.* 53-54 (1-2): 129–52. doi:10.1007/BF00225250. PMID 6137760
31. Meloni R, Biguet NF, Mallet J (2002). "Post-genomic era and gene discovery for psychiatric diseases: there is a new art of the trade? The example of the HUMTH01 microsatellite in the Tyrosine Hydroxylase gene.". *Mol. Neurobiol.* 26 (2-3): 389–403. doi:10.1385/MN:26:2-3:389. PMID 12428766.
32. Joh TH, Park DH, Reis DJ (1979). "Direct phosphorylation of brain tyrosine hydroxylase by cyclic AMP-dependent protein kinase: mechanism of enzyme activation.". *Proc. Natl. Acad. Sci. U.S.A.* 75 (10): 4744–8. doi:10.1073/pnas.75.10.4744. PMID 33381.
33. Haycock JW, Ahn NG, Cobb MH, Krebs EG (1992). "ERK1 and ERK2, two microtubule-associated protein 2 kinases, mediate the phosphorylation of tyrosine hydroxylase at serine-31 in situ.". *Proc. Natl. Acad. Sci. U.S.A.* 89 (6): 2365–
9. doi:10.1073/pnas.89.6.2365. PMID 1347949.
35. Haycock JW (1990). "Phosphorylation of tyrosine hydroxylase in situ at serine 8, 19, 31, and 40.". *J. Biol. Chem.* 265 (20): 11682–91. PMID 1973163.
36. Craig SP, Buckle VJ, Lamouroux A, et al. (1986). "Localization of the human tyrosine hydroxylase gene to 11p15: gene duplication and evolution of metabolic pathways.". *Cytogenet. Cell Genet.* 42 (1-2): 29– doi:10.1159/000132246. PMID 2872999.
37. Grima B, Lamouroux A, Boni C, et al. (1987). "A single human gene encoding multiple tyrosine hydroxylases with different predicted functional characteristics.". *Nature* 326 (6114): 707–11. doi:10.1038/326707a0. PMID 2882428.
38. Kaneda N, Kobayashi K, Ichinose H, et al. (1987). "Isolation of a novel cDNA clone for human tyrosine hydroxylase: alternative RNA splicing produces four kinds of mRNA from a single gene.". *Biochem. Biophys. Res. Commun.* 146 (3): 971–5. doi:10.1016/0006-291X(87)90742-X. PMID 2887169.
39. Kobayashi K, Kaneda N, Ichinose H, et al. (1987). "Isolation of a full-length cDNA clone encoding human tyrosine hydroxylase type 3.". *Nucleic Acids Res.* 15 (16): 6733. doi:10.1093/nar/15.16.6733. PMID 2888085.

## CodY regulon in Bacillaceae

Ekaterina Ermakova<sup>1</sup>, Mikhail Gelfand<sup>1</sup> and Dmitry Rodionov<sup>1,2</sup>,

<sup>1</sup> Institute for information transmission problems RAS, Bolshoy Karetny per. 19,  
127994 Moscow, Russia

<sup>2</sup> Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road,  
92037 La Jolla, CA, USA

[ermakova@iitp.ru](mailto:ermakova@iitp.ru), [gelfand@iitp.ru](mailto:gelfand@iitp.ru), [rodionov@burnham.org](mailto:rodionov@burnham.org)

**Keywords:** comparative genomics, regulation of transcription, *Bacillaceae*.

Transcription factor CodY is a global regulator of nutrient limitation and amino acid metabolism in *Firmicutes* studied mostly in *B. subtilis*. Starting with a set of 42 experimentally verified binding sites in *B. subtilis* ([1] and B. Belitsky, personal communication), CodY regulon was analysed in *Bacillaceae* using comparative genomics methods. We show that CodY regulon in *Bacillaceae* consists of at least 135 clusters of orthologous operons having conserved binding sites.

### References

1. Belitsky, B.R., Sonenshein, A.L.: Genetic and Biochemical Analysis of CodY-binding Sites in *Bacillus subtilis*. *J. Bacteriol.* 190, 1224—1236 (2008)

## Close Upper Bound for rSPR distance between two rooted binary phylogenetic trees

Bo Hou and Hang He

Department of Computer Science and Engineering,  
University of Connecticut,  
Storrs, CT, 06269, U.S.A.  
bo.hou@huskymail.uconn.edu, Hang.He@enr.uconn.edu

**Abstract.** Phylogenetic tree is a commonly used model to denote the evolutionary process for a set of species. However, some biology events, say hybridization event, would lead to some genes deriving from different ancestors. In other words, the evolution history same set of species may have different phylogenetic tree to represent. This introduces the problem of measuring difference between two phylogenetic trees. One of metric describing such difference is *rooted subtree prune and regraft* distance (rSPR distance). In this work, we present an algorithm to get a close upper bound for rSPR distance.

**Keywords:** Phylogenetic tree, rooted subtree prune and regraft distance, a close upper bound

### 1 Introduction

Although rSPR distance is a good metric to measure difference of two rooted binary phylogenetic trees, it is not easily to obtain because it is a NP-hard problem.

Previous study has given some algorithms to obtain exact rSPR distance. Paper [1, 3, 5] use *Fixed-parameter algorithm* to convert the optimal problem to decision problem to deal with rSPR problem. Paper [4] encodes this problem into the Integer Linear problem. Although these algorithms can get exact rSPR distance, these algorithms can not handle large-size phylogenetic tree data. We try all these algorithms on two phylogenetic trees which have one hundred and twelve leaves and none of them can give an exact distance in a reasonable time. Fortunately for us, some efficient approximation algorithms are proposed to provide a lower bound of rSPR distance. Paper [1, 2, 3] provide a 3-approximations of rSPR distance but with different time complexity. This give us an inspiration that if the close upper bound of rSPR distance can be calculated efficiently, branch and bound algorithm may be designed to obtain the exact rSPR distance quickly even for large-size data. So in this study, we propose an efficient greed algorithm to obtain a close upper bound for rSPR distance.

## 2 Method

### 2.1 BACKGROUND

A binary phylogenetic tree is such a tree that its leaves compose of a set  $X$  and degree of its internal node is three except root. The binary phylogenetic tree is also called a *binary phylogenetic X-tree*, if such a tree has  $n$  leaves set  $X$  which is  $\{1, 2, 3 \dots n\}$  and each elements in  $X$  represents a leaf in the tree. Assuming  $V$  is subset of  $X$ ,  $T(V)$  denotes the smallest subtree of  $T$  that connects all nodes in  $V$ .  $T|V$  represents a tree from  $T(V)$  by forced contractions. Forced contraction will replace a two degree node  $A$  and its two connecting edges  $(B, A)$  and  $(A, C)$  with a single edge  $(B, C)$ .

As described above, rooted Subtree prune and regraft (rSPR) distance is a metric to measure two rooted binary phylogenetic  $X$ -trees. Given an  $X$ -tree  $T$ , *Subtree prune and regraft* operation cuts an edge  $(a, b)$  in  $T$  and separates  $T$  into two parts: subtrees  $T_a$  containing node  $a$  and  $T_b$  containing node  $b$ . Next, it divides an edge of  $T_b$  using a new vertex  $b'$  and construct an edge connecting node  $a$  and node  $b'$  so that two subtrees  $T_a$  and  $T_b$  is in same component again. The remaining node  $b$  is eliminated by forced contraction. See figure 1.

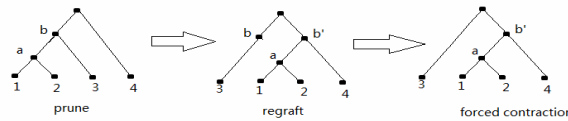


Figure 1 the subtree prune and regraft operation

Maximum agreement forest (MAF) is often used to obtain rSPR distance. Through agreement forest, the problem of a series of SPR operation transformed into the problem of the set of edges cutting to obtain MAF of two evolutionary trees. For two phylogenetic trees  $T_a, T_b$ , if we add a virtual root  $\rho$  to each of the tree, then rSPR distance equals to the number of trees in MAF minus one [4]. Definition of maximum agreement is given below:

For two rooted phylogenetic  $X$ -trees  $T_a$  and  $T_b$ , collection  $T_1, T_2, T_3 \dots T_k$  are a series of rooted trees with their leaf sets  $L_1, L_2, L_3 \dots L_k$ . We call collection  $T_1, T_2 \dots T_k$  is a agreement forest of  $T_a$  and  $T_b$ , if

- (i) The leaf sets  $L_1, L_2, L_3 \dots L_k$  partition  $X \cup \{\rho\}$ , in other words,  $L_i \cap L_j = \emptyset$  and  $\cup L_i = X \cup \{\rho\}$ .
- (ii) For all  $i$  equals  $1, 2 \dots k$ ,  $T_a|L_i = T_b|L_i$ .
- (iii) The trees in each of sets  $\{T_a(L_i)|i=\rho, 1, 2 \dots k\}$  and  $\{T_b(L_i)|i=\rho, 1, 2 \dots k\}$  are vertex-disjoint rooted trees.

An MAF is an agreement forest such that  $k$  is the minimum, i.e. no other agreement of  $T_a$  and  $T_b$  could be found whose number of trees is less than this  $k$ .

## 2.2 PROPOSED ALGORITHM

Assuming in certain time, collection  $\{T_1, T_2 \dots T_m\}$  is agreement forest (AF) of two trees  $T_a$  and  $T_b$ , we choose two components of this collection, say  $T_i, T_j$ , and merge them together if after merging, new collection is still AF of  $T_a$  and  $T_b$ . Repeating this process, until there is no two components that can be merged to form a less number AF. This is the basic idea of algorithm. There are two key technical for this algorithm. Firstly, some rules must be constructed so that it can be efficiently judge whether the merging of two components lead to an AF of  $T_a$  and  $T_b$ . Secondly, how can we choose two components in the collection so that when the algorithm stops, the number of final AF can be as small as possible?

For the first one, let's assume that the collection is  $L_1, L_2 \dots L_x$  and we select two components  $L_i, L_j$  to see if they can be merged. We first judge whether  $T_a|_{L_i \cup L_j}$  equals  $T_b|_{L_i \cup L_j}$ . If this is not true, they can not merge. If this is true, according to definition of MAF, we need to check whether  $T_a(L_i \cup L_j)$  and  $T_b(L_i \cup L_j)$  have vertex that joins with  $T_a(L_k)$  and  $T_b(L_k)$  and  $k$  does not equal  $i$  and  $j$ . If such vertex does not exist, then these two components can be merged. According to definition, this guarantees new collection is AF of  $T_a$  and  $T_b$ .

When there is more than one pair of components to choose for merge, which pair should be selected? Here a greed strategy is used. Firstly, we get the score for each pair. Secondly, the pair that has minimum score is merged. The score is calculated in this way:

For two components  $L_i$  and  $L_j$ , trees  $T_a(L_i \cup L_j)$  and  $T_b(L_i \cup L_j)$  is obtained. For every node in  $T_a(L_i \cup L_j)$ , we get every children of that node in  $T_a$  such that these children do not exist in  $T_a(L_i \cup L_j)$ . Let's denote  $e(T_a(L_i \cup L_j))$  the set that contains all such branches in  $T_a$ , so does  $e(T_b(L_i \cup L_j))$ . Moreover,  $\forall x \in e(T_a(L_i \cup L_j)) \cup e(T_b(L_i \cup L_j))$ , we denote  $s(x)$  be the number of different components belonging to the collection under edge  $x$  in tree  $T_a$  or  $T_b$ . Then score of pair  $L_i$  and  $L_j$  is the sum of all  $s(x)$  for every  $x$  in  $e(T_a(L_i \cup L_j)) \cup e(T_b(L_i \cup L_j))$ .

## 3 EXPERIMENT

Here we use two different data sets to evaluate greed algorithm. The first data is simulated data which has large number of leaves and second data set is coming from real data set of EEEB [6]. Experiment results are shown in table 1 and table 2. From table 1, it can be seen that all the result of Greed algorithm equals to the exact rSPR distance. For table 2, we can see although five greed solutions are different from the exact rSPR distance, the difference between greed solution and rSPR distance is small (not more than 2).

From the experiment result, it can be concluded that this greed algorithm gives a close upper bound of rSPR distance. For the future work, the performance is still needed to be improved and furthermore, we would design an efficient branch and bound algorithm based on the greed algorithm in this paper to handle large phylogenetic tree.



Table 1 experiment result of simulate data

	rSPR distance	Greed solution
G1_100_10	10	10
G2_100_10	10	10
G3_100_10	9	9
G4_100_10	9	9
G5_100_10	10	10
G6_100_10	9	9
G7_100_10	10	10
G8_100_10	10	10
G9_100_10	10	10
G10_100_10	10	10

Table 2 experiment result of EEEB data

	rSPR distance	Greed solution
Ndhf phyB	12	12
ndhF rbcL	10	10
ndhF rpoC2	11	11
ndhF waxy	7	9
ndhF ITS	19	20
phyB rbcL	4	4
phyB rpoC2	6	6
phyB waxy	3	3
phyB ITS	8	9
rbcL rpoC2	11	12
rbcL waxy	6	6
rbcL ITS	13	13
rpoC2 waxy	1	1
rpoC2 ITS	14	15

## References

1. Magnus Bordewich, Catherine McCartin, Charles Semple. : A 3-approximation algorithm for the subtree distance between phylogenies. *Journal of Discrete Algorithms*. Volume 6, Issue 3 458-471 (2008)
2. Estela M. Rodrigues, Marie-France Sagot, Yoshiko Wakabayashi.: The maximum agreement forest problem: Approximation algorithms and computational experiments. *Theoretical Computer Science*, 374(1-3):91-110,2007
3. Chris Whidden and Norbert Zeh.: A Unifying View on Approximation and FPT of Agreement Forests. *ALGORITHMS IN BIOINFORMATICS*, Volume 5724/2009, 390-402.(2009)
4. Yufeng Wu.: A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, 25(2):190-196, 2009
5. Chris Whidden, Robert G. Beiko and Norbert Zeh.: Fast FPT Algorithms for Computing Rooted Agreement Forests: Theory and Experiments (Extended Abstract). Accepted to SEA 2010
6. Grass Phylogeny Working Group.: Phylogeny and subfamilial classification of the grasses (poaceae). *Ann. Mo. Bot. Gard.* 88, 373 - 457 (2001)

## Analytical Comparison of Information Criteria for Systems Biology Model Selection

Chanchala Kaddi<sup>1</sup>, Chang F. Quo<sup>1</sup>, May D. Wang<sup>1</sup>

<sup>1</sup> The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA

{gtg538v, gtg801f}@mail.gatech.edu, maywang@bme.gatech.edu

**Abstract.** Existing information criteria for model selection emphasize accuracy and prediction value, but neglect ‘well-posed’-ness and economy in terms of the number of model parameters. Because systems biology models generally involve a large number of parameters that often must be estimated, ‘well-posed’-ness and economy are important factors for model selection. To address this problem, we propose two scoring methods to quantify these factors, and compare them analytically to existing well-known criteria. Results indicate that the proposed criteria are useful to illustrate potential costs of parameter estimation, i.e., inherent model uncertainty, in terms of the proportion of free parameters with respect to the total number of parameters. This is an initial step to develop an information criterion for model selection that is targeted for the systems biology community, so as to complement efforts to standardize model annotation and facilitate model sharing.

**Keywords:** model selection, systems biology, meta-data

### 1 Introduction

Model selection is difficult because different models may describe the same system, sometimes without definitive difference in terms of model results, e.g., accuracy and prediction value. From literature, some recurring themes for developing information criteria for model selection may be identified. These include (a) accuracy, (b) prediction value, (c) ‘well-posed’-ness, and (d) economy of parameters [1, 2]. Some criteria, in terms of accuracy and prediction, are relatively well-defined, e.g., using statistical tests [3], and experimental validation [4], while others, in terms of ‘well-posed’-ness and economy, are more ambiguous. In particular, systems biology models generally involve a large number of parameters that need to be estimated, increasing model uncertainty in the process, so ‘well-posed’-ness and economy are important for model selection. Furthermore, and especially for the systems biology community, the lack of clear information criteria for model selection may hinder efforts to standardize model annotation, e.g., using MIRIAM [5], and to facilitate model sharing, e.g., using SBML [6].

Here, we propose alternative scoring methods to quantify models in terms of ‘well-posed’-ness and economy, and compare these analytically to existing information

criteria that focus on model accuracy. This effort is a continuation of previous work [7, 8], which did not include an analytical treatment.

## 2 Methods

We propose two potential criteria to quantify ‘well-posed’-ness, i.e., in terms of the number of free parameters ( $k$ ) and economy of parameters, i.e., in terms of the total number of parameters ( $r$ ). These methods are then compared analytically to existing information criteria that focus on accuracy, i.e., Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

### 2.1 Proposed criteria for ‘well-posed’-ness and economy

Both the p-score and s-score are designed to emphasize the disparity between the total number of parameters ( $r$ ) and the number of free parameters ( $k$ ). These criteria aim to illustrate the amount of uncertainty involved in parameter estimation during model development. Less uncertainty is preferred.

The p-score is defined as:

$$p = \frac{k \cdot \ln(k)}{r}$$

where  $k$  is the number of free parameters, and  $r$  is the total number of parameters. A model with lower p-score indicates that the proportion of free parameters with respect to the total number of parameters, effectively degrees-of-freedom, is smaller than that of a model with higher p-score.

The s-score is defined as:

$$s = \frac{1}{r} + \left| \frac{k}{r} - 1 \right|$$

where  $k$  is the number of free parameters, and  $r$  is the total number of parameters. A model with higher s-score indicates that the proportion of free parameters with respect to the total number of parameters is smaller than that of a model with lower s-score.

### 2.2 Akaike and Bayesian Information Criteria

Both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are measures of the goodness-of-fit, i.e., accuracy, of a statistical model. Models with minimum AIC and BIC values are preferred.

The AIC is defined as:

$$AIC = 2k - 2\ln(L)$$

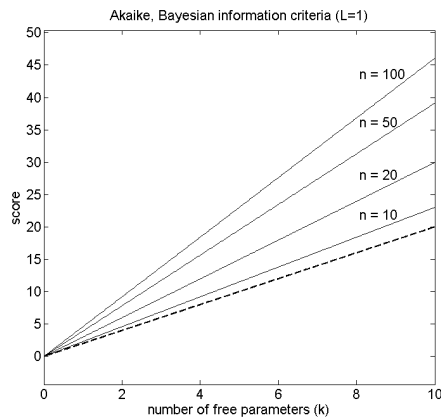
where  $k$  is the number of free parameters, and  $L$  is the maximized value of the likelihood function for the estimated model.

The BIC is defined as:

$$BIC = k \ln(n) - 2\ln(L)$$

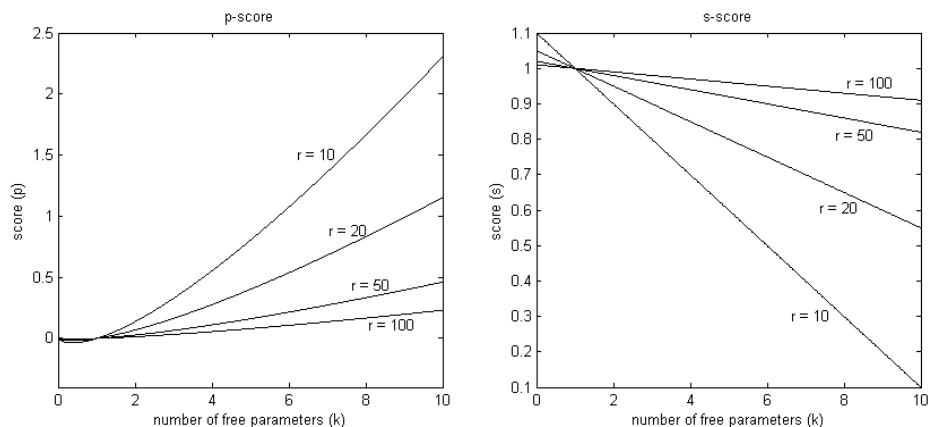
where  $k$  is the number of free parameters,  $n$  is the number of observations, and  $L$  is the maximized value of the likelihood function for the estimated model.

### 3 Results and Discussion



**Fig 1:** Akaike (dashed line) and Bayesian (solid lines) information criteria, as a function of the number of free parameters ( $k$ ) with fixed maximized value of likelihood function ( $L=1$ ).

For a given  $k$ , differences in BIC reflect changes in the number of observations ( $n$ ) (Fig. 1), but the fraction of parameters which are unknown is not considered in either. On the other hand, differences in p- and s-score reflect changes in the total number of parameters ( $r$ ) (Fig. 2). Thus, the potential cost of parameter estimation is apparent when for the same number of free parameters, a model with more total parameters is selected over another with fewer total parameters. In addition, the s-score, when plotted over higher numbers of free parameters (not shown), provides an opportunity to consider how potential users, who with preference for particular models, may use it to choose from publicly available models, e.g., on the BioModels database [9].



**Fig 2:** The p-score (left), and s-score (right), as a function of the number of free parameters ( $k$ ) with varying total number of parameters ( $r = \{10, 20, 50, 100\}$ ).

These scoring methods are only an initial step to developing information criteria for model selection, in particular systems biology models, in terms of ‘well-posed’-ness and economy. The proposed criteria illustrate potential costs of parameter estimation when the proportion of free parameters increases with respect to the total number of parameters. These scoring methods are currently being tested using public models.

**Acknowledgments:** This research has been supported by grants from the Parker H. Petit Institute for Bioengineering and Bioscience (IBB), Johnson & Johnson, Georgia Tech BIMS Initiative, NIH (BRP R01CA108468, CCNE U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), Microsoft Research, and the NSF (Graduate Research Fellowship Award to CK).

## 5 References

1. Klipp E, Liebermeister W et al: Systems biology: a textbook. Wiley-VCH (2009)
2. Smye S and Clayton R: Mathematical modelling for the new millenium: medicine by numbers. *Med. Eng. Phys.* 24: 565–574 (2002)
3. Cedersund G and Roll J: Systems biology: model based evaluation and comparison of potential explanations for given biological data. *FEBS J.* 276: 903–922 (2009)
4. Cedersund G, Roll J et al: Model-based hypothesis testing of key mechanisms in initial phase of insulin signaling. *PLoS Comput. Biol.* 4(6) (2008)
5. Le Novère N, Finney A et al: Minimum information required in the annotation of models (MIRIAM). *Nat. Biotechnol.* 23: 1509–1515 (2005)
6. Hucka M, Bergmann FT et al: The systems biology markup language (SBML): language specification for level 3 version 1 core. *Nature Precedings*: <http://dx.doi.org/10.1038/npre.2010.4959.1> (2010)
7. Kaddi C, Oden ED et al: Exploration of quantitative scoring metrics to compare systems biology modeling approaches. In: *IEEE EMBC*. 1121--1124 (2007)
8. Kaddi C, Quo CF, and Wang MD: Quantitative metrics for bio-modeling algorithm selection. In: *IEEE EMBC*. 4613--4616 (2008)
9. Li C, Donizelli M et al: BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Bio.* 4: 92 (2010)

## Predicting non-classical secretory proteins by using Gene Ontology terms and physicochemical properties

Wen-Lin Huang

Department of Management  
Information System  
Asia Pacific Institute of Creativity  
Miaoli, Taiwan  
[wenlinhuang2001@gmail.com](mailto:wenlinhuang2001@gmail.com)

Chyn Liaw

Institute of Bioinformatics and  
Systems Biology,  
National Chiao Tung University  
Hsinchu 30068, Taiwan  
[chynliaw@gmail.com](mailto:chynliaw@gmail.com)

Shinn-Ying Ho

Institute of Bioinformatics and  
Systems Biology,  
National Chiao Tung University  
Hsinchu 30068, Taiwan  
[syho@mail.nctu.edu.tw](mailto:syho@mail.nctu.edu.tw)

**Abstract**—Eukaryotic secretory proteins that traverse classical ER-Golgi pathway are usually characterized by short N-terminal signal peptides. However, several secretory proteins lacking the signal peptides are found to be exported by a non-classical secretion pathway. Therefore, predicting non-classical secretory proteins regardless of the N-terminal signal peptides is necessary for developing a critical computational approach. Several prediction methods have been proposed by using various types of features to predict secretory proteins. However, prediction performance seems not acceptable. This study proposes an SVM-based prediction method, namely ProSec-iGOX, which uses a major set of informative Gene Ontology (GO) terms and a minor set of assistance features. Physicochemical properties as the assistance features are useful when a query protein sequence without homologous protein with annotated GO terms. Two data sets, S25 and S40, having the identity 25% and 40%, respectively, are adopted for performance comparisons. The ProSec-iGOX yields test accuracies of 95.1% and 96.8% when adopting on the data sets S25 and S40 respectively. The latter accuracy (96.8%) is significantly higher than that of SPRED (82.2%), which uses frequency of tri-peptides and short peptides, secondary structure, physicochemical properties as input features to a random forest classifier. The experimental results show that GO terms are effective features for predicting non-classical secretory proteins.

**Keywords**—Gene Ontology; secretory; physicochemical properties; non-classical secretion; signal peptides

### I. INTRODUCTION

Eukaryotic protein secretion normally routes through the endoplasmic reticulum (ER) and Golgi, ending up in a secretory vesicle fusing to the cell membrane (Fig. 1) [1]. This pathway that traverses the endoplasmic reticulum (ER) and Golgi apparatus is classical secretory pathway [2]. The secretory proteins are usually characterized by short N-terminal signal peptides with intrinsic signals for their transport and localization in the cell [3]. However, several secretory proteins lacking a signal peptide such as fibroblast growth factors (FGF-1, FGF-2), interleukins (IL-1 alpha, IL-1 beta) galectins, thioredoxin, viral proteins and parasitic surface proteins have been found to be exported by a non-classical secretion pathway (Fig. 1) [1, 4]. Additionally, the molecular mechanisms of non-classical secretion in eukaryotes are still unknown even though the phenomenon of non-classical secretion was discovered more than a decade ago [5]. Therefore, an automated approach regardless of the N-terminal signal peptides is necessary to predict non-classical secretory proteins.

Several methods without using N-terminal signal peptides have been proposed for the identification of secretory proteins via the classical [6] and non-classical secretory pathways [2, 5]. For example, SecretomeP uses the number of atoms, positively charged residues, propeptide cleavage sites, protein sorting, low complexity regions, and transmembrane helices as an input for a neural network [2]. The SRTPRED utilizes amino acid composition (AAC), their order and similarity search to predicts secretory proteins irrespectively of N-terminal signal peptides [7]. The SPRED uses frequency of tri-peptides and short peptides, secondary structure, physicochemical properties with a random forest classifier [5]. These prediction methods use many types of sequence-based features but it is difficult to assess which feature type is the most informative except IBCGA-SVM [8] extracting a small set of physicochemical properties with SVM classifier. However, physicochemical properties seem not to be discriminative features for sequence-based prediction projects [9, 10].

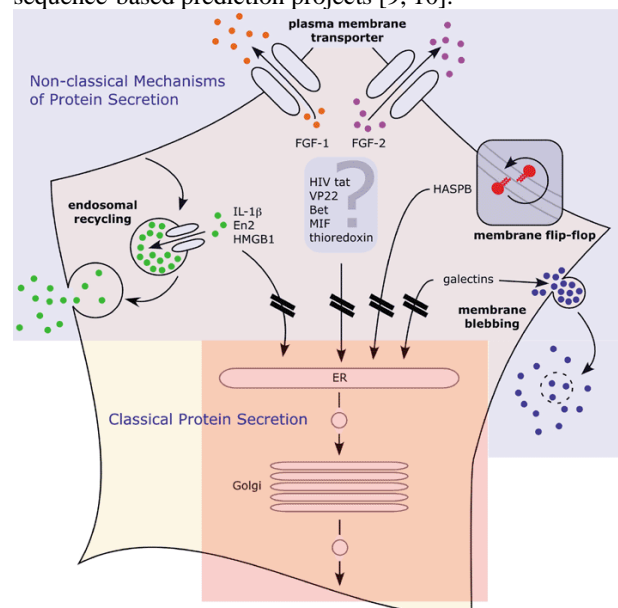


Figure 1 Potential export routes of non-classical protein secretion

Contrarily, Gene-Ontology (GO) based methods perform well, compared with some sequence-based and structure-based methods to predict subcellular localization [9-11], where GO is a controlled vocabulary of terms split into three related ontology consisting of molecular function, biological processes and cellular components [12]. The growth of GO databases in size has increased accuracies of GO-based prediction methods [13, 14]. Additionally, few sequences

without homologous protein can be annotated by GO so that assistance features (e.g. physicochemical properties) are useful. Therefore, this study utilizing GO terms and physicochemical property composition (PCC) features proposes an SVM-based prediction method ProSec-GOX for discriminating classical and non-classical secretions of eukaryotic proteins. Compared with SPRED[5], ProSec-GOX without further using feature selection have high performance in predicting non-classical secretory proteins.

## II. MATERIALS

### A. Data sets

A data set S40 with 40% sequence identity obtained from another work [5] has 780 non-classical (positive data set) and 1980 classical (negative data set) secretory protein sequences. The protein sequences are taken from the SWISS-PROT [15] protein sequence database according to the annotation information in the CC (comment or notes) and ID (identification) fields. The proteins in the data set were screened strictly using the following criteria: 1) only the sequences annotated with “mammalian” in the ID field are collected; 2) sequences with uncertain annotation labels such as ‘probable’, ‘potential’ and ‘by similarity’ are removed; 3) sequences annotated with keywords “extracellular” are collected as a positive data set; 4) signal peptides are removed from the positive data set; and 5) sequences annotated in cytoplasm and/or nucleus subcellular locations are taken as a negative data set; and 6) sequences with 40% identity were operated by a culling program [16].

Additionally, we established another data set S25 of 372 non-classical and 1011 classical secretory proteins with 25% sequence identity using a culling program [16] to evaluate the proposed method. All of the proteins in the data set S25 are divided randomly into two separated sets with sizes in the ratio 1:1, for training and independent testing, respectively. However, for comparison, the numbers of sequences for training and independent testing in the data set S40 are the same as those of SPRED [5]. Table I presents the numbers of proteins within non-classical and classical classes in the data sets S25 and S40.

TABLE I. NUMBERS OF PROTEINS IN THE DATA SETS S40 AND S25

Class	S40L	S40T	S25L	S25T
classical	600	180	186	186
Non-classical	600	1380	506	505
Total	1200	1560	692	691

### B. Gene Ontology annotation

The newest version of the GO database [17] (released on Jan. 14, 2011) contains 58,932 terms in the three branches of biological process, molecular function and cellular component. This study utilized the GOA database including GO annotations for non-redundant proteins from many species in the GOA/UniProt database [18]. The GOA database was downloaded directly from GOA [19] (released on August 24, 2010) [19].

Protein accession numbers are necessary when querying the GOA database to obtain their annotated GO terms. For novel proteins, BLAST [20, 21] was used to obtain homologies with known accession numbers from the query protein to retrieve GO terms. The parameter  $e$ -value  $e$  and the number  $h$  of homologies in BLAST are critical to the quality of the homologies and the retrieval of GO terms. We tested the following  $e \in \{10^0, 10^{-1}, 10^{-2}, \dots, 10^{-10}\}$  and found that  $e=10^{-9}$  is a good trade-off value [22]. Even with this threshold value, BLAST with the different values of the parameter  $h$  could retrieve different numbers of hits. Sequentially, the best value of  $h$  was determined from  $h \in \{1, 2, \dots, 5\}$  and  $e=10^{-9}$  using a step-wise method with the  $k$ -nearest-neighbor ( $k$ -NN) classifier, where  $k=1$  [23]. Figure 2 shows the best accuracies obtained by using  $(h, e) = (1, 10^{-9})$  for the data sets S40 and S25

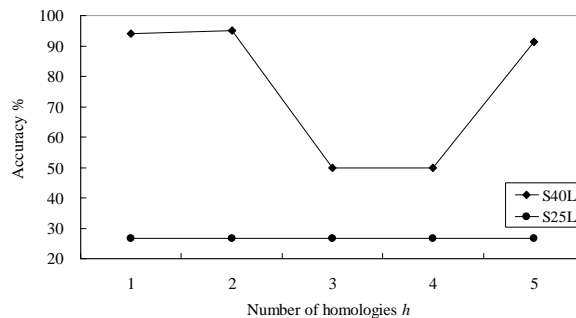


Figure 2 Results obtained by using various numbers of homologies

## III. METHODS

The prediction method ProSec-GOX consists of two parts, model training and prediction of query sequences. The model training comprises two classifiers, SVM-iGO and SVM-PCC, due to that ProSec-GOX uses a major set of GO terms and a minor set of assistance features to predict non-classical secretory proteins, where PCC features as the assistance features for query sequences without corresponding to GO terms.

### A. Model training

TABLE II. TEN TOP-RANK FREQUENCY DESCRIPTORS OF GO TERMS

Rank	S25L Go term	No. of sequences	Rank	S40L Go term	No. of sequences
1	GO:0005576	164	1	GO:0005576	568
2	GO:0005615	73	2	GO:0005634	421
3	GO:0005515	34	3	GO:0005515	415
4	GO:0005125	26	4	GO:0005737	363
5	GO:0005179	24	5	GO:0005615	258
6	GO:0006810	24	6	GO:0045449	181
7	GO:0006955	22	7	GO:0003677	176
8	GO:0042742	20	8	GO:0006350	173
9	GO:0006952	19	9	GO:0046872	155
10	GO:0016787	18	10	GO:0016787	142

Let  $n$  be the total number of GO terms that have ever appeared for all training proteins. From the  $n$  GO terms, determine the 500 top-rank frequency descriptors of GO terms. The occurrence frequency  $f_k$  is the number of the sequences annotated by the  $k^{\text{th}}$  GO term where  $k = 1, 2, \dots, n$ .

Table II shows the 10 top-rank frequency descriptors of GO terms; for example, GO:0005576 (Extracellular region) annotated by 164 and 568 sequences in the data set S25L and S40L, respectively.

With regard to the PCC features, they are as the assistance features for query sequences without corresponding to GO terms. Protein sequence is represented as a 531-dimensional feature vector. The 531 features are derived from the 531 physicochemical properties of AAindex [24] by averaging over the protein sequence.

Consequently, the 500 top-rank GO terms and 531 PCC features are in conjunction with a series of binary classifiers of LIBSVM to design two SVM-based classifiers, SVM-iGO and SVM-PCC, respectively. A radial basis kernel function  $\exp(-\gamma \|x^i - x^j\|^2)$  is adopted, where  $x^i$  and  $x^j$  are training samples, and  $\gamma$  is a kernel parameter. There are two parameters  $\gamma$  and a cost parameter  $C$  to be tuned in using the SVM. In this study, the best values of parameters  $C$  and determined using a step-wise approach were employed to the two SVM-based methods, SVM-GO and SVM-PCC, where  $\gamma \in \{2^{-7}, 2^{-6}, \dots, 2^8\}$  and  $C \in \{2^{-7}, 2^{-6}, \dots, 2^8\}$ .

The leave-one-out cross-validation (LOOCV) is considered to be the most rigorous and objective test that can always yield a unique result for a given data set [23]. Although bias-free, this test is very computationally demanding and is often impractical for large data sets. The N-fold cross-validation not only provides a bias-free estimation of the accuracy at a much reduced computational cost, but is also considered as an acceptable test for evaluating prediction performance of an algorithm [25]. Additionally, for comparison to SPRED, SVM-iGO and SVM-PCC use the prediction accuracy of 5-fold cross-validation (5-CV) as the fitness function on the whole training sets of proteins under consideration of computation cost.

### B. Prediction method ProSec-GOX

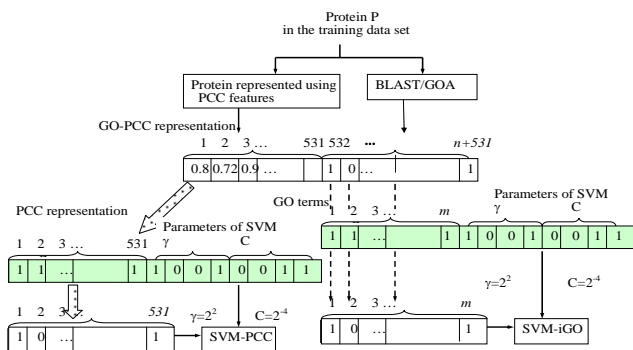


Figure 3 Protein representation and the training model of ProSec-GOX

Figure 3 illustrates the prediction flowchart of ProSec-GOX using SVM classifiers for predicting non-classical secretory proteins. For a query sequence, the BLAST with  $(h, e)=(1, 10^{-9})$  is first performed on the Swiss-Prot database to obtain its homologies with known accession numbers. Subsequently, the obtained accession numbers were used to retrieve the corresponding  $k$  GO terms, GO:1, GO:2, ..., GO:k. The query protein is represented as an  $m$ -dimensional

GO feature vector  $[p_1, p_2, \dots, p_m]$  as an input to the SVM-iGO classifier, where  $m=500$ . The GO terms as the input to the SVM-iGO classifier and the output is 'classical' or 'non-classical' label. If  $k=0$  means the query sequence not annotated by the  $m$  GO terms, the query sequence is represented as a 531-dimensional PCC feature vector and input to the SVM-PCC classifier to predict non-classical secretory proteins.

## IV. RESULTS AND DISCUSSION

### A. Effective GO term features

To evaluate effectiveness of the GO term features used in ProSec-GO, SVM with three additional feature sets: 1) 20 AAC features, 2) 531 PCC feature, 3)  $n$  GO terms, and 4) 500 GO terms (named SVM-AAC, SVM-PCC, SVM-GO, and SVM-iGO, respectively) were individually evaluated in terms of prediction accuracy of 5-fold cross-validation (5-CV) using S25L and S40L. The best values of parameters  $\gamma$  and  $C$  in the SVM-based classifiers were determined using a step-wise approach from  $\gamma \in \{2^{-7}, 2^{-6}, \dots, 2^8\}$  and  $C \in \{2^{-7}, 2^{-6}, \dots, 2^8\}$ . The best classifier is ProSec-GO, yielding training accuracies of 94.1% and 98.9% on S25L and S40L, respectively (Table III). For the training data set S25L, SVM-iGO performs better than that of SVM-ACC (82.5%) and SVM-PCC (84.8%) features. So does for the training data set S40L. Additionally, the method SVM-GO using all GO terms as features obtains the same accuracy as that of SVM-iGO, which reveals this feature selection mechanism can not effectively extract the best feature subset to enhance prediction performance. Therefore, further work aims to find the best feature subset for designing prediction method.

TABLE III. PERFORMANCE COMPARISON USES PREDICTION ACCURACY (%) OF 5-CV

Method	No. Of features	S25L		S40L	
		Accuracy ( $C, \gamma$ )		Accuracy ( $C, \gamma$ )	
SVM-AAC	AAC (20)	82.5	$(2^1, 2^{-5})$	83.2	$(2^1, 2^{-5})$
SVM-PCC	PCC (531)	84.8	$(2^{-7}, 2^{-7})$	86.1	$(2^{-5}, 2^{-5})$
SVM-GO	GO terms (all)	94.1	$(2^2, 2^{-7})$	98.9	$(2^0, 2^{-5})$
SVM-iGO	GO terms (500)	94.0	$(2^1, 2^{-7})$	98.9	$(2^3, 2^{-5})$
ProSec-GOX	GO terms (500) PCC (531)	94.1	$(2^1, 2^{-7})$ $(2^{-7}, 2^{-7})$	98.9	$(2^3, 2^{-5})$ $(2^{-5}, 2^{-5})$

### B. Performance of ProSec-GOX

TABLE IV. ACCURACIES (%) AND MCC PERFORMED ON S25 AND S40

Class	5-CV S25L	Independent test S25T	5-CV S40L	Independent test S40T
Non-classical	88.1 (0.919)	90.3 (0.933)	99.8 (0.983)	98.3 (0.867)
Classical	100 (0.919)	100 (0.933)	98.0 (0.983)	97.6 (0.867)
Overall accuracy % (MCC)	94.1 (0.919)	95.1 (0.933)	98.9 (0.983)	96.8 (0.867)

The Matthew correlation coefficient (MCC) [26] is typically employed to evaluate the performance on unbalanced data sets. Table IV shows detailed results for individual classes that consist of MCC and the accuracies when applied to S25 and S40. The MCC performances of ProSec-GO are 0.919 and 0.933 for S25L and S25T,



respectively, and the corresponding overall accuracies are 94.1% and 95.1%. Additionally, the training and test accuracies for classical secretory proteins are all true predictions.

TABLE V. PERFORMANCE COMPARISON TO OTHER EXISTING METHODS ON THE DATA SET S40

Method	No. of features	Independent Test	
		Accuracy (%)	MCC
Na Bayes*	50	77.8	(0.264)
IBK*	50	80.9	(0.234)
SPRED*	50	82.2	(0.504)
ProSec-GOX	1031	96.8	(0.867)

The ProSec-GO is applied to the whole data set S40 to compare it with existing prediction methods. Table V presents the results of the performance comparison in terms of MCC and the independent test accuracy. The Naïve Bayes classifier [27] and IBK algorithm [28] using the same features as those of SPRED [5] only have accuracies 79.8% (MCC=0.264) and 80.9% (MCC=0.234), respectively. ProSec-GOX has the highest accuracy of 96.8% and MCC=0.867, which is better than 77.8% of Na Bayes, 80.9% of IBK and 82.2% of SPRED, using the top 50 of the frequency of tri-peptides and peptides, secondary structure, physicochemical properties features [5]. The results reveal that the GO terms are effective features for predicting non-classical secretory proteins.

#### V. CONCLUSIONS

Several of eukaryotic secretory proteins lacking the signal peptides are found to be traversed by a non-classical secretion pathway. Therefore, predicting non-classical secretory proteins regardless of signal peptides is necessary for developing a critical computational approach. The ProSec-iGOX yields test accuracies of 95.1% and 96.8% when adopting on the data sets S25 and S40 respectively. However, the method SVM-GO using all GO terms as features obtains the same accuracies as those of SVM-iGO using 500 top-rank frequency descriptors of GO terms, indicating this feature selection mechanism can not effectively extract the best feature subset to improve prediction performance. Therefore, further work is to find the best feature subset for designing prediction method.

#### ACKNOWLEDGMENT

The authors would like to thank the National Science Council of Taiwan for financially supporting this research under the contract Numbers NSC 99-2221-E-243 -005.

#### REFERENCES

[1] I. Prudovsky, A. Mandinova, R. Soldi et al., "The non-classical export routes: FGF1 and IL-1{alpha} point the way," *Journal of Cell Science*, vol. 116, pp. 4871-4881, 2003.

[2] J. D. Bendtsen, L. J. Jensen, N. Blom et al., "Feature-based prediction of non-classical and leaderless protein secretion," *Protein Engineering Design and Selection*, vol. 17, pp. 349-356, 2004.

[3] G. Blobel, "Protein Targeting (Nobel Lecture)," *ChemBioChem*, vol. 1, pp. 86-102, 2000.

[4] W. Nickel, "The mystery of nonclassical protein secretion," *European Journal of Biochemistry*, vol. 270, pp. 2109-2119, 2003.

[5] K. K. Kandaswamy, G. Pugalenti, E. Hartmann et al., "SPRED: A machine learning approach for the identification of classical and non-

classical secretory proteins in mammalian genomes," *Biochemical and Biophysical Research Communications*, vol. 391, pp. 1306-1311, 2010.

[6] J. D. Bendtsen, H. Nielsen, G. von Heijne et al., "Improved Prediction of Signal Peptides: SignalP 3.0," *Journal of Molecular Biology*, vol. 340, pp. 783-795, 2004.

[7] A. Garg, and G. P. S. Raghava, "A Machine Learning Based Method for the Prediction of Secretory Proteins Using Amino Acid Composition, Their Order and Similarity-Search," *In Silico Biology*, vol. 8, pp. 129-140, 2008.

[8] C. H. Hung, H. L. Huang, K. T. Hsu et al., "Prediction of non-classical secreted proteins using informative physicochemical properties" *Interdisciplinary Sciences: Computational Life Sciences* vol. 2, pp. 263-270, 2010.

[9] T. Blum, S. Briesemeister, and O. Kohlbacher, "MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction," *BMC Bioinformatics*, vol. 10, 2009.

[10] W. L. Huang, C. W. Tung, S. W. Ho et al., "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," *BMC Bioinformatics*, vol. 9, pp. 80, 2008.

[11] K. C. Chou, and H. B. Shen, "A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0," *PLoS ONE*, vol. 5, 2010.

[12] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet.*, pp. 25-29, 2000.

[13] W. M. Dahdul, J. P. Balhoff, J. Engeman et al., "Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology Literature," *PLoS ONE*, vol. 5, pp. e10708, 2010.

[14] S. Srivastava, L. Zhang, R. Jin et al., "A Novel Method Incorporating Gene Ontology Information for Unsupervised Clustering and Feature Selection," *PLoS ONE*, vol. 3, pp. e3860, 2008.

[15] A. Bairoch, and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, pp. 45-48, 2000.

[16] Y. Huang, B. Niu, Y. Gao et al., "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, 2010.

[17] S. Carbon, A. Ireland, C. J. Mungall et al., "AmiGO: online access to ontology and annotation data," *Bioinformatics*, vol. 25, pp. 288-289, 2009.

[18] R. Apweiler, A. Bairoch, C. H. Wu et al., "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res.*, vol. 32, pp. D115-D119, 2004.

[19] "GOA. [[ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/](http://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/)]"

[20] S. F. Altschul, W. Gish, W. Miller et al., "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403-410, 1990.

[21] S. F. Altschul, T. L. Madden, A. A. Schaffer et al., "Gapped BLAST and PSIBLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389-3402, 1997.

[22] Z. Lei, and Y. Dai, "Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction," *BMC Bioinformatics*, vol. 7, pp. 491, 2006.

[23] W. L. Huang, C. W. Tung, H. L. Huang et al., "Predicting protein subnuclear localization using GO-amino-acid composition features," *Biosystems*, vol. 98, pp. 73-79, 2009.

[24] S. Kawashima, and M. Kanehisa, "AAindex: Amino Acid index database," *Nucleic Acids Research*, vol. 28, pp. 374-374, 2000.

[25] M. Stone, "Cross-validators: choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, vol. 36, pp. 111-147, 1974.

[26] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta.*, vol. 405, pp. 442-451, 1975.

[27] H. J. George, and P. Langley, "Estimating continuous distributions in bayesian classifiers," *Eleventh Conference. Uncertainty Artif. Intell. San Mateo*, pp. 338-345, 1995.

[28] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37-66, 1991.

## A Network-based Approach for Hospital Capacity Management in a Pandemic

*Jiang Zhang*

*School of Business, Adelphi University, Garden City, NY 11530  
zhang@adelphi.edu*

*Lihui Bai*

*Department of Industrial Engineering, University of Louisville, Louisville, KY  
40292  
lihui.bai@louisville.edu*

The recent outbreaks of H1N1 influenza have shown that the demand for intensive care unit resources and ventilators can overtake the hospital's capacities. Indeed, during a pandemic, the drastic surge in patient volume will cause hospitals to operate beyond their capacities in many other resources. Thus, managing available resources efficiently becomes critical. In the literature of healthcare capacity management, most assume that patients will go to their assigned hospitals. However, we consider that during a pandemic, patients will go to their choice of hospitals (e.g., nearest hospitals) on their own. Consequently, the surge of patient volume will be greater in hospitals of some (e.g., populated) areas than in those of other (e.g. remote) areas. This project proposes an incentive-based approach to help direct patients to alternative hospitals so that capacity shortages across all hospitals are balanced. In other words, under this approach the hospital resources for the community as a whole are utilized most efficiently. Examples of incentives include offering financial discount for services and offering on-site pharmacy within a hospital.

Two optimization models are used in the study. One is an equilibrium model for describing the patients' behavior that everyone maximizes their utility when selecting a hospital. The other is a nonlinear optimization model for the health authority to maximize the total utility of all patients as a whole. Usually, the optimal solutions to the two models are different, because of patients' non-cooperative "selfish" decisions. However, in the presence of incentive programs at certain hospitals, patients change their choice of hospitals. The goal is to select appropriate hospitals to offer the incentive programs, so that the patients' "new" choice of hospital matches the one desired by the health authority, which utilizes the system resources most efficiently.

Keywords: Hospital Capacity Management, Pandemic Emergency, Network Optimization

## Dynamical Inactivation/Enhancement of the Catalytic Machinery of the SARS 3C-like Protease by its Evolutionarily Acquired Extra-domain

*Jiahai Shi and Jianxing Song*

*National University of Singapore, Singapore*  
[shijh@nus.edu.sg](mailto:shijh@nus.edu.sg), [bchsj@nus.edu.sg](mailto:bchsj@nus.edu.sg)

Severe acute respiratory syndrome (SARS) is the first emerging infectious disease of the 21st century which has not only caused rapid infection and death, but also triggered a dramatic social crisis. Its 3C-like protease plays a vital role in processing two viral polyproteins and thus represents a top target for drug design. Intriguingly, the SARS protease evolutionarily acquired a C-terminal extra domain in addition to the chymotrypsin fold sufficient to host the complete catalytic machinery of the 3C protease such as from picorovirus [1]. The functional role of this extra domain had been previously unknown but shortly after the SARS outbreak, we revealed that it plays a key role in mediating the dimerization essential for catalysis [2]. By determining the high-resolution structure of an inactive and monomeric R298A mutant, we further established the mechanism how the extra domain controls the dimerization which can be eventually coupled to the catalytic machinery [3]. On the other hand, we also identified several other mutants on the extra domain which have no significant alteration of the dimerization properties but their activities are either significantly attenuated (N214A) or enhanced (STI/A and STIF/A). Surprisingly their crystal structures we just determined are almost identical to that of the wild-type, thus strongly implying that the enzyme dynamics are extremely critical to the catalysis. Therefore, we launched Molecular Dynamics (MD) simulations for WT, R298, N214, STI and STIF mutants, as well as several artificial monomers. The results show that different proteases display distinctive dynamical behaviors. While in WT, the catalytic machinery stably retains in the activated state, in R298A it remains largely collapsed in the inactivated state, implying that two states are not only structurally very distinguishable [4], but also dynamically well separated. Unbelievably, in N214A the catalytic dyad becomes dynamically unstable and many residues constituting the catalytic machinery jump to sample the conformations highly resembling those of R298A. Therefore, the N214A mutation appears to trigger the dramatic change of the enzyme dynamics in the context of the dimeric form which ultimately inactivates the catalytic machinery [4]. Our MD simulations represent the

longest reported so far for the SARS-CoV 3CLpro, unveiling that its catalysis is critically dependent on the dynamics, which can be amazingly modulated by the extra domain. Consequently, mediating the dynamics may offer a potential avenue to inhibit the SARS-CoV 3CLpro.

Keywords: SARS, 3C-like protease, X-ray crystallography, Molecular dynamics simulation

#### References

1. Shi J, Wei Z, Song J (2004) *J Biol Chem.* 279, 24765-73
2. Shi J and Song J (2006) *FEBS J.* 273, 1035-1045
3. Shi J., Sivaraman J. and Song J. (2008) *J. Virol.* 82, 4620-4629.
4. Shi J, Han N, Lim L, Lua S, Sivaraman J. Wang L, Mu Y and Song J. (2011) *PLoS Computational Biology.* In press.

## Genome-wide association studies with a mother/father-child paired design

Guicheng Zhang\*, Jack Goldblatt, Peter N LeSouëf

School of Paediatrics and Child Health, Faculty of Medicine and Dentistry, University of Western Australia, Perth, Australia

\*: Supported by NHMRC; gczhang@meddent.uwa.edu.au

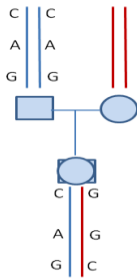
**Abstract.** We propose an analytic strategy to interpret GWAS data with a single parent-child design. The proposed analytic strategy has several advantages, including replicating the GWAS findings in the same population by investigating both the within and between family components of linkage and association, avoiding the confounding effects of population admixture and dissecting parent-of-origin allelic effects of genes on complex traits. With the strategy we propose, more meaningful interpretations of genotype data from a mother/father-child paired study would be expected. Genome-wide association studies with a mother/father-child paired design may hold the key to dissecting the aetiology of complex traits.

A recent publication in *Science* [1] uncovered parent-of-origin allelic effects of many loci in murine brain tissue. In contrast to fewer than 100 imprinted genes hitherto identified, this study suggested that gene imprinting was a common phenomenon underlying the mechanism of physiological function of many genes. Developments in bioinformatics and technological advances in genome-wide analyses have facilitated the analysis of large volumes of data generated by genotyping large sets of single nucleotide polymorphisms (SNPs). It is timely to investigate the parent-of-origin effects of imprinted genes on human disease using the genome-wide association study (GWAS) data.

GWASs are gaining popularity in genetic analyses of complex traits with a commonly used case-control design which recruits unrelated individuals. Although many promising results are being reported there remain many challenges in interpreting the findings of GWA studies [2-3]. The massive multiple testing inherent with GWAS is resulting in these studies having progressively larger sample sizes to try and restrict the false positive rates. However, in studies with a large sample size the modest genetic effects are vulnerable to subtle differences in ancestry between cases and controls [4]. Attempting replication by using still larger sample sizes may replicate imperfect matching of cases and controls and duplicate the false positive finding [4]. A family based study design is robust to population stratification and has been employed for GWA studies with corresponding analytical statistical methods [5]. A family based design allows both linkage and association analyses, investigates inter-familial as well as intra-familial information and has the potential of investigating parent-of-origin allelic effects. However, most published family based studies with GWAS data were limited to replicating or screening genetic variants based on an individual variant or haplotype [6]. Although they tried to maximally utilize the within and between information for the analysis the available statistical analytic programs have not sufficiently utilised the information derived from millions of SNP genotyped in family members [7-9] except a recent attempt in which a isolated population (Icelanders) was investigated for the allele parental origin using GWA data [10]. We believe that implementation of this strategy in GWA studies to determine the allele parental origin would facilitate better understanding of genotype-phenotype relationships.

In a paired single parent-child design with millions of SNPs genotyped it should be possible to ascertain the parental allele origin of a child using a mathematical algorithm, similar to the above [10]. The ability to directly allocate the parent of origin allele becomes easier as the individual

chromosome data become denser and more detailed through linkage maps. With GWAS data, mathematical approaches have been described to inferring membership [11] and to accurately and robustly determining whether individuals are in a complex genomic DNA mixture [12]. An ethical concern has arisen in revealing individual-level information in GWAS [13]. Using a Bayesian inference method, a flexible statistical model has been established to infer missing genotype and haplotype phase for large scale population genotype data [14]. If we infer the haplotypes in parents and their children separately using the information on millions of genotype data, compare the genotypes and haplotype structure between the parent and his/her child, and also incorporate the information on physical distance and recombination rates between SNPs, it is expected we can determine the allele origin of the child for most alleles of GWA data with a mother/father-child paired design. For example, if a mother-child pair has the genotype of AA and Aa for the mother and child, respectively, we know that the child's 'A' allele is from mother. We can also infer the haplotype including the maternal 'A' allele in the child is from mother. As shown in Figure 1, based on the genotype information we know that the CAG haplotype is from the father and GGC from the mother, thus resolving the parental origin of alleles linked with these haplotypes. A mathematical algorithm should be developed to infer the haplotype in a sliding window manner, compare the haplotypes and genotypes between the mother/father-child pair, and then to discern the parental origins of the child's alleles. In most situations the probability of the parental origin of the child's alleles would be determined with a likelihood close to one for mother/father-child paired GWA data.

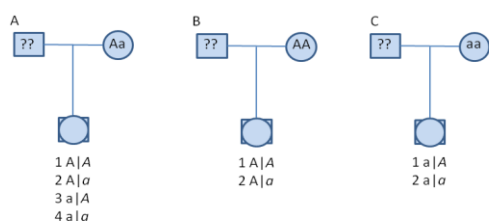


**Figure 1: The haplotype transmission between the father and his child.** In the three loci, the father has the genotypes of CC, AA and GG. His child has the genotypes of CG, AG and GC. Therefore we know the CAG haplotype of the child is from her/his father and the GGC haplotype, from her/his mother.

With the parental origin of the child's alleles resolved for mother/father-child paired GWA data, the imprinting effect of genes or alleles can be investigated in the relation to complex traits [15-16]. A traditional, straightforward, analysis strategy: Transmission Disequilibrium Test (TDT) can also be used for linkage and association analyses with a complex trait of interest. In the family based genetic study, the TDT is the basis to examine linkage and association between a marker locus and disease-susceptibility locus [17]. The test compares the transmission pattern of alleles from heterozygous parents to their affected offspring. When the allele origin of the child is known, using a similar concept to the TDT, we can simply test linkage and association with complex traits for an allele from either parent. To explain this concept we have selected mother-child paired data as an example. In Figure 2 A, the mother has a genotype of Aa and if the affected child has the genotype A|A (letter in italics denotes the allele from father) or A|a, the maternal 'A', and not 'a', allele has been transmitted to the affected child. If the affected child has the genotype of a|A or a|a, the maternal 'a', and not the 'A', allele has been transmitted to the affected child. The number of 'A' or 'a' alleles transmitted to the affected child can be counted and tested to measure which maternal allele is over-transmitted. This would also be used for the mother/father paired with an unaffected child to compare the parental allele transmission patterns



in unaffected children (controls). In this manner, the within-family component can be examined by using a similar method to TDT which is inherently robust in considering population stratification. Analysis of the information from family B and C in Figure 2 can contribute to between family and total association tests for mother/father-child paired GWA data. In this example we do not know the paternal genotypes; however, based on the genotype frequencies in the general population, we can also test the linkage and association for the paternal allele with a complex trait of interest. Mathematically the maternal and paternal effects can be combined together with the overall effects of the allele reported. Although this example is for a binary outcome for a single SNP, the described method could easily be extended to quantitative traits and other genetic variants [9, 17-18].



**Figure 2: Allele transmission patterns in three mother-child pairs.** In Pair A with the mother's genotype of Aa, her child has the possibility of having A|A (letter in italics denotes the allele from father), A|a, a|A and a|a. In Pair B with the mother's genotype of AA, her child has the possibility of having A|A and A|a. In Pair C with the mother's genotype of aa, her child has the possibility of having a|A and a|a.

In summary, we propose an analytic strategy to interpret GWAS data with a single parent-child design. A major obstacle is to accurately clarify the parent-of-origin of the child's alleles. An attempt has been published for GWA data in an isolated population with family relationships between individuals and the parental origin of most alleles can be determined. Therefore, we propose a mathematical strategy could be developed to resolve this for mother/father-child paired data with millions of SNPs genotyped. This analytic strategy has several advantages, including replicating the GWAS findings in the same population by investigating both the within and between family components of linkage and association, avoiding the confounding effects of population admixture and dissecting parent-of-origin allelic effects of genes on complex traits. Compared to recruiting cohorts of unrelated subjects, it may be more difficult to recruit family members for these epidemiological studies. However, an increasing number of genetic studies are using mother-child designs [19-20]. With the strategy we propose, more meaningful interpretations of genotype data from a mother/father-child paired study would be expected. Genome-wide association studies with a mother/father-child paired design may hold the key to dissecting the aetiology of complex traits.

### References

1. Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G.P., Haig, D., Dulac, C.: High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* 329, 643-648 (2010)
2. Zhang, G., Goldblatt, J., LeSouef, P.: The era of genome-wide association studies: opportunities and challenges for asthma genetics. *J Hum Genet* 54, 624-628 (2009)
3. Zhang, G., Goldblatt, J., Lesouef, P.: Findings in genome-wide association studies on asthma lack generalisation. *Clin Respir J* 4, e8-9 (2010)

4. McClellan, J., King, M.C.: Genetic heterogeneity in human disease. *Cell* 141, 210-217 (2010)
5. Rabinowitz, D., Laird, N.: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 50, 211-223 (2000)
6. Benyamin, B., Visscher, P.M., McRae, A.F.: Family-based genome-wide association studies. *Pharmacogenomics* 10, 181-190 (2009)
7. Lasky-Su, J., Won, S., Mick, E., Anney, R.J., Franke, B., Neale, B., Biederman, J., Smalley, S.L., Loo, S.K., Todorov, A., Faraone, S.V., Weiss, S.T., Lange, C.: On genome-wide association studies for family-based designs: an integrative analysis approach combining ascertained family samples with unselected controls. *Am J Hum Genet* 86, 573-580 (2010)
8. Macgregor, S.: Optimal two-stage testing for family-based genome-wide association studies. *Am J Hum Genet* 82, 797-799; author reply 799-800 (2008)
9. Chen, M.H., Larson, M.G., Hsu, Y.H., Peloso, G.M., Guo, C.Y., Fox, C.S., Atwood, L.D., Yang, Q.: A three-stage approach for genome-wide association studies with family data for quantitative traits. *BMC Genet* 11, 40 (2010)
10. Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K.T., Frigge, M.L., Gylfason, A., Olason, P.I., Gudjonsson, S.A., Sverrisson, S., Stacey, S.N., Sigurgeirsson, B., Benediktsdottir, K.R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J.H., Johannsson, O.T., Hreidarsson, A.B., Sigurdsson, G., Ferguson-Smith, A.C., Gudbjartsson, D.F., Thorsteinsdottir, U., Stefansson, K.: Parental origin of sequence variants associated with complex diseases. *Nature* 462, 868-874 (2009)
11. Manichaikul, A., Mychaleckyj, J., Rich, S., Daly, K., Sale, M., Chen, W.-M.: Robust Relationship Inference in Genome Wide Association Studies. Oxford University Press (2010)
12. Homer, N., Szlinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4, e1000167 (2008)
13. Lumley, T., Rice, K.: Potential for revealing individual-level information in genome-wide association studies. *JAMA* 303, 659-660 (2010)
14. Scheet, P., Stephens, M.: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78, 629-644 (2006)
15. Weinberg, C.R., Wilcox, A.J., Lie, R.T.: A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 62, 969-978 (1998)
16. Whittaker, J.C., Gharani, N., Hindmarsh, P., McCarthy, M.I.: Estimation and testing of parent-of-origin effects for quantitative traits. *Am J Hum Genet* 72, 1035-1039 (2003)
17. Laird, N.M., Lange, C.: Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7, 385-394 (2006)
18. Lazzeroni, L.C., Lange, K.: A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* 48, 67-81 (1998)
19. Shi, M., Umbach, D.M., Vermeulen, S.H., Weinberg, C.R.: Making the most of case-mother/control-mother studies. *Am J Epidemiol* 168, 541-547 (2008)
20. Wang, S., Zheng, T., Chanock, S., Jedrychowski, W., Perera, F.P.: Methods for detecting interactions between genetic polymorphisms and prenatal environment exposure with a mother-child design. *Genet Epidemiol* 34, 125-132 (2010)



# Predicting multiplex subcellular location of proteins using protein-protein network: A comparative study

Jonathan Q. Jiang

Department of Computer Science, City University of Hong Kong  
 Tat Chee Avenue, Kowloon, Hong Kong  
 qiajiang@cityu.edu.hk

## 1 Methods

Recent investigations, both experimental [3] and computational [4], show that physical interaction seems an important hint for co-localization of proteins. This discovery provides us new opportunities to reveal protein subcellular locations in the context of PPI network. Unfortunately, so far, no systematic efforts have been made in this regard except for a few investigations focused on the "mono-location" case only. Actually, proteins may often simultaneously exist at, or move between two or more different subcellular compartments.

These motivate us to design a comparative study for associating proteins with multiple locations based on PPI network. Two local methods, Majority [3] and  $\chi^2$ -score [1], and two global methods, GenMultiCut(GMC) [8] and FunFlow [2], originally proposed for protein function prediction, are exploited since these two topics belong to the same type of problem, i.e., classifying nodes in a partially labeled network. We compiled a *Saccharomyces cerevisiae* PPI network, consisting of 3179 proteins with 12413 interactions, from BioGRID database (version 3.1.73, released 25-Jan-2011)[5], and extracted the 22 experimentally observed locations for each proteins from Yeast Gtp Fusion Localization Database[9]. With these sources, we systematically analyze the four algorithms by performing a large-scale cross validation on this PPI network and comparing their predictions. Furthermore, we build an ensemble classifier based on these four approaches and give assignments to 529 unlabeled and 137 ambiguous annotated proteins with multiplex subcellular locations. Among these predictions, most of them have been previously characterized in UniPort [7] database.

## 2 Results and discussions

### 2.1 5-fold cross validation

We test the performance of these four algorithms using 5-fold cross validation on the obtained PPI network. To deal with the *partially correct* problem, for the first time, we adopt Average Precision (AP)[6] to evaluate and compare the approaches on each subcellular location. The mean average precision (MAP)

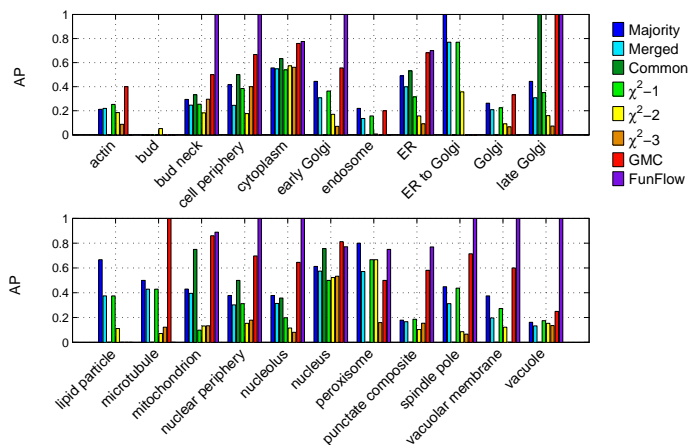
is the overall evaluation, which for these approaches are summarized in Table 1. From the table, we see clearly that the global methods, GMC and FunFlow significantly outperform the local counterparts, Majority and  $\chi^2$ -score.

**Table 1.** MAP of four algorithms for 5-fold cross validation.

Algorithms	MAP (%)	Algorithms	MAP (%)	Algorithms	MAP (%)	Algorithms	MAP (%)
Majority	42.13	Common	24.36	$\chi^2-2$	19.77	GMC	53.43
Merged <sup>1</sup>	32.53	$\chi^2-1^2$	33.07	$\chi^2-3$	14.59	FunFlow	62.07

<sup>1</sup> Merged and Common, two variants of Majority were proposed in [4]. We also list them here for comparison.

<sup>2</sup>  $\chi^2-k$  denotes the  $\chi^2$ -score method with radius  $k$ .

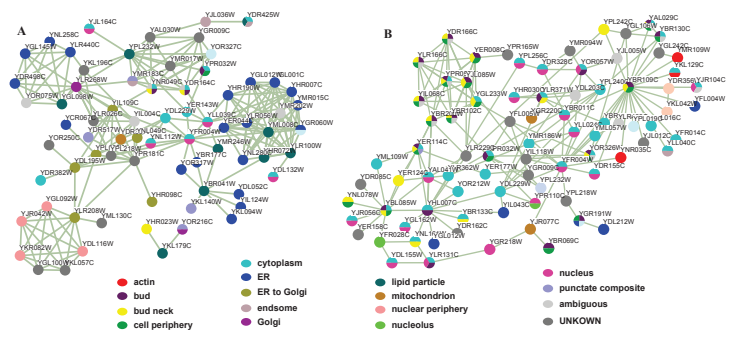


**Fig. 1.** AP for each subcellular location.

We further check the AP for each subcellular location (Figure 1). All these methods achieve a competitive performance for two subcellular locations "cytoplasm" and "nucleus" which there are a large number of proteins experimentally annotated with. For another 11 locations, i.e., "Bud neck", "cell periphery", "Early Golgi", "Late Golgi", "Microtubule", "Mitochondrion", "Nuclear periphery", "Punctate composite", "Spindle pole", "Vacuolar membrane" and "Vacuole", two global methods always, sometimes significantly, outperform two local approaches. The superior performance of global methods is expected since the GMC algorithm takes the full structure of the PPI network into account, and FunFlow considers both the global and local effects. Moreover, we are surprised to find that two local methods as well as their variants achieved better performance for two locations, "ER to Golgi" and "lipid particle" which are involved

in protein transport and secretion. Finally, it is astonishing that almost all the methods cannot successfully recover the "Bud" location for proteins except for the  $\chi^2 - 2$  algorithm with a very low AP value. We design case studies to further analyze these two unexpected phenomena in the following section.

2.2 Case studies



**Fig. 2.** Case studies: (A) The subnetwork consists of the interactions among proteins annotated with locations "ER to Golgi" and "Lipid particle" and their immediate neighbors. (B) The subnetwork contains the proteins labeled with "Bud" location and its immediate neighbors as well as their interactions.

We extracted the proteins annotated with locations "ER to Golgi" and "Lipid particle" as well as their immediate neighbors and the physical interactions among them from our PPI network. The subnetwork, containing 72 unique proteins and 204 unique interactions, is illustrated in Figure 2A. Clearly, the 6 proteins experimentally annotated with "ER to Golgi" location scatter in the subnetwork to bridge the gap between two protein cliques that localized in "endoplasmic reticulum" and "Nuclear periphery". For example, the protein YLR208W, 4 "Nuclear periphery" proteins and 2 "Unknown" proteins are joined together with it in a tightly-knit fashion (the lower left corner of Figure 2A). Obviously, it is misclassified into "Nuclear periphery" group if the global methods are applied. By contrast, if we adopt the local method, the "ER to Golgi" location is one of the two subcellular locations that are common among its neighbors. Similar phenomenon can be found for locations "Lipid particle" and "Bud" (Figure 2B). From the above analysis, we assert that the superiority of the local algorithms for these two locations is totally due to the neighborhood topology of these proteins annotated with corresponding locations.

2.3 Assign subcellular locations to uncharacterized proteins

Considering that the local methods and global methods have respective advantages and disadvantages, we build an ensemble classifier to assign subcellular

4 J.Q. Jiang

locations to 527 unlabeled and 139 ambiguous annotated proteins in our PPI network. Most of them were previously validated in Uniprot database. Only the first 20 predictions are listed in Table 2 due to the page limits.

**Table 2.** Predictions of subcellular locations for 666 uncharacterized proteins in our PPI network.

Protein (ORF)	Prediction	UniProt
Q0045	mitochondrion	Mitochondrion inner membrane; Multi-pass membrane protein.
Q0080	mitochondrion	Mitochondrion membrane; Single-pass membrane protein.
Q0085	mitochondrion	Mitochondrion inner membrane; Multi-pass membrane protein.
Q0105	mitochondrion	Mitochondrion inner membrane; Multi-pass membrane protein.
Q0120		Mitochondrion.
Q0130	mitochondrion	Mitochondrion membrane; Multi-pass membrane protein.(Potential)
Q0250	mitochondrion	Mitochondrion inner membrane; Multi-pass membrane protein.
Q0275	mitochondrion	Mitochondrion inner membrane; Multi-pass membrane protein.(By similarity)
YAL003W	cytoplasm	
YAL013W	nucleus	Cytoplasm. Nucleus.
YAL013W	nucleus	Cytoplasm. Nucleus.
YAL020C	cytoplasm	
YAL028W	cytoplasm;nucleus	Endoplasmic reticulum membrane.
YAL029C	bud	Bud.
YAL030W	lipid particle	Endomembrane system.
YAL034C	nucleus	
YAL040C	cytoplasm	
YAL042W	ER	Endoplasmic reticulum membrane. Golgi apparatus membrane.
YAL062W	actin;cytoplasm	
YAR018C	spindle pole	

## References

1. Hishigaki H., Nakai K., Ono T., Tanigami A., Takagi T.: Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* 18, 523–531 (2001)
2. Nabieva E., Jim K., Agarwal A., Chazelle B., Singh M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21 (Suppl 1), i302–i310 (2005)
3. Schwikowski, B., Uetz, P., Field, S.: A network of protein- protein interactions in yeast. *Nature Biotechnology* 18, 1257–1261 (2000).
4. Shin, C.J., Wong, S., Davis, M.J., Ragan, M.A.: Protein-protein interaction as a predictor of subcellular location. *BMC Syst Biol.* 3, 28 (2009).
5. Biological General Repository for Interaction Datasets. <http://thebiogrid.org/>
6. TRECVID. <http://www-nlpir.nist.gov/projects/trecvid/>
7. UniProt Database. <http://www.uniprot.org/>
8. Vazquez A., Flammini A., Maritan A., Vespignani A.: Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* 21, 697–700 (2003)
9. Yeast Gfp Fusion Localization Database. <http://yeastgfp.yeastgenome.org/>

# Support Vector Machine for Classification of DNA repair genes

Hao Jiang and Wai-Ki Ching

Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong  
haohao@hkusuc.hku.hk wching@hku.hk

**Abstract.** Human longevity is a complex phenotype that has a significant genetic predisposition. Intensive research has been carried out to elucidate the role of DNA repair systems in the ageing process. Decision trees and naive Bayesian algorithm are two data-mining based classification methods for systematically analyzing data about human DNA repair genes. In this paper we develop a linearly combined kernel with Support Vector Machine (SVM) to analyze the ageing related data. The popular supervised learning algorithm enables better discrimination between ageing-related and non-ageing-related DNA repair genes. Through training on the whole data set, we can identify the same important genes that target essential pathways as well. What's more, novel genes are detected which may reveal possible insights for biologists in ageing research.

**Key words:** SVM; Kernel Combination; Classification

## 1 Introduction

Ageing research has benefited a lot from the application of genetics in the past decades. It has been argued that regulatory genes which affect multiple pathways and processes are most likely to have significant effects on longevity [1]. DNA repair genes could be designated as a major type of genes associating with ageing process [2, 3]. To date, over 150 DNA repair genes have been identified [4, 5].

However, conceptual approaches have not quite caught up with the technology. This creates an opportunity for the application of bioinformatics approaches in ageing research. Decision tree learning and naive Bayesian algorithms stand as the first application of data-mining based methods for the analysis of DNA repair genes [6]. Two types of ageing related datasets are used to testify the robustness of the above two algorithms. One type of datasets includes only gene expression attributes. The other type of dataset involves multiple types of attributes rather than gene expression attributes. Results show that for the former type of dataset, two algorithms tend to exhibit weak performance in classification, achieving 51.1% and 52.1% AUC values respectively. More robust methods can be introduced to improve the classification accuracy for the data set.

Support Vector Machine (SVM) has successfully applied to many different areas [13]. For example, in [7], physico-chemically weighted kernel was constructed

in conjunction with SVMs for the classification of protein datasets and glycan data set [8]. Recent development of kernel methods emphasized on the need to consider a combination of multiple kernels in real-world applications. An evolutionary approach was proposed for finding the optimal weights of a combined kernel used by SVMs [9]. In this paper, we apply SVM in the classification of the gene expression based ageing data. Using the linear combination of **linear kernel** and **polynomial kernel of degree 3**, better discrimination performance can be achieved. Moreover, not only the significant genes identified can target the well-known pathway involved in ageing, but also, novel genes are detected. This gives potential clues for biologists for the investigation of the specific function of the selected genes.

## 2 Methods

### 2.1 Combination of Kernels

One of the most important steps in SVM classification systems is the construction of appropriate kernel functions. In the case of linearly separable data, linear kernel is the most straightforward choice. Polynomial Kernel is suitable for problems where all the training data are in normalized form [14]. As RBF kernels use the Euclidean distance, they are not robust to outliers.

Real world applications have posed a need for emphasis on the combination of kernels. Here we propose to consider a combination of **linear kernel** and **polynomial kernel of degree 3** in fulfilling the task of classification of the normalized ageing data. In this context, the hyperplane can be presented in the following form:

$$f(\mathbf{x}) = \left( \sum_{i=1}^m \alpha_i y_i [\langle \mathbf{x}, \mathbf{x}_i \rangle + \langle \mathbf{x}, \mathbf{x}_i \rangle^3] + b \right).$$

Here  $\mathbf{x}_i$ ,  $i = 1, 2, \dots, m$  are the support vectors obtained from training and  $\alpha_i$ ,  $i = 1, 2, \dots, m$  are the corresponding coefficients for the support vectors,  $y_i$ ,  $i = 1, 2, \dots, m$  are the corresponding classes they belong to, with  $b$  being the bias part and  $m$  being the number of the training data.

### 2.2 Selection of Important Genes

Important genes are selected through the procedure of training the whole data set and according to the ranking of the decision function value. The selected genes are then compared with the current biological results to see if they can target some essential pathways associated with ageing process.

## 3 Results and Discussions

### 3.1 Classification Results

The effectiveness of our proposed method was evaluated in comparison with J48 and naive Bayesian algorithm in terms of classification accuracy. A set of

DNA repair genes was obtained from [6] with 148 instances in total. We have 33 positive data instances and the remaining are negative. The number of attributes is 108 and 10-fold cross-validation was utilized for training and testing of the data set. Table 1 presents the performance of the SVM classifier for the combined kernel for 10 times of 10-fold cross-validation. The values in the table represent the averaged AUC value. It is clearly that for the ten time 10-fold cross-

**Table 1.** 10-time AUC values by Combination of Kernels

10-time AUC Values									
0.643	0.6422	0.6678	0.6379	0.6555	0.6424	0.6477	0.6643	0.6552	0.6486

validation, the AUC values can reach 65% most of the times. This is a significant improvement when compared to the previous two data-mining approaches: J48 and naive Bayesian algorithm. For the two methods, with the same standard of 10-fold cross-validation, they can only get 51.1% and 52.1% respectively.

**3.2 Selection of Important Genes**

Five Important genes are selected according to their scores ranking and they are : PCNA, PARP1, APEX1, MLH1, XRCC6. Compared to the genes selected by J48 and Naive Bayesian Algorithm, in the significant pathway identified, we have targeted APEX1, XRCC6 as well. Moreover, PCNA is not included by J48 and Naive Bayes Algorithm but is detected by our method. This further validates the robustness of our proposed kernel.

The novel genes not associated in the pathway are PARP1 and MLH1. PARP1 and WRN interact physically and co-operate functionally in preventing carcinogenesis in vivo [10] when the WRN protein is associated with Werner’s syndrome that is the one of most representative characteristics of accelerated ageing [11]. PARP1 has also been shown to link with DNA double-strand break pathway, exhibiting various symptoms of accelerated ageing [12].

As for MLH1, it was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). Alternative splicing results in multiple transcript variants encoding distinct isoforms. Additional transcript variants have been described, but their natures have not been fully determined. This fact would provide some potential clues for the biologists in further investigation of the specific role MLH1 played in ageing process.

**4 Conclusions**

We proposed a linearly combined kernel in SVM classification for DNA repair genes data set. Compared to J48 and naive Bayesian algorithm, not only the AUC value for the classification has been improved to 10-15%, but still, the robust kernel can identify the same genes associated with the important pathways

targeted by the two algorithms. In a further perspective, our method detects other genes like PCNA that plays critical role in the same pathway while the two methods failed to identify. The promising perspective lies in that, we have also detected novel genes associated with ageing while the full natures of which are expecting to be explored. This would give a clue for the biologists in further investigation of the specific roles they played in ageing.

**Acknowledgments.** Research supported in part by GRF Grant No. 7017/07P, HKU Strategy Research Theme fund on Computational Sciences, National Natural Science Foundation of China Grant No. 10971075 and Guangdong Provincial Natural Science Grant No. 9151063101000021.

## References

1. Jazwinski S: The RAS genes: a homeostatic device in *Saccharomyces cerevisiae* longevity, Vol. 20. *Neurobiol. Aging* (1999) 471–478.
2. Arking R: *The Biology of Ageing: Observations and Principles*. Oxford: Oxford University Press (2006).
3. Hasty P, Campisi J, Heijmakers J, van Steeg H, Vijg J: Aging and genome maintenance: lessons from the mouse? Vol. 299. *Science* (2003) 1355–1359.
4. Wood R, Mitchell M, Sgouros J, Lindahl T: Human DNA repair genes. Vol. 291. *Science* (2001) 1284–1289.
5. Wood R, Mitchell M, Lindahl T: Human DNA repair genes. Vol. 577. *Mutation Research* (2005) 275–283.
6. Freitas A, Vasieva O, de Magalhaes JP: A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related. Vol. 12. *BMC Genomics* (2011) 1–11.
7. Jiang H, Ching W: Physico-Chemically Weighted Kernel for SVM protein classification. In *Proceedings of the 2nd International Conference on Biomedical Engineering and Computer Science: 23-24 April 2011; Wuhan, China* (2011) 12–15.
8. Kuboyama T, Hirata K, Aoki-Kinoshita K, Kashima H, Yasuda H: A Gram Distribution Kernel Applied to Glycan Classification and Motif Extraction. Vol. 17. *Genome Informatics* (2006) 25–34.
9. Dios L, Oltean M, Rogozan A, Pecuchet J: Improving SVM Performance Using a Linear Combination of Kernels. Vol. 4432. *Lecture Notes in Computer Science* (2007) 218–227.
10. Lebel M, Lavoie J, Gaudreault I, Bronsard M, Drouin R: Genetic cooperation between the Werner syndrome protein and poly(ADPribose) polymerase-1 in preventing chromatid breaks, complex chromosomal rearrangements, and cancer in mice. Vol. 162. *Am J Pathol* (2003) 1559–69.
11. Hasty P, Vijg J: Accelerating aging by mouse reverse genetics: a rational approach to understanding longevity. Vol. 3. *Aging Cell* (2004).
12. Mvd V, Andressoo J, Holcomb V, Lindern M, Jong W, Zeeuw C, Suh Y, Hasty P, Heijmakers J, GTJvd Horst ea: Adaptive response in segmental progeria resembles long-lived dwarfism and calorie restriction in mice. Vol. 2. *PLoS Genetics* (2006).
13. Schölkopf B: *Kernel Methods in Computational Biology*. New York: MIT Press (2004).
14. Kernel Functions for Machine Learning Applications. [<http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html>].



# Novel Phylogenomic Method for Prokaryotes

*Katerina Korenblat, Zeev Volkovich*

*Software Engineering Department, ORT Braude Academic College, Karmiel,  
Israel*

*[katerina@braude.ac.il](mailto:katerina@braude.ac.il), [vlvolkov@braude.ac.il](mailto:vlvolkov@braude.ac.il)*

*and*

*Alexander Bolshoy*

*Department of Evolutionary and Environmental Biology, University of Haifa,  
Israel*

*[bolshoy@research.haifa.ac.il](mailto:bolshoy@research.haifa.ac.il)*

Here we present a novel phylogenomic method of genome-tree construction on the basis of gene lengths of orthologous genes presented in completely sequenced genomes of prokaryotic organisms using Clusters of Orthologous Groups (COGs). Every single element of our input data is a median protein length related to a pair (COG, genome). In principle, the method is so fast that input data may consist of median protein lengths related to thousands of COGs and hundreds of genomes. Clustering is performed using an application of the information bottleneck method for unsupervised clustering.

Two main strategies in the field of a species tree phylogenomic reconstruction were developed to this end: the supertree and the supermatrix. One group of the supermatrix methods is associated with a Boolean matrix based on the presence and absence of gene families in genomes. Even though the method we present here is closely related to a group of methods based on the presence and absence of genes, it uses the information related to the lengths of genes, and this addition makes a significant difference.

In introducing a novel supermatrix phylogenomic method, we have had several primary goals:

- First, to propose a fast method that allows the use of whole proteome characters for reliable construction of genome trees. We show that the method is fast and reliable, indeed.

- Second, to show the robustness of the proposed algorithm. For robustness evaluation, we applied jackknife technique to input data. The aim of this approach is to show that tree structure based on different subsets of COGs is sufficiently stable. We have conducted extensive experiments to validate the performance of bootstrapping and jackknifing in order to estimate how robust the phylogenies produced by the proposed methodology are. These experiments show that randomization as part of the bootstrap procedure substantially decreases stability of the obtained trees and that jackknifing is very useful to determine the confidence level of a phylogeny.
- Thirdly, to reveal the phylogenetic nature of these trees on the basis of a few empirical case studies. We demonstrate that a selected small group of genomes is distributed reasonably along a produced phylogenetic tree. Although our comprehensive genome clustering is independent of phylogenies based on the level of homology of individual genes, it correlates well with the standard "tree of life" based on sequence similarity of 16s rRNA. This, together with successful jackknifing for the determination of confidence levels signifies that the method may be truly classified as phylogenomic. We have also examined several of the methodological issues involved in going from a large sequence database to a useable phylogeny. In particular, we integrate (semi) automated solutions to rogue taxon identifications and jackknifing measurements of tree stability in a single study to examine the phylogenetic signal contained in large sparse supermatrices.
- On the basis of a few empirical case studies, we intended to fix the parameters of the method. We considered three parameters to choose the most appropriate values of the parameters. (1) A bootstrapping parameter that designates a fraction of randomly resampling columns (COGs) of the input dataset. (2) A jackknifing parameter that designates a fraction of randomly deleted columns. (3) A preprocessing parameter (threshold) to consider only those columns of the supermatrix containing more elements than a certain threshold.

To summarize, we are confident in our proposal to construct prokaryotic phylogenetic trees using the fast and reliable method described in this manuscript with parameter values equal to 15% of the maximal COG size for the preprocessing parameter and equal to 80% for the jackknifing parameter.

Keywords: species tree, information bottleneck approach, phylogenetic signal, robustness of clustering, clusters of orthologous genes

# An exponential algorithm for the Discretizable Molecular Distance Geometry Problem is polynomial on proteins

LEO LIBERTI<sup>1</sup>, CARLILE LAVOR<sup>2</sup>, ANTONIO MUCHERINO<sup>3</sup>

<sup>1</sup> LIX, École Polytechnique, 91128 Palaiseau, France  
liberti@lix.polytechnique.fr

<sup>2</sup> Dept. of Applied Maths (IME-UNICAMP), State Univ. of Campinas, 13081-970,  
Campinas - SP, Brazil clavor@ime.unicamp.br

<sup>3</sup> CERFACS, Toulouse, France antonio.mucherino@cerfacs.fr

**Abstract.** An important application of distance geometry to biochemistry studies the embeddings of the vertices of a weighted graph in the three-dimensional Euclidean space such that the edge weights are equal to the Euclidean distances between corresponding point pairs. When the graph represents the backbone of a protein, one can exploit the natural vertex order to show that the search space for feasible embeddings is discrete. The corresponding decision problem can be solved using a binary tree based search procedure which is exponential in the worst case. We discuss assumptions that bound the search tree width to a polynomial size, and show empirically that they apply to proteins.

**Keywords:** Branch-and-Prune, symmetry, distance geometry.

## 1 Introduction

The MOLECULAR DISTANCE GEOMETRY PROBLEM, which asks to find the embedding in  $\mathbb{R}^3$  of a given weighted undirected graph, is a good model for determining the structure of proteins given a set of inter-atomic distances [2]. Its generalization to  $\mathbb{R}^K$  is called DISTANCE GEOMETRY PROBLEM (DGP). In general, the MDGP and DGP implicitly require a search in a continuous Euclidean space. Proteins, however, have further structural properties that can be exploited to define subclasses of instances of the MDGP and DGP whose solution set is finite [1]. These instances can be solved with an algorithmic framework called Branch-and-Prune (BP) [1]: this is an iterative algorithm where the  $i$ -th atom of the protein can be embedded in  $\mathbb{R}^3$  using distances to at least three preceding atoms. Since the intersection of three 3D spheres contains in general two points, the BP gives rise to a binary search tree. In the worst case, the BP is an exponential time algorithm, which is fitting because the MDGP and DGP are NP-hard [Saxe, 1979]. Compared to continuous search algorithms, the performance of the BP algorithm is impressive from the point of view of both efficiency and reliability. In this paper we show that the BP has a polynomial worst-case under assumptions found in proteins.

## 2 Discretizable instances and the BP algorithm

### Notation

For all integers  $n > 0$ , we let  $[n] = \{1, \dots, n\}$ . Given an undirected graph  $G = (V, E)$  with  $|V| = n$ , for all  $v \in V$  we let  $N(v) = \{u \in V \mid \{u, v\} \in E\}$  be the set of vertices adjacent to  $v$ . Given a positive integer  $K$ , an *embedding* of  $G$  in  $\mathbb{R}^K$  is a function  $x : V \rightarrow \mathbb{R}^K$ . If  $d : E \rightarrow \mathbb{R}_+$  is a given edge weight function on  $G = (V, E, d)$ , an embedding is *valid* for  $G$  if  $\forall \{u, v\} \in E \ \|x_u - x_v\| = d_{uv}$ . For any  $U \subseteq V$ , an embedding of  $G[U]$  (i.e. the subgraph of  $G$  induced by  $U$ ) is a *partial embedding* of  $G$ . If  $x$  is a partial embedding of  $G$  and  $y$  is an embedding of  $G$  such that  $\forall u \in U \ (x_u = y_u)$  then  $y$  is an *extension* of  $x$ . For a total order  $<$  on  $V$  and for each  $v \in V$ , let  $\rho(v) = |\{u \in V \mid u \leq v\}|$  be the *rank* of  $v$  in  $V$  with respect to  $<$ . The rank is a bijection between  $V$  and  $[n]$ , so we can identify  $v$  with its rank and extend arithmetic notation to  $V$  so that for  $i \in \mathbb{Z}$ ,  $v + i$  denotes the vertex  $u \in V$  with  $\rho(u) = \rho(v) + i$ . For all  $v \in V$  and  $\ell < \rho(v)$  we denote by  $\gamma_\ell(v)$  the set of  $\ell$  immediate predecessors of  $v$ . If  $U \subseteq V$  with  $|U| = h$  such that  $G[U]$  is a clique, let  $D'(U)$  be the symmetric matrix whose  $(u, v)$ -th component is  $d_{uv}^2$  for  $u, v \in U$ , and let  $D(U)$  be  $D'(U)$  bordered by a left  $(0, 1, \dots, 1)^T$  column and a top  $(0, 1, \dots, 1)$  row (both of size  $h + 1$ ). Then the Cayley-Menger formula states that the volume in  $\mathbb{R}^{h-1}$  of the  $h$ -simplex defined by  $G[U]$  is given by  $\Delta_{h-1}(U) = \sqrt{\frac{(-1)^h}{2^{h-1}((h-1)!)^2} |D(U)|}$ .

Generalized DISCRETIZABLE MOLECULAR DISTANCE GEOMETRY PROBLEM ( $K$ DMDGP). Given an integer  $K > 0$ , a weighted undirected graph  $G = (V, E, d)$  with  $d : E \rightarrow \mathbb{Q}_+$ , a total order  $<$  on  $V$  and an embedding  $x' : [K] \rightarrow \mathbb{R}^K$  such that:

1.  $x'$  is a valid partial embedding of  $G[[K]]$  (START)
2.  $G$  contains all  $(K + 1)$ -cliques of  $<$ -consecutive vertices as induced subgraphs (DISCRETIZATION)
3.  $\forall v \in V$  with  $v > K$ ,  $\Delta_{K-1}(\gamma_K(v)) > 0$  (STRICT SIMPLEX INEQUALITIES),

is there a valid embedding  $x$  of  $G$  in  $\mathbb{R}^K$  extending  $x'$ ?

We denote by  $X$  the set of embeddings solving a  $K$ DMDGP instance;  $X$  is a finite set [1]. The  $K$ DMDGP is NP-hard by reduction from the DMDGP [1]. For a partial embedding  $x$  of  $G$  and  $\{u, v\} \in E$  let  $S_{uv}^x$  be the sphere centered at  $x_u$  with radius  $d_{uv}$ . The BP algorithm, used for solving the  $K$ DMDGP and its

---

#### Algorithm 1 BP( $v, \bar{x}, X$ )

---

**Require:** A vtx.  $v \in V \setminus [K]$ , a partial emb.  $\bar{x} = (x_1, \dots, x_{v-1})$ , a set  $X$ .

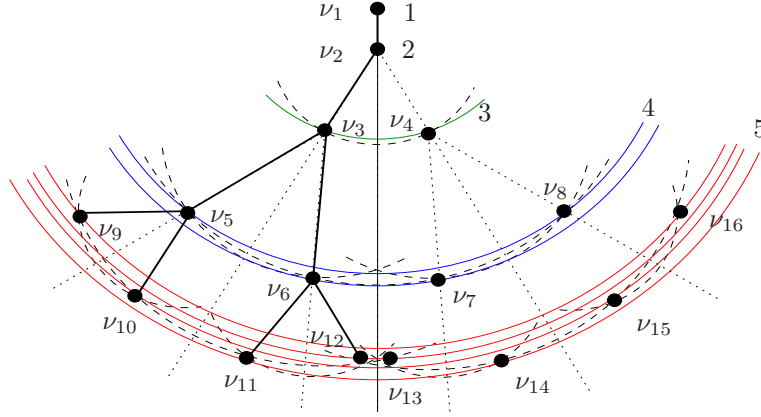
- 1:  $P = \bigcap_{\substack{u \in N(v) \\ u < v}} S_{uv}^{\bar{x}}$ ;
  - 2:  $\forall p \in P \ (x \leftarrow (\bar{x}, p))$ ; **if**  $(\rho(v) = n) \ X \leftarrow X \cup \{x\}$  **else** BP( $v + 1, x, X$ ).
- 

restrictions, is BP( $K + 1, x', \emptyset$ ) (see Alg. 1). By STRICT SIMPLEX INEQUALITIES,  $|P| \leq 2$ . At termination,  $X$  contains all embeddings extending  $x'$  [1].

### 3 BP tree geometry

Since the definition of the  $K$ DMDGP requires  $G$  to have at least those edges used to satisfy the DISCRETIZATION axiom, we partition  $E$  into the sets  $E_D = \{\{u, v\} \mid |\rho(v) - \rho(u)| \leq K\}$  and  $E_P = E \setminus E_D$ . With a slight abuse of notation we call  $E_D$  the *discretization distances* (guaranteeing that a DGP instance is in  $K$ DMDGP) and  $E_P$  the *pruning distances* (used to reduce the search space by pruning the BP tree). Pruning distances might make the set  $P$  in Alg. 1 empty or a singleton.

Let  $G$  be a YES instance of the  $K$ DMDGP,  $G_D = (V, E_D, d)$  and let  $X_D$  be the set of embeddings of  $G_D$ ; since  $G_D$  has no pruning distances, the BP search tree for  $G_D$  is a full binary tree and  $|X_D| = 2^{n-K}$ . The discretization distances arrange the embeddings so that, at level  $\ell$ , there are  $2^{\ell-K}$  possible embeddings  $x_v$  for the vertex  $v$  with rank  $\ell$ . Furthermore, when  $P = \{x_v, x'_v\}$  and the discretization distances to  $v$  only involve the  $K$  immediate predecessors of  $v$ , we have that  $x'_v = R_x^v(x_v)$  [3], the reflection of  $x_v$  w.r.t. the hyperplane through  $x_{v-K}, \dots, x_{v-1}$ . This also implies that the partial embeddings encoded in two BP subtrees rooted at reflected nodes  $\nu, \nu'$  are reflections of each other. This situation is shown in the picture below.

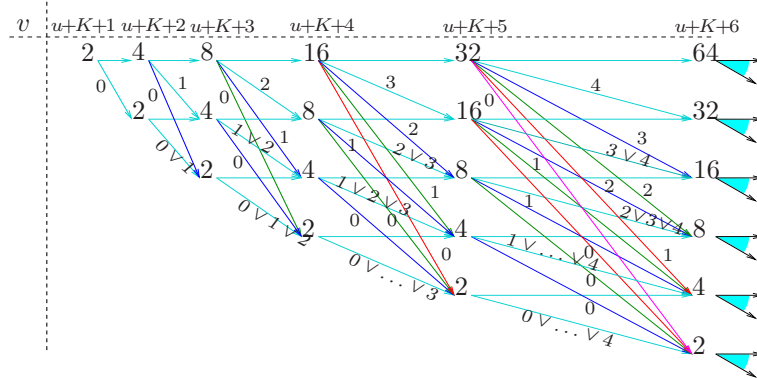


More precisely, with probability 1 we have  $\forall v > K, u < v - K \exists H^{uv} \subseteq \mathbb{R}$  s.t.  $|H^{uv}| = 2^{v-u-K}$  and  $\forall x \in X \|x_v - x_u\| \in H^{uv}$ ; also  $\forall x \in X \|x_v - x_u\| = \|R_x^{u+K}(x_v) - x_u\|$  and  $\forall x' \in X (x'_v \notin \{x_v, R_x^{u+K}(x_v)\} \rightarrow \|x_v - x_u\| \neq \|x'_v - x_u\|)$ .

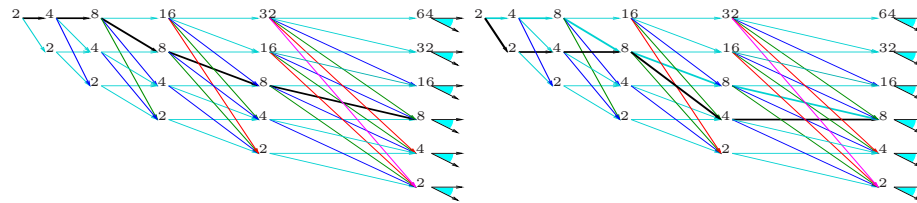
### 4 BP search trees with bounded width

Consider the BP tree for  $G_D$  and assume that there is a pruning distance  $\{u, v\} \in E_P$ ; at level  $u$  there are  $\max(2^{u-K}, 1)$  nodes, each of which is the root of a subtree with  $2^{v-\max(u, K)}$  nodes at level  $v$ . By the above remarks, for each such subtree only two nodes will encode a valid embedding for  $v$  (we call such nodes *valid*). Thus the number of valid nodes at level  $v > K$  is  $2^{\max(u-K+1, 1)}$ .

Consider the following Directed Acyclic Graph (DAG)  $\mathcal{D}_{uv}$ , used to compute the number of BP nodes in function of pruning distances  $\{u, v\}$  with  $u < v - K$ .



Nodes, arranged vertically, show the number of BP nodes in function of the rank of  $v$  w.r.t.  $u$  (first line). An arc is labelled with  $i_1, \dots, i_h$  if one of  $\{u + i_j, v\}$  (for  $j \leq h$ ) is a pruning distance, and is unlabelled if no such pruning distance exists. A path  $p$  in this DAG represents the set of pruning distances between  $u$  and  $v$ : each node  $p_\ell$  in this path shows the number of valid nodes in the BP search tree at level  $\ell$ . For example, following unlabelled arcs corresponds to no pruning distance between  $u$  and  $v$  and leads to a full binary BP search tree with  $2^{v-K}$  nodes at level  $v$ . Each set of pruning distances  $E_P$  corresponds to a longest path in  $\mathcal{D}_{1n}$ . BP trees have bounded width when these paths are below a diagonal with constant node labels. For example, if  $\exists v_0 \in V \setminus [K]$  s.t.  $\forall v > v_0 \exists! u < v - K$  with  $\{u, v\} \in E_P$  then the BP search tree width is bounded by  $2^{v_0-K}$ . This situation is pictured below (left). Another polynomial class of cases is shown on the right.



Out of a set of 16 protein instances from the Protein Data Bank (PDB), all yield BP trees of bounded width (with  $v_0 = 4$ ). This empirically illustrates the polynomiality of BP on real proteins.

### References

1. C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. The discretizable molecular distance geometry problem. *Comp. Opt. Appl.*, to appear.
2. L. Liberti, C. Lavor, A. Mucherino, and N. Maculan. Molecular distance geometry methods: from continuous to discrete. *Int. Trans. Op. Res.*, 18:33–51, 2010.
3. L. Liberti, B. Masson, C. Lavor, J. Lee, and A. Mucherino. Technical Report 1010.1834v1[cs.DM], arXiv, 2010.

## A Conditional Random Fields Method for RNA Sequence-Structure Relationship Modeling and Conformation Sampling

Zhiyong Wang<sup>1</sup> and Jinbo Xu<sup>1,\*</sup>

E-mail: zywang@ttic.edu jinbo.xu@gmail.com

<sup>1</sup>Toyota Technological Institute at Chicago  
6045 S Kenwood, Chicago, IL, 60637, USA

### Introduction

Accurate tertiary structures are very important for the functional study of non-coding RNA molecules. However, predicting RNA tertiary structures is extremely challenging because of a large conformation space to be explored and lack of an accurate scoring function differentiating the native structure from decoys. The fragment-based conformation sampling method (e.g., FARNA [1]) bears shortcomings that the limited size of a fragment library makes it infeasible to represent all possible conformations well. A recent dynamic Bayesian network method BARNACLE [2] overcomes the issue of fragment assembly. In addition, neither of these methods makes use of sequence information in sampling conformations. MC-Sym [3] is a motif assembly method for RNA 3D structure prediction, which uses a library of nucleotides cyclic motifs (NCM) to construct an RNA structure. Its high time complexity (respect to RNA length) prevents it being used to predict tertiary structure for a large RNA.

Here, we present a new probabilistic graphical model, Conditional Random Fields (CRFs) [4], to model RNA sequence-structure relationship, which enables us to accurately estimate the probability of a RNA conformation from sequence. Coupled with a novel tree-guided sampling scheme, our CRF model is then applied to RNA conformation sampling. Experimental results show that our CRF method can model RNA sequence-structure relationship well and sequence information is important for conformation sampling. Our method, named as TreeFolder, generates a much higher percentage of native-like decoys than FARNA and BARNACLE, although we use the same simple energy function as BARNACLE [2]. An extended version will appear in *Bioinformatics* published by Oxford University Press.

### Method

**Structure representation.** We represent an RNA 3D structure using a sequence of torsion angles, as shown in Fig. 1. Every nucleotide has seven bonds that rotate freely. Six of them lie on the backbone: P-O5', O5'-C5', C5'-C4', C4'-C3', C3'-O3' and O3'-P. The seventh bond connects a base to atom C1'. We use a simplified representation to reduce the number of torsion angles needed for the local conformation of a nucleotide [5-8]. In particular, we use the torsions  $\tau_1$  and  $\tau_2$  on pseudo-bonds P-C4' and C4'-P (see pink lines in Fig.1). However, to determine coordinates of the six backbone atoms of a nucleotide, we also need two planar angles  $\theta$ ,  $\psi$  and another torsion  $\alpha$  on bond P-O5'. Overall, we use a five-tuple  $(\tau_1, \tau_2, \theta, \psi, \alpha)$  to represent the local conformation of a nucleotide, similar as previous works [5-8].

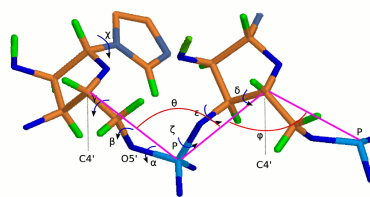


Fig. 1. Conformation of a nucleotide is represented by angles.

**Conformation state.** We use a Gaussian distribution to describe the local conformation preference of one nucleotide. First, we cluster all the angles collected from the experimental

\*Corresponding author

2 Zhiyong Wang<sup>1</sup> and Jinbo Xu<sup>1,\*</sup>

structures into dozens of groups (20~100). Then we fit a Gaussian distribution to each group. Each group (or cluster) and its Gaussian distribution are identified by an index, which is also denoted as a conformation state. Given the conformation state of a nucleotide, we can sample its real-valued angles from the corresponding distribution.

**Conditional Random Fields Model.** Our CRF method can estimate the probability of RNA conformation states from the primary sequence and secondary structure. A CRF model consists

of two major components: input features (nucleotide types, base pairing states) and output labels (conformation states), as in Fig.2. In contrast to BARNACLE [2] estimating the generative probability of an RNA structure, our CRF model estimates the conditional probability of an RNA structure, represented as a conformation state vector  $y$ , from the input feature vector  $x$  as Equation (1).  $Z(x)$  is the partition function;  $x_i$  is the feature vector at position  $i$ ;  $y_i$  is the label at position  $i$ ;  $w_{i,j}$  is the weight for transition from state  $i$  to  $j$ ;  $v_i$  is the weight factor for predicting state  $i$  from input feature  $x$ ;  $L$  is the length of RNA. The function  $\psi$  describes dependency between a conformation state and the input features and thus, called a label feature function. The function  $\Phi$  describes dependency between two adjacent states and thus, called an edge feature function. We extend the 1<sup>st</sup>-order CRF model to the 2<sup>nd</sup>-order model so that we can capture dependency among three adjacent nucleotides. As in Fig.2, two adjacent positions are combined to a single super-node.

$$P(Y = \bar{y} | X = \bar{x}) = \frac{1}{Z(\bar{x})} \exp\left[\sum_{i=1}^L \psi(y_i, \bar{x}) + \sum_{i=1}^{L-1} \Phi(y_i, y_{i+1})\right] \quad (1)$$

$$\bar{y} = (y_1 \dots y_L), \quad \psi(y_i, \bar{x}) = V_i^T \bar{x}, \quad \Phi(y_i, y_{i+1}) = W_{y_i, y_{i+1}}$$

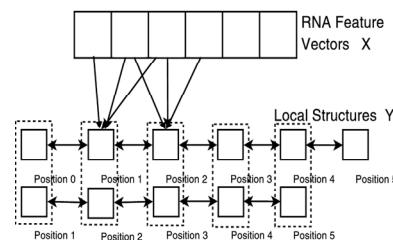


Fig. 2. The 2<sup>nd</sup>-order CRF model describes RNA sequence-structure relationship. A super-node in this model contains the conformation states in two adjacent positions.

**Conformation sampling with a guide tree.** A guide tree represents the base pairing information in an RNA and determine the order of conformation sampling. Given the base pairing information, we build the guide tree as follows. The root node in the tree corresponds to the whole RNA. Given a base pair  $(i, j)$ , we have one node in the tree corresponding to the segment between  $i$  and  $j$ . One node  $A$  is the child of the other node  $B$  if and only if the segment corresponding to  $B$  is the minimal segment containing the segment corresponding to  $A$ . We always can add some intermediate nodes to make each node have at most two children. Pseudoknots are removed by eliminating minimal base pairings in a guide tree.

To sample conformations of an RNA, we first mark all the nodes in the guide tree as “undone” at beginning. The torsion angles of the RNA are sampled using a bottom-up method along the tree as follows. We randomly pick up an “undone” node  $A$  in the tree which is either a leaf node or a node with all the child nodes being marked as “done”. (1) If  $A$  is a leaf node, we sample the angles for the segment corresponding to  $A$  using the segment conformation sampling algorithm. (2) If  $A$  has one or two child nodes, by cutting out the segments corresponding to the child nodes, we have at most three separate segments left in  $A$ , for which we use the segment conformation sampling algorithm to generate angles separately.

This segment conformation sampling algorithm consists of two steps: sampling a label for each nucleotide in the segment by the probability calculated from the CRF model and sampling real-valued angles from the Gaussian distribution corresponding to a label. We use a forward-backward algorithm [9] to sample the label sequence of a segment.

The new conformation is accepted if its energy is lower. Otherwise it is accepted by a probability  $\exp(\Delta E/T)$  where  $\Delta E$  is the energy difference between current and the new conformations and  $T$  is the annealing temperature. This sampling procedure is repeated 3000



times and then node A is marked as “done”. The folding simulation process ends when the root node is marked as “done”.

**Energy function.** Different from the complex energy function in FARNA, we adopt a simple energy function used by BARNACLE [2] as right. Where H is the number of hydrogen bonds formed in the secondary structure,  $\hat{d}_k$  is the distance

$$E = \sqrt{\frac{1}{|H|} \sum_{k=1}^{|H|} (\hat{d}_k - d_k)^2}$$

between the donor and the acceptor of the  $k^{th}$  hydrogen bond, and  $d_k$  is the average length of hydrogen bonds of the same type. The smaller this value is, the more the decoy is consistent with its secondary structure.

## Result

We build our training dataset from the RNA structure classification database DARTS [10]. Then we use 11 RNAs tested by both BARNACLE and FARNA to benchmark our method TreeFolder. These RNAs contain 12~46 nucleotides and are not homologous to any structures in our training dataset. In case that an RNA has multiple NMR structures, we use the first structure in the PDB file as its native structure.

It is not very reliable to compare two methods simply using the decoys with the lowest RMSD since they may be generated by chance and also depend on the number of decoys to be generated. The more decoys are generated, the more likely the lowest-RMSD decoy has lower RMSD from the native. Therefore, a better strategy is to compare the RMSD distributions of decoys.

**Our TreeFolder generates better decoys than FARNA.** We compare FARNA and TreeFolder in terms of the quality of the decoy clustering centroids. Similar to FARNA clustering only on the top 1% decoys with the lowest energy, we run MaxCluster to cluster the top 1% of our decoys with the lowest energy into 5 clusters. As shown in Table 1, TreeFolder can generate decoys with better cluster centroids for 9 RNAs: 1a4d, 1esy, 1kka, 1q9a, 1xjr, 1zih, 28sp, 2a43, and 2f88. By the way, even if a significantly smaller number of decoys are generated by us, the lowest RMSD decoys by our TreeFolder for 1a4d, 1zih and 28sp still have smaller RMSD than those by FARNA.

**Our TreeFolder generates better decoys than BARNACLE.** Table 2 displays the 5% and 25% quantiles of the RMSD distributions for decoys generated by BARNACLE and TreeFolder. The quantiles by BARNACLE are taken from Table S4 in [2]. BARNACLE considers only decoys with energy less than 1 since this kind of decoys are likely to have more correct base pairings. We use exactly the same energy function as BARNACLE, so we also consider only decoys with energy less than 1

**Table 1.** Comparison between FARNA and our method TreeFolder. The results of FARNA are taken from Table 1 in [1]. Column “Best cluster centroid” lists the RMSD of the best cluster centroid of the top 1% decoys with the lowest energy. Column “#decoys” is the number of decoys generated by the methods. Bold fonts indicate better results.

PDB ID	Method	Len	FARNA			TreeFolder		
			Best cluster centroid	Lowest RMSD decoy	#decoys	Best cluster centroid	Lowest RMSD decoy	#decoys
1a4d	NMR	41	6.48	3.43	28949	<b>3.65</b>	<b>2.69</b>	7168
1esy	NMR	19	3.98	<b>1.44</b>	69103	<b>2.00</b>	1.52	22529
1kka	NMR	17	4.14	<b>2.08</b>	81492	<b>3.71</b>	2.40	24934
1l2x	X-ray	27	<b>3.88</b>	<b>3.11</b>	47958	8.07	3.97	15360
1q9a	X-ray	27	6.11	<b>2.65</b>	48817	<b>4.76</b>	3.50	15415
1qwa	NMR	21	<b>3.71</b>	<b>2.01</b>	65977	3.77	2.49	18838
1xjr	X-ray	46	9.82	<b>6.25</b>	24646	<b>9.26</b>	7.05	7168
1zih	NMR	12	1.71	1.03	117104	<b>1.19</b>	<b>0.73</b>	40960
28sp	NMR	28	3.20	<b>2.31</b>	46034	<b>2.96</b>	<b>1.91</b>	17117
2a43	X-ray	26	4.93	<b>2.79</b>	49972	<b>4.52</b>	3.47	18432
2f88	NMR	34	3.63	<b>2.41</b>	36664	<b>3.33</b>	2.70	12230

**Table 2.** The 5% and 25% quantiles of the RMSD distributions for decoys generated by our method TreeFolder and BARNACLE. Bold numbers indicate better distributions. Columns “#energy<1” and “#energy<2” list the number of decoys with energy less than 1 and 2, respectively. “Bps” is the number of base pairings.

PDB ID	Len	Bps	BARNACLE		TreeFolder					
			5%	25%	5%	25%	#energy<1	5%	25%	#energy<2
1esy	19	6	2.99	3.28	<b>2.19</b>	<b>2.60</b>	577	<b>2.25</b>	<b>2.78</b>	1102
1kka	17	6	4.40	5.02	<b>3.75</b>	<b>4.30</b>	349	<b>3.8</b>	<b>4.39</b>	776
1l2x	27	8	5.43	6.88	-	-	0	5.44	8.08	5
1q9a	27	6	4.80	5.42	<b>4.55</b>	<b>5.05</b>	486	<b>4.61</b>	<b>5.07</b>	1025
1qwa	21	8	4.06	4.64	<b>3.65</b>	<b>4.26</b>	407	<b>3.9</b>	<b>4.51</b>	884
1xjr	46	15	10.41	11.01	<b>8.50</b>	<b>9.43</b>	22	<b>8.84</b>	<b>9.79</b>	540
1zih	12	4	1.72	2.16	<b>1.32</b>	<b>1.84</b>	1721	<b>1.36</b>	<b>1.88</b>	1931
28sp	28	8	3.23	3.76	<b>2.88</b>	<b>3.43</b>	152	<b>2.93</b>	<b>3.58</b>	563
2a43	26	7	4.72	6.08	-	-	0	<b>4.64</b>	<b>5.48</b>	26
2f88	34	13	3.82	4.41	<b>3.73</b>	<b>3.73</b>	1	3.85	4.57	130

4 Zhiyong Wang<sup>1</sup> and Jinbo Xu<sup>1,\*</sup>

to ensure a fair comparison. We did not generate as many decoys as BARNACLE and thus, for some test RNAs we do not have many decoys with energy less than 1. In this case we use decoys with energy less than 2. On the 10 RNAs shown in Table 2, TreeFolder yields better RMSD distributions for 8 of them: 1esy, 1kka, 1q9a, 1qwa, 1xjr, 1zih, 28sp, 2a43 and 2f88.

## Conclusions

We have presented a new method TreeFolder for modeling RNA sequence-structure relationship and conformation sampling using conditional random fields (CRFs) and a tree-guided sampling scheme. Our CRF method not only captures the relationship between sequence and angles, but also models the interdependency among the angles of three adjacent nucleotides. Our conformation sampling method distinguishes from FARNA in that we do not use fragments to build RNA conformations so that we do not need to worry about if there are a sufficient number of structure fragments to cover all the possible local conformations. Our TreeFolder also differs from both FARNA and BARNACLE in that we use primary sequence to estimate the probability of backbone angles while the latter two do not. In addition, we also use a tree, built from (predicted) 2<sup>nd</sup> structure, to guide conformation sampling so that at one moment we can simultaneously sample conformations for two segments far away from each other along the RNA sequence. By contrast, both FARNA and BARNACLE can only sample conformations for a single short segment at any time. The results indicate that our TreeFolder indeed models sequence-structure relationship well and compares favorably to both FARNA and BARNACLE, even if we use only the same simple energy function as BARNACLE.

**Acknowledgements** This work is financially supported by the National Institute of Health grant R01GM089753 (to JX) and the National Science Foundation grant DBI-0960390 (to JX). The authors are also grateful to the Open Science Grid and to TeraGrid for the computational resources of grants TG-MCB100062 and TGCCR100005 (to JX).

## Reference

1. Das, R., Baker, D.: Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences* 104, 14664-14669 (2007)
2. Frellsen, J., Ida, M., Martin, T., V., M.K., Jesper, F.-B., Thomas, H.: A Probabilistic Model of RNA Conformational Space. *PLoS Computational Biology* 5, 1000406 (2009)
3. Parisien, M., Major, F.: The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452, 51-55 (2008)
4. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. In: *ICML 2001: Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282-289. Morgan Kaufmann Publishers Inc., (Year)
5. Hershkovitz, E., Sapiro, G., Tannenbaum, A., Williams, L.D.: Statistical Analysis of RNA Backbone. *IEEE/ACM Transaction on Computational Biology and Bioinformatics* 3, 33 (2006)
6. Zhang, J., Lin, M., Chen, R., Wang, W., Liang, J.: Discrete state model and accurate estimation of loop entropy of RNA secondary structures. *The Journal of Chemical Physics* 128, 125107 (2008)
7. Cao, S., Chen, S.: Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA* 11, 1884-1897 (2005)
8. Duarte, C.M., Pyle, A.M.: Stepping through an RNA structure: a novel approach to conformational analysis. *Journal of Molecular Biology* 284, 1465-1478 (1998)
9. Zhao, F., Li, S., Sterner, B.W., Xu, J.: Discriminative learning for protein conformation sampling. *Proteins: Structure, Function, and Bioinformatics* 73, 228-240 (2008)
10. Abraham, M., Dror, O., Nussinov, R., Wolfson, H.J.: Analysis and classification of RNA tertiary structures. *RNA* 14, 2274-2289 (2008)

## Searching for Important Amino Acids in DNA-binding Proteins for Histogram Methods

Andrea Szabóová<sup>1</sup>, Ondřej Kuželka<sup>1</sup>, Filip Železný<sup>1</sup>, and Jakub Tolar<sup>2</sup>

<sup>1</sup> Czech Technical University, Prague, Czech Republic

<sup>2</sup> University of Minnesota, Minneapolis, USA  
szaboand@fel.cvut.cz

**Abstract.** We develop a method capable to identify important amino acids for histogram-based methods predicting DNA-binding propensity. This method can be used both for prediction from sequence information (*Tube Histograms*) and prediction from structural information (*Ball Histograms*). We validate our method in prediction experiments using only proteins' primary structure, achieving favourable accuracies. Moreover, the histogram-based methods equipped with this new searching method also provide interpretable features involving distributions of amino acids.

**Keywords:** Feature construction, Proteomics, DNA-binding proteins

### 1 Introduction

The process of protein-DNA interaction has been an important subject of recent bioinformatics research, however, it has not been completely understood yet. DNA-binding proteins have a vital role in the biological processing of genetic information like DNA transcription, replication, maintenance and the regulation of gene expression. Several computational approaches have been proposed for the prediction of DNA-binding function from protein structure.

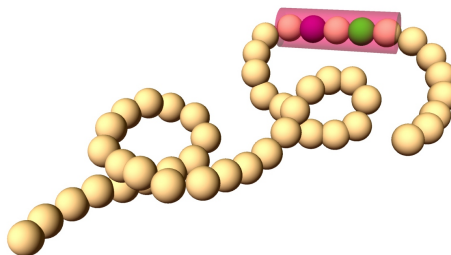
In this paper we will be concerned with prediction of DNA-binding propensity from sequence information. Previously developed methods for DNA-binding-propensity prediction can be divided into two main groups: alignment-based approaches [5] and physicochemical-property-based approaches [8, 10]. Gao and Skolnick [5] developed a threading-based method for the prediction of DNA-binding domains and associated DNA-binding protein residues. Ofra et al. [8] used only protein sequence information, without requiring any additional experimental or structural information. Their method relies on sequence environment, evolutionary profiles and predicted structural features (secondary structure, solvent accessibility and globularity). Patel et al. [10] used artificial neural network for prediction from amino acid sequences. Yan et al. [4] started with a Naive Bayes classifier trained to predict whether a given amino acid residue is a DNA-binding residue based on its identity and the identities of its four sequence neighbours on each side of the target residue.

Recently, we have introduced histogram-based methods which are able to predict DNA-binding propensity either from sequence information (*Tube Histograms*) or from structural information (*Ball Histograms* [11]). In this paper we develop a method capable to identify important amino acids for histogram-based methods predicting DNA-binding propensity.

## 2 Method

We propose the following approach to predict DNA-binding propensity. It consists of four main parts. First, so-called *templates* are found, which determine amino acids whose distributions should be captured by *tube histograms*. In the second step *tube histograms* are constructed for all proteins in a training set. Third, a transformation method is used to convert these histograms to a form usable by standard machine learning algorithms. Finally, a random forest classifier [3] is learned on this transformed dataset and then it is used for classification.

A *template* is a list of names of some Boolean amino acid properties. Given a template and a location in the primary structure of a protein, we infer a list of binary values indicating the truth values of the respective properties in the template for the amino acid at the position. For example, the template (*Arg, Lys, Positive, Negative, Neutral*) acquires the value (1, 0, 1, 0, 0) if the amino acid at the inquired position is an Arginine. A *tube* of size  $s$  represents a part of an amino-acid sequence containing  $s$  consecutive amino acids (see Fig. 1).



**Fig. 1.** Illustration of the Tube Histogram Method - Amino acids are shown as small balls in sequence forming an amino acid chain. They have different colors according to their type.

Given a protein, a template  $\tau = (f_1, \dots, f_k)$  and a sampling-tube size  $s$ , a *tube histogram* is a  $k$ -dimensional histogram constructed as follows. Starting by placing the sampling tube on the first  $s$  amino acids we get the first *sample* for the histogram. When a sample is collected the numbers of amino acids complying with the particular properties listed in the given template are extracted from it and stored. In further steps the tube is moved by one amino acid at time along the protein sequence and the samples are continuously stored for subsequent histogram construction. This process ends when the last amino acid is reached.

Finally, the histogram constructed from the collected samples is normalized. Intuitively, tube histograms capture the joint probability that a randomly picked *sampling tube* (Fig. 1) will contain exactly  $t_1$  amino acids complying with  $f_1$ ,  $t_2$  amino acids complying with  $f_2$  etc.

In our previous study [11] we used pre-fixed templates with charged amino acids selected according to [9, 7, 6]. Here, we introduce a method for automatic selection of templates which are sufficiently discriminative to distinguish DNA-binding proteins from non-DNA-binding proteins. The basic idea of the method is to find templates which maximize *distance* between average histograms from the two classes (DNA-binding and non-DNA-binding proteins). Intuitively, such templates should allow us to construct classifiers with good discriminative ability.

We construct the templates in a heuristic way using best-first search algorithm to maximize Bhattacharyya distance [2] between the average histograms from the two classes. In order to avoid repeated construction of histograms from the whole datasets, we construct a histogram corresponding to the biggest possible template (containing all amino acid properties), then, during the search, we construct histograms for the other templates by marginalising this biggest histogram.

### 3 Results

In this section we present experiments performed on real-life data (PD138 [12]/NB110 [1]). We decided to study distribution of amino acids (represented by *tube histograms*). We constructed histograms with automatically discovered templates (with maximum length 5) and three different sampling-tube sizes: 5, 10 and 15. We trained random forest classifiers selecting optimal sampling-tube size and an optimal number of trees for each fold by internal cross-validation. The estimated accuracy and area under ROC is shown in Table 1. As we can see, the accuracy of our method exceeds the accuracy obtained by the method used in [12].

Method	Accuracy	AUC
Szilágyi et al.	81.4	0.92
Tube Histogram	<b>86.3</b>	<b>0.94</b>

**Table 1.** Accuracies and AUCs estimated by 10-fold cross-validation on PD138/NB110.

The four most informative automatically selected templates are: (*Arg, Cys, Lys, Gly, Ala*), (*Arg, Cys, Lys, Gly, Asp*), (*Arg, Cys, Lys, Gly, Glu*), (*Arg, Cys, Lys, Gly, Leu*). It is noteworthy that each charged amino acids (under normal circumstances *Arg* and *Lys* are positively charged, whereas *Glu* and *Asp* are charged negatively) is contained at least one of these templates.

## 4 Conclusions

We developed a method capable to identify important amino acids for histogram-based methods predicting DNA-binding propensity. We validated our method in prediction experiments using only proteins' primary structure, achieving favourable accuracies. In future work we plan to validate this method in prediction experiments using proteins' structural information (*Ball Histograms*).

**Acknowledgement:** Andrea Szabóová and Filip Železný were supported by project ME10047 granted by the Czech Ministry of Education. Andrea Szabóová was further supported by the Czech Technical University internal grant #10-801940. Ondřej Kuželka was supported by the Czech Technical University internal grant #10-811550.

## References

1. Shandar Ahmad and Akinori Sarai. Moment-based prediction of dna-binding proteins. *Journal of Molecular Biology*, 341(1):65 – 71, 2004.
2. A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* 35, pages 99–109, 1943.
3. L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
4. C. Yan C, M. Terribilini, F. Wu abd R. L. Jernigan, D. Dobbs D, and V. Honavar. Predicting dna-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, 19(7):262, 2006.
5. M. Gao and J. Skolnick. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. *PLoS Computational Biology*, 5(11), 2009.
6. S. Jones, P. van Heyningen, H. M. Berman, and J. M. Thornton. Protein-DNA interactions: a structural analysis. *Journal of Molecular Biology*, 287(5):877–896, 1999.
7. Y. Mandel-Gutfreund, O. Schueler, and H. Margalit. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *Journal of Molecular Biology*, 253(2):370–382, 1995.
8. Y. Ofran, V. Mysore, and B. Rost. Prediction of dna-binding residues from sequence. *Bioinformatics*, 23(13):347–53, 2007.
9. C. O. Pabo and R. T. Sauer. Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry*, 61:1053–1095, 1992.
10. A. K. Patel, S. Patel, and P. K. Naik. Binary classification of uncharacterized proteins into DNA binding/non-DNA binding proteins from sequence derived features using ann. *Digest Journal of Nanomaterials and Biostructures*, 4(4):775–782, 2009.
11. A. Szabóová, O. Kuželka, S. Morales E., F. Železný, and J. Tolar. Prediction of dna-binding propensity of proteins by the ball-histogram method. In *ISBRA 2011, LNBI 6674*, pages 358–367, 2011.
12. András Szilágyi and Jeffrey Skolnick. Efficient prediction of nucleic acid binding function from low-resolution protein structures. *Journal of Molecular Biology*, 358(3):922 – 933, 2006.

# Finding Deletions with Exact Break Points from Noisy Low Coverage Paired-end Short Sequence Reads

Jin Zhang and Yufeng Wu

Department of Computer Science and Engineering,  
 University of Connecticut,  
 Storrs, CT 06269, U.S.A.  
 {jinzhang, ywu}@engr.uconn.edu

**Abstract.** While there is great interest in structural variants (SV), it is very challenging to identify them using noisy low coverage next-generation sequencing short reads. Methods of analysing discordant insert sizes and split-reads mapping have been used but are not completely satisfying due to the low quality of data. We present a two-stages approach combining the two methods, which tolerates the low quality data to find the exact break points of deletions. Experiments suggest that our approach is more accurate and efficient than an alternative approach.

**Keywords:** Structural variant, Next-generation sequencing, Burrows-Wheeler transform

## 1 Introduction

High throughput sequencing technologies have generated huge amount of sequence data. One application is using these sequence data to discover SVs. A major challenge in this application is developing efficient and accurate algorithms to find SVs from the large amount of sequence data. Refer to [2, 4, 6, 8] for surveys on the latest methods. While many current methods work well with high quality data (high coverage and low errors), in practice most existing sequence data has low quality. For example, the 1000 genomes project [1] uses low-coverage sequencing to sequence hundreds of individuals from several human populations. Sequence errors, substitutions and small indels may cause problems in mapping these sequence reads onto the reference genome. In this paper, we develop a two-stage approach for discovering deletions using noisy low coverage data. The first step is applying an enhanced split-reads mapping approach to identify *candidate* deletion sites from population sequence reads. The second step is finding mapped paired-end reads which *span* candidate deletion sites and have insert length matching the candidate deletions. Our approach exploits more information in the sequence data than existing approaches by using both insert length information of mapped paired-end reads and break points information from mapped split-reads.

## 2 Method

Candidate deletions with exact break points were discovered by utilizing the paired-end reads that with only one end mapped. The mapped end is an anchor, and the other end is the split-read. Like several well known reads mapping tools, such as Bowtie [3] and BWA [5], our method uses Burrows-Wheeler transform as basic utilities to achieve higher efficiency and tolerates certain errors. Searching locally (near the anchor) can even speed up the aligning process and give more accurate results. To call deletions from the candidates, insert size changes of paired-end reads that are mapped spanning the candidate deletion are examined to find supports.

## 3 Results

To evaluate the accuracy of our method when applied to low coverage data with noise, we test our method using 1000 genomes project pilot one data, and compared our results with those of Pindel [6] viewing the releases [7, 9] of 1000 genomes project as benchmarks. The experiment suggests that our approach is more accurate and efficient.

## 4 Acknowledgement

Experiment is run on workstations supported by NSF grant IIS-0916948. Research partly supported by NSF grant IIS-0953563.

## References

1. The 1000 genomes project consortium. <http://www.1000genomes.org/>.
2. F. Hormozdiari, C. Alkan, EE. Eichler and SC. Sahinalp Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* 19: 1270–1278, 2009.
3. B. Langmead, C. Trapnell, M. Pop and SL. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome *Genome Biology*, 10:R25, 2009.
4. P. Medvedev, M. Stanciu and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* 6, S13 - S20, 2009
5. H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.
6. K. Ye, MH. Schulz, Q. Long, R. Apweiler and Z. Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25:2865–2871, 2009.
7. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.
8. S. Lee, E. Xing and M. Brudno. MoGUL: Detecting Common Insertions and Deletions in a Population. RECOMB 2010, LNBI 6044 356–368, 2010.
9. RE. Mills et al. Mapping copy number variation by population-scale genome sequencing *Nature* 470, 5965, 2011



## Genome and Systems Evolution in *Chlamydiae*

Hong Cai<sup>1, #</sup>, Jianying Gu<sup>2, #</sup>, Zhan Zhou<sup>1,3, #</sup>, Yufeng Wang<sup>1, \*</sup>,

<sup>1</sup> Department of Biology, South Texas Center for Emerging Infectious Diseases  
University of Texas at San Antonio, San Antonio, TX 78249, USA  
hong.cai@utsa.edu, zhan.zhou@utsa.edu, yufeng.wang@utsa.edu (\*corresponding author)

<sup>2</sup> Department of Biology, College of Staten Island  
City University of New York, Staten Island, NY 10314, USA  
jianying.gu@csi.cuny.edu (# Equally contributed)

<sup>3</sup> College of Life Sciences, Zhejiang University, Hangzhou 310058, P.R. China,

**Abstract.** *Chlamydiae* represents a unique group of obligate intracellular bacteria that are the causative agents of a variety of human and animal infectious diseases, including the most common sexually transmitted disease. In this study, we investigated the genome plasticity of 14 evolutionarily closely related *Chlamydiae* strains and identified the components whose functions range from fundamental biological processes to complex networks specific to intracellular parasitism.

**Keywords:** *Chlamydiae*, genomics, evolution

### 1 Introduction

*Chlamydiae* is a phylum of gram-negative obligate intracellular bacteria, which encompasses two genera of human and animal pathogens: (1) the genus *Chlamydia* includes three species: *C. trachomatis* is the causative agent for the most common bacterial sexually transmitted infections in humans and the leading cause of infectious blindness globally [1-5]; *C. muridarum* infects mice and hamsters, causing pharyngitis, bronchitis, and pneumonitis [6]; *C. suis* infects swine, causing pneumonia, enteritis, conjunctivitis, pericarditis, perinatal mortality, and reproductive disorders; (2) the genus *Chlamydophila* includes six species: *C. abortus* is a common cause of infectious abortion in sheep, goats, cattle and pigs, and represents a significant risk to pregnant women [7]; *C. caviae* is the causative agent for guinea pig conjunctivitis [8]; *C. felis* causes pneumonia and conjunctivitis in cats [9]; *C. pecorum* infects cattle, sheep and goats, koalas, and swine, and is associated with abortion, conjunctivitis, pneumonia, and polyarthritis; *C. pneumoniae* infects humans. It is a major cause of pneumonia and is associated with atherosclerosis [5, 6, 10, 11]; *C. psittaci* is the causative agent of psittacosis in birds and humans [8].

Despite sharing a common developmental cycle that alternates between an extracellular, infectious elementary body (EB) stage and an intracellular, noninfectious reticulate body (RB) stage, the bacteria in the *Chlamydiae* phylum

## ISBRA 2011 Short Abstracts

exhibit striking difference in their host specificity and disease outcome. In this study, we report a comprehensive survey of the complete genomes of 14 *Chlamydiae* strains.

**Table 1.** Genomic sequences used in the comparative analysis of *Chlamydiae*. The inter-genomic search yielded a core genome comprised of 764 orthologous proteins.

Strains	Accession ID	No. Genes in genome	No. Protein coding genes	% core in genome
<i>Chlamydia muridarum</i> Nigg	NC_002620	955	904	84.62
<i>Chlamydia trachomatis</i> 434/Bu	NC_010287	934	874	87.53
<i>Chlamydia trachomatis</i> A/HAR-13	NC_007429	955	911	84.08
<i>Chlamydia trachomatis</i> B/Jali20/OT	NC_012686	936	875	87.43
<i>Chlamydia trachomatis</i> B/TZ1A828/OT	NC_012687	937	880	86.93
<i>Chlamydia trachomatis</i> D/UW-3/CX	NC_000117	940	895	85.47
<i>Chlamydia trachomatis</i> L2b/UCH-1/proctitis	NC_010280	934	874	87.53
<i>Chlamydophila abortus</i> S26/3	NC_004552	1003	932	82.30
<i>Chlamydophila caviae</i> GPIC	NC_003361	1053	998	76.95
<i>Chlamydophila felis</i> Fe/C-56	NC_007899	1046	1005	76.82
<i>Chlamydophila pneumoniae</i> AR39	NC_002179	1167	1112	69.33
<i>Chlamydophila pneumoniae</i> CWL029	NC_000922	1122	1052	73.19
<i>Chlamydophila pneumoniae</i> J138	NC_002491	1110	1069	72.12
<i>Chlamydophila pneumoniae</i> TW-183	NC_005043	1155	1113	69.27

## 2 Data and Methods

### 2.1 Data

We collected the complete genomes of 14 *Chlamydiaceae* strains (Table 1). The Genbank RefSeq annotation was integrated with genome information collected from the J. Craig Venter institute's (JCVI) Comprehensive Microbial Resources Genomics database (<http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi>) and NCBI.

### 2.2 Cluster of gene families and functional classification analysis

To identify the presence of orthologous and paralogous genes, we merged all proteins of 14 *Chlamydiaceae* genomes and conducted an exhaustive all-against-all BLASTP search; genes were defined as orthologous or paralogous if (1) they had a E-score  $< e^{-10}$ ; (2) their similarity  $I$  was  $\geq 30\%$  if the length of the alignable region  $L \geq 150$  amino acid residues, or  $I = 0.01n + 4.8L(-0.32(1 + \exp(-L/1000)))$ , if  $L < 150$  aa, where  $n$  = the number of sequences; (3) the length of the alignable region between the two sequences was  $>50\%$  of the longer protein [12]. A Markov cluster algorithm, OrthoMCL, was used to cluster genes into gene clusters [13]. Multiple alignments of each clusters were obtained by the program ClustalX [14] and T-coffee [15], followed by manual inspection and editing. Phylogenetic trees were inferred by the neighbor-joining method, using MEGA5 (<http://www.megasoftware.net/>). A hierarchical functional classification was performed for each *Chlamydiaceae* sequence by searching against the Clusters of Orthologous Groups (COG) database [16]. The classification of specific supergene families including transporters, kinases, and proteases was based on the standard nomenclature defined in the Transporter Classification (TC) system, the Kinase Classification System, and Merops.

## 3 Results and Discussion

The OrthoMCL analysis revealed that the core genome of the 14 *Chlamydiae* strains we examined is comprised of 764 orthologous genes, accounting for about 69-87% of the genome complements (Table 1). The proportions of core genome components in the *Chlamydia* genus are very similar (84-87%). *Chlamydomphila* has a slightly lower proportion of the core genome (69-82%).

617 (81%) of the 764 orthologous clusters in the core genome were predicted to fall into a COG, while the remaining 147 (19%) appear to have no identifiable functions. The core genome contains the components for fundamental biology such as genetic information processing (replication, transcription and translation), and metabolism. It also includes abundant components that have been implicated in pathogenesis such as Type III secretion system. A better understanding of the genome plasticity and evolution in *Chlamydiae* can bring new insights into the mechanism underlying pathogenesis, tissue tropism, and niche adaptation.

**Acknowledgments.** This work is supported by NIH grants AI067543, GM081068 and AI080579 to YW, and the PSC-CUNY Research Award PSCREG-39-497 to JG. ZZ is supported by the government scholarship from China Scholarship Council.

## References

1. Stephens, R.S., et al.: Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282, 754-759 (1998)
2. Carlson, J.H., et al.: Comparative genomic analysis of *Chlamydia trachomatis* oculotropic and genitotropic strains. *Infect Immun* 73, 6407-6418 (2005)
3. Thomson, N.R., et al.: *Chlamydia trachomatis*: genome sequence analysis of lymphogranuloma venereum isolates. *Genome Res* 18, 161-171 (2008)
4. Seth-Smith, H.M.B., et al.: Co-evolution of genomes and plasmids within *Chlamydia trachomatis* and the emergence in Sweden of a new variant strain. *Bmc Genomics* 10, - (2009)
5. Kalman, S., et al.: Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat Genet* 21, 385-389 (1999)
6. Read, T.D., et al.: Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res* 28, 1397-1406 (2000)
7. Thomson, N.R., et al.: The *Chlamydia abortus* genome sequence reveals an array of variable proteins that contribute to interspecies variation. *Genome Res* 15, 629-640 (2005)
8. Read, T.D., et al.: Genome sequence of *Chlamydia caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. *Nucleic Acids Res* 31, 2134-2147 (2003)
9. Azuma, Y., et al.: Genome sequence of the cat pathogen, *Chlamydia felis*. *DNA Res* 13, 15-23 (2006)
10. Shirai, M., et al.: Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Res* 28, 2311-2314 (2000)
11. Shirai, M., et al.: Comparison of outer membrane protein genes *omp* and *pmp* in the whole genome sequences of *Chlamydia pneumoniae* isolates from Japan and the United States. *J Infect Dis* 181 Suppl 3, S524-527 (2000)
12. Gu, Z., et al.: Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol Biol Evol* 19, 256-262 (2002)
13. Li, L., et al.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13, 2178-2189 (2003)
14. Larkin, M.A., et al.: Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947-2948 (2007)
15. Poirot, O., et al.: Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res* 31, 3503-3506 (2003)
16. Tatusov, R.L., et al.: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41 (2003)

## Accurate Reconstruction of Microbial Community From Environmental Shotgun Sequences Avoiding Primer Bias

Lu Fan<sup>1</sup>, Staffan Kjelleberg<sup>1</sup> and Torsten Thomas<sup>1</sup>

<sup>1</sup> Centre for Marine Bio-Innovation and School of Biotechnology and Biomolecular Sciences,  
University of New South Wales, Sydney, Australia  
leanderfan@gmail.com

**Abstract.** Metagenomic studies employing high-throughput, next-generation sequencing technologies have brought a new era of microbial ecology and evolution. Metagenomics provides access to great taxonomic diversity and enable characterization of part of the biosphere currently hidden from polymerase chain reaction (PCR) based surveys. In this study, we developed a novel workflow to accurately reconstruct the phylogenetic profile of microbial communities from metagenomic shotgun-sequencing data. We assess those profiles by phylogenetic analysis and cross-comparison with PCR generated sequences. We utilized replicate samples with shared microbial populations to generate nearly full-length SSU-rRNA sequences containing sufficient information for accurate phylogenetic analysis. Using this workflow, we identified high level of primer bias in a set of human twin-gut data, which lead to unreliable conclusions. We also indicated different levels of bias for three primer sets in samples with various community compositions.

**Keywords:** metagenomics; microbial community; primer bias; high-throughput sequencing

### 1 Introduction

As about 99% of the microorganisms in the environment are currently unculturable in laboratory. Direct amplification and sequencing of marker genes, often the small subunit ribosomal RNA genes (SSU-rRNA), from the environmental DNA samples has become a popular and powerful approach to assess the microbial diversity. However, as the “universal” primers used in PCR are designed based of groups of already known species, a skewed picture of community composition is potentially obtained for environmental samples containing divergent bacterial lineages [1]. Metagenomic approaches directly sequence randomly sheared (i.e. shotgun) DNA from environmental samples and hence provide a direct assess to the microbial community on a genomic level without the PCR primer bias problem.

Here we developed a novel workflow to accurately reconstruct the microbial communities composition down to species level using the SSU-rRNA sequences from replicated metagenomic samples. This workflow employs a stringent sequence assembly of samples with shared microbial populations to generate nearly full-length SSU sequences for accurate phylogenetic assignment. Through phylogenetic distance

based OTU clustering we also performed a comparison of our approach with PCR-based community assessments. In a test analysis of 18 metagenomic samples for a previous human twin-gut study [2], we observed a bias of the V2 PCR primer approach, which led to misinterpretation for the role of the human gut fauna in the original publication.

## **2 The Workflow**

### **2.1 Unique Sequence Generation and Coverage Calculation**

Metagenomic sequences produced by Sanger sequencing or 454 platforms (FLX or Titanium) are suitable for this workflow. Metagenomic reads were searched against the SILVA SSU databases [3] and hits with the E-value lower than  $1e-10$  and with alignment length  $> 45$  bp were retrieved. SSU-rRNA sequences from replicate samples were pooled and assembled by Newbler using the “cDNA” option (99% overlapping similarity and 43 bp minimal overlapping length). The generated contigs and unassembled singletons are defined as Overall Unique Sequences (OUSs) across samples. Reads from each sample were then mapped back against the OUSs to determine the relative abundance in each sample.

### **2.2 Phylogenetic Distance-Based Operational Taxonomic Unit (OTU)**

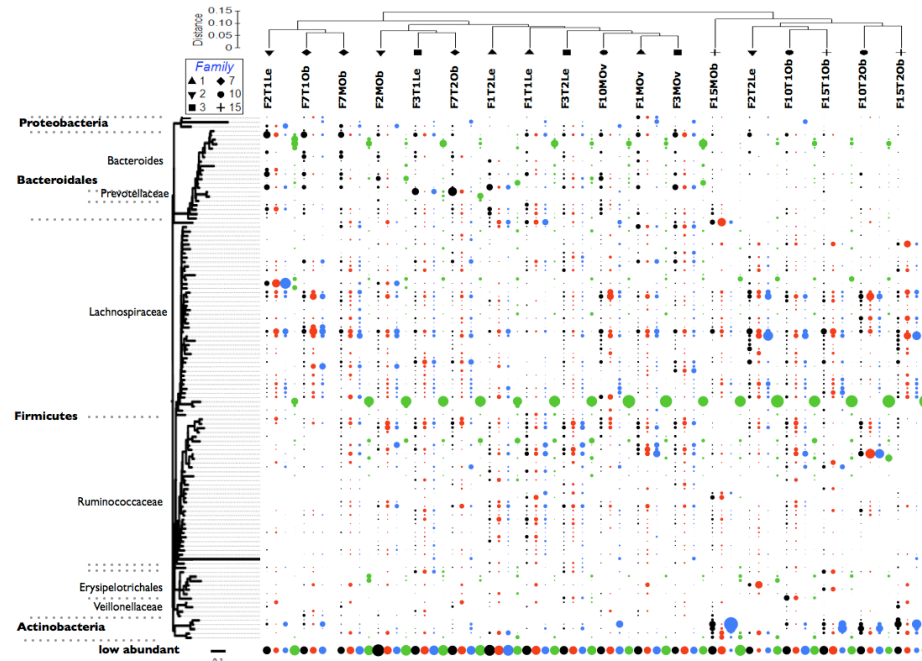
OUSs and sequences from PCR-based SSU-rRNA clone libraries or 454-tag sequencing data were clustered by CD-Hit (98% identity). Representatives of all unique clusters were aligned against the SILVA SSU seed alignment using Mothur [5]. Pairwise distances were calculated and sequences were clustered by Mothur (furthest neighbor algorithm and distance cutoff of 0.03). Representatives were then inserted into the SILVA SSURef tree in ARB. The pairwise phylogenetic distances of the representatives were calculated and clustered by Mothur (average linkage and distance cutoff of 0.03). Representatives of these clusters were then defined as the final operational taxonomic units (OTUs) and their abundance in each sample was calculated.

## **3 Analysis of the Human Twin-Gut Samples**

### **3.1 OTU Generation**

The 18 metagenomic data sets generated from pyrosequencing of the gut microbiome of human twins were downloaded from the NCBI Trace database [2]. Their corresponding 16S rRNA gene sequences were amplified by primers 8F/1391R targeting the full-length gene, by primers 8F/338R targeting the V2 region, and by a pool of 5 forward and 4 reverse primers targeting the V6 region and were retrieved. SSU-rRNA sequences from the metagenomic data sets were processed according to

our workflow and assembled into unique sequences. OTUs were generated from the pool of the metagenomic SSU unique sequences and all the 16S rRNA gene amplicons.



**Fig. 1.** Distribution of the 120 most abundant OTUs across samples. Proportion for the most abundant OTUs in each sample and the sum remaining low abundant OTUs were presented with the relative size of bubbles. Black bubbles, GS samples processed with the novel workflow; red, full-length gene; blue, V6 and green, V2 regions. The phylogenetic tree of the OTUs was extracted from the SILVA SSURef database and clades were taxonomic annotated according to their position in the SSURef tree. Host individuals are clustered according to their GS samples by Unifrac distance [4]. Hosts from the same family are labeled with the same shape.

### 3.2 Metagenomic Versus PCR-Based Community Profiles

We clustered the metagenomic samples and plotted the distribution of the most abundant OTUs according to this cluster (Fig. 1). Microbial communities of gut samples are generally diverse across all the 18 samples. Bacteria from the Bacteroidetes and Firmicutes groups comprise the majority of the lineages in most of the samples. One-factor analysis using PERMANOVA in PRIMER-E shows that metagenomic samples' community structure are significantly different among families ( $P = 0.002$ ) for the novel workflow analysis, but not among different body mass index (lean, overweight and obese) and between twins and mothers.

The three sets of primers targeting different regions of the 16S rRNA gene displayed different levels of bias (Fig. 1). Primers targeting the full length and the V6 region recovered most of the OTUs found in metagenomic shotgun sequencing data,

but varied in the relative abundance of certain OTUs. The V2 primers exhibited a substantial bias and shared few OTUs with the other three approaches. For the V2 primer only two species (belonging to phyla Bacteroidetes and Firmicutes) were presented in high abundance across samples. While the overall proportion of sequences assigned to the phyla Bacteroidetes and Firmicutes is relatively consistent between the four approaches used, our analysis shows that assignment to lower taxonomic ranks (e.g. species or genera) varies a lot and can not be considered reliable.

## 4 Conclusion

Our novel workflow is easily implemented into current metagenomic pipelines and gives reliable assessment of microbial community structure. Our re-analysis of the human twin gut microbiome study [2] showed a significant correlation between microbial community composition and family, which was not previously noted. Our analysis also reveals substantial bias in community profiling employing commonly used PCR primers (in particular the V2 primer set), highlighting the need to either develop better primers system or for direct assessment of community structure by metagenomics.

## References

1. Hong, S., Bunge, J., Leslin, C., Jeon, S., Epstein, S.S.: Polymerase Chain Reaction Primers Miss Half of Rrna Microbial Diversity. *ISME. J.* 3, 1365–1373 (2009)
2. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., Egholm, M., Henrissat, B., Heath, A.C., Knight, R., Gordon, J.I.: A Core Gut Microbiome in Obese and Lean Twins. *Nature.* 457, 480--484 (2009)
3. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glöckner, F.O.: SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB. *Nucleic. Acids. Res.* 35, 7188--7196 (2007)
4. Lozupone, C., Hamady, M., Knight, R.: Unifrac--An Online Tool for Comparing Microbial Community Diversity in a Phylogenetic Context. *BMC. Bioinformatics.* 7, 371 (2006)
5. Schloss, P., Westcott, S., Ryabin, T., Hall, J., Hartmann, M., Hollister, E., Lesniewski, R., Oakley, B., Parks, D., Robinson, C.: Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* 75, 7537--7541 (2009)



## Nonlinear Model-Based Clustering for Periodically Expressed Gene Profiles

Li-Ping Tian<sup>1</sup>, Hengyao Lu<sup>3</sup>, and Fang-Xiang Wu<sup>2,3\*</sup>

<sup>1</sup>School of Information, Beijing Wuzi University,

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

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

\*Corresponding author: faw341@mail.usask.ca

**Abstract:** Cluster analysis has played an important role in analyzing gene expression data. Many distance/correlation- and static model-based clustering techniques have been applied to time-course expression data. However, these techniques did not account for the dynamics of such data and thus may not be effective for time-course gene expression data, especially those periodically expressed gene data. In this paper, we propose a nonlinear model-based clustering method for periodically expressed gene profiles. As periodically expressed genes are associated with periodic biological processes, in the proposed method it is naturally assumed that a periodically expressed gene dataset is generated by a number of periodical processes. Each process is modeled by a linear combination of trigonometric sine and cosine functions in time plus a Gaussian noise term. A two stage method is proposed to estimate the model parameter and a relocation-iteration algorithm is employed to assign each gene to an appropriate cluster. A bootstrapping method and an average adjusted Rand index (AARI) are employed to measure the quality of clustering. The results of preliminary study show that our method allows the better quality clustering than other clustering methods (e.g. k-means) for periodically expressed gene data, and thus it is an effective cluster analysis method for periodically expressed gene data.

**Keywords:** clustering, periodically expressed gene, nonlinear parameter estimation, average adjusted rand index

### I. INTRODUCTION

Many biological processes such as cell-cycle division exhibit periodic behaviors. Gene expression profiles associated with these periodic biological processes exhibits periodic. Cluster analysis on periodically expressed gene could help understand the molecular mechanism of periodic biological processes. In past decades, a number of clustering methods have been proposed for cluster analysis on gene expression data. These include distance/correlation-based clustering methods (e.g., hierarchical clustering,<sup>1</sup> k-means clustering,<sup>2</sup> and self-organizing maps<sup>3</sup>) and static model-based clustering methods.<sup>4,5</sup> In these methods, gene expression profiles are viewed as multi-dimensional vectors. Distance/correlation-based clustering methods cluster genes based on the distance/correlation among their expression profiles. Static model-based clustering methods assign genes to one of clusters if their expression profiles may be generated by a multivariate normal distribution. These methods do not take the

dynamic of time-course gene expression data and thus are not efficient for periodically expressed gene data.

Recently, some dynamic model-based clustering methods have been proposed to analyze time-course gene expression data<sup>6,7</sup>. These methods employ autoregressive models to describe the dynamics of time-course gene expression data. As periodically expressed genes are associated with periodic biological processes, it is natural to model a periodically expressed gene data by periodic (nonlinear) function. This paper proposes a nonlinear model based method for clustering periodically expressed genes from their time-course expression profiles.

## II. METHODS

### 2.1 Model for periodically expressed gene profiles

Let  $x(t)$  ( $t=1,2,\dots, m$ ) be a time-course gene expression profile generated from a periodical biological process, where  $m$  is the number of time points at which gene expression is measured. After shifting the mean of gene expression profiles to 0, the periodicity of this time-course gene expression profile can be model by a linear combination of trigonometric sine and cosine functions in time plus a Gaussian noise term as follows [8]

$$x(t) = a \cos(\omega t) + b \sin(\omega t) + \varepsilon(t) \quad (1)$$

Where  $a$  and  $b$  are the coefficients of sine and cosine function, respectively;  $\omega$  is the frequency of periodic expression data; and  $\varepsilon(t)$  represent random errors. This study assumes that the errors have a normal distribution independent of time with the mean of 0 and the variance of  $\sigma^2$ . This model is equivalent to sinusoidal function model [9-14]

$$x(t) = A \sin(\omega t + \Phi) + \varepsilon(t) \quad (2)$$

which are widely used to generate the synthetic periodic gene expression profiles [9] and to detect the periodically expressed genes[10-15]. In model (2),  $A = \sqrt{a^2 + b^2}$  is called magnitude and  $\Phi = \arctan(a / b)$  is called the phase.

Given a time-course gene expression profile  $x(t)$  ( $t=1, 2,\dots, m$ ), estimating parameters  $a$ ,  $b$  and  $\omega$  in model (1) is a nonlinear estimation problem as  $\omega$  is nonlinear in the model. In general, all nonlinear optimization programs can be used to estimate parameters in model (1), for example, Gauss-Newton iteration method and its variants such as Box-Kanemasu interpolation method, Levenberg damped least squares methods, and Marquardt's method [16]. However, these iteration methods are sensitive to initial values. Another main shortcoming is that these methods may converge to the local minimum of the least squares cost function, and thus cannot find the real values of the parameters.

Our observation is that noise free model (1)

$$x(t) = a \cos(\omega t) + b \sin(\omega t) \quad (3)$$

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

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

and that  $\omega^2$  is linear in equation (4) which is independent of a and b. Therefore, we propose the following two-step parameter estimation methods to estimate parameters a, b and  $\omega$  in model (2):

**Step1:** numerically calculate the second derivative of  $x(t)$ . Then based on equation (4), use linear least squares method to estimate parameter  $\omega^2$ . In details, let

$$X2 = [\ddot{x}(1), \dots, \ddot{x}(l)] \quad \text{and} \quad X1 = [x(1), \dots, x(l)]$$

then by the least squares method  $\omega^2$  is estimated as

$$\hat{\omega}^2 = X1^T X2 / X1^T X1 \quad \text{and} \quad \hat{\omega} = \sqrt{\hat{\omega}^2} \quad (5)$$

as time-course gene expression data are discrete, the second derivative  $\ddot{x}(t)$  is estimated by the central finite difference formula as follows

$$\ddot{x}(t) = \frac{x(t+1) + x(t-1) - 2x(t)}{\Delta^2} \quad \text{for } t=2, \dots, m-1 \quad (6)$$

where  $\Delta$  is time difference between two consecutive gene expression data points. From equation (7), the length of vectors X2 and X1 is m-2.

**Step2:** Substitute the estimated value of  $\omega$  into equation (2). Apply the maximum likelihood method to model (1) to estimate parameters a and b. In detail, let

$$X = [x(1), \dots, x(m)] \quad \text{and} \quad A = \begin{bmatrix} \cos(\Delta\hat{\omega}), \dots, \cos(m\Delta\hat{\omega}) \\ \sin(\Delta\hat{\omega}), \dots, \sin(m\Delta\hat{\omega}) \end{bmatrix}$$

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

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (AA^T)^{-1} (AX^T) \quad (7)$$

## 2.2 Nonlinear model-based clustering

*The mixture model:* In this study, it is assumed that a time-course gene express dataset is a collection of periodically expressed gene profiles which belongs to several clusters and profiles in each cluster can be described by model (1) or (2) with different parameters. Let  $\theta_k = [a_k, b_k, \omega_k, \sigma_k^2]$  be parameters of model (1) for the  $k^{\text{th}}$  cluster. Then the task of nonlinear model-based clustering is: for a given number of cluster K, divide a time-course gene expression dataset into a partition  $C = \{C_1, \dots, C_k, \dots, C_K\}$  using model (1) with parameters  $\theta_k = [a_k, b_k, \omega_k, \sigma_k^2]$  ( $k = 1, \dots, K$ ) which minimize

$$f(C | \Theta) = \sum_{k=1}^K \sum_{x \in C_k} \sum_{i=1}^m [x(i) - a_k \cos(i\Delta\omega_k) - b_k \sin(i\Delta\omega_k)]^2 \quad (8)$$

where the parameters  $\Theta$  consists of  $\{\theta_k, k = 1, \dots, K\}$ .

*Estimation of model parameters:* According to the parameter estimation method proposed in previous section for a single time-course expression profiles, for the  $k^{\text{th}}$  cluster parameters  $\theta_k = [a_k, b_k, \omega_k, \sigma_k^2]$  can be estimated as:

$$\hat{\omega}_k^2 = \sum_{x \in C_k} X1^T X2 / \sum_{x \in C_k} X1^T X1 \quad \text{and} \quad \hat{\omega}_k = \sqrt{\hat{\omega}_k^2} \quad (9)$$

$$\begin{bmatrix} \hat{a}_k \\ \hat{b}_k \end{bmatrix} = \frac{1}{|C_k|} \sum_{x \in C_k} (AA^T)^{-1} (AX^T) \quad (10)$$

$$\hat{\sigma}_k^2 = \frac{1}{m|C_k|} \sum_{x \in C_k} \sum_{i=1}^m [x(i) - \hat{a}_k \cos(i\Delta\hat{\omega}_k) - b_k \sin(i\Delta\hat{\omega}_k)]^2 \quad (11)$$

where  $|C_k|$  represents the number of time series in cluster  $C_k$ ,  $\sum_{k=1}^K |C_k| = N$ .

*Algorithm:* This study employs a relocation-iteration algorithm as shown in Figure 1 to estimate the parameters such that the cost function (8) is minimized. In 2(a) of Figure 1,  $\Theta^t$  represents the estimated parameters in cost function (8) at iteration  $t$  while in 2(b), parameters  $\hat{a}_k^t, b_k^t$ , and  $\hat{\omega}_k^t$  represent the parameters of model  $k$  at iteration  $t$ .

1. Select an initial partition for given the number of clusters,  $K$ ;
2. Iteration ( $t = 1, 2, \dots$ ):
  - (a) Estimate the parameter  $\Theta^t$  based on the present partition by using Eqs. (9)–(11);
  - (b) Generate a new partition by assigning each sequence  $x$  to cluster  $k$  for which the value of  $s^2 = \sum_{i=1}^m [x(i) - \hat{a}_k^t \cos(i\Delta\hat{\omega}_k^t) - b_k^t \sin(i\Delta\hat{\omega}_k^t)]^2$  is minimum;
3. Stop if the improvement of the cost function (8) is below a given threshold, the cluster memberships of time series do not change.

Figure 1. Algorithm for nonlinear model-based clustering

### III. EXPERIMENT RESULTS AND CONCLUSION

Both the one synthetic dataset and one biological dataset are employed to investigate the performance of the proposed method. The results of preliminary study show that our method allows the better quality clustering than other clustering methods (e.g. k-means) for periodically expressed gene data in terms of average adjusted rand index. Therefore the proposed method is an effective cluster analysis method for periodically expressed gene data.

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

#### References

1. M. B. Eisen, *et al.* "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc Natl. Acad. Sci. USA* **95**, 14863-14868(1998).

2. Toronen, P., et al. (1999) Analysis of gene expression data using self-organizing maps, *FEBS Letter*, **451**, 142-146.
3. Yeung, K.Y., et al. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987.
4. Ghosh, D. and Chinnaiyan, A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275-286.
5. McLachlan, G.J., et al. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422.
6. Ramoni, M.F., et al. (2002) Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA*, **99**, 9121-9126.
7. FX Wu, WJ Zhang, and AJ Kusalik (2005): Dynamic model-based clustering for time-course gene expression data, *Journal of Bioinformatics and Computational Biology* 3: 821-836
8. FX Wu (2010): Identification of Periodically Expressed Genes from Their Time-Course Expression Profiles, ISBRA10(short paper): 12-15
9. Harmer S., Hogenesch J. B., Straume M., Chang H. S., Han B., Zhu T., Wang X., Kreps J. A., and Kay S. A.: Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* 290 (2000) 2110-2113
10. Wichert S., Fokianos K. and Strimmer K.: Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* 20 (2004) 5-20
11. Chen J.: Identification of significant period genes in microarray gene expression data. *BMC bioinformatics* 6 (2005) 286
12. Glynn E. F., Chen J., and Mushegian A.R.: Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle Periodograms. *Bioinformatics* 22 (2006) 310-316
13. Chen J. and Chang K. C.: Discovering Statistically significant period Gene expression. *International Statistical Review* 76 (2008) 228-246
14. Liew A. W. C., Law N. F. Cao X. Q., and Yan H.: Statistical power of fisher test for the detection of short periodic gene expression profiles. *Pattern Recognition* 42 (2009) 549-556
15. Spellman, P.T., et al.: Comprehensive identification of cell cycle-regulated genes of the Yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* 9 (1998) 3273-3297
16. Beck J.V. and Arnold K. J.: *Parameter Estimation in Engineering and Science*. New York: John Wiley & Sons, 1977

## Study of the Xylose Isomerase Reveals Certain Fingerprint of Its Evolution

Yixiang Shi<sup>1</sup>, Yuanyuan Li<sup>1,2</sup>

<sup>1</sup> Shanghai Center for Bioinformation Technology,  
Shanghai, 200235, P. R. China

<sup>2</sup> Bioinformatics Center, Key Laboratory of Systems Biology, Shanghai Institutes for  
Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P.R. China.

{yxshi, yyli}@scbit.org

**Abstract.** Xylose isomerase plays an important role in species which can utilize xylose as the carbon source. In this study, we identified xylose isomerase gene and its TIM barrel components in the genome of *Clostridium beijerinckii* NCIMB 8052. A phylogenetic tree constructed by using the xylose isomerase genes from *C. beijerinckii* and several other species showed certain interesting findings. The origination of the TIM barrels domain proteins in *C. beijerinckii* remains mystery, which we would like to explore further.

**Keywords:** xylose isomerase, phylogenetic tree, TIM barrel, evolution.

### 1 Introduction

As a member of pentose family, xylose is commonly existed in the world. A lot of the agricultural and forestry products contain polysaccharides which can be degraded into xylose. Therefore, using xylose as the major carbon source for industrial fermentation to produce chemical products such as fuels can be both cost-saving and environmental-friendly. Another reason xylose is desirable is that it is a good complement to the glucose-based fermentation, since co-metabolism of pentose and hexose, can raise the utilization efficiency for both in the industrial production process[1].

*Clostridium beijerinckii* NCIMB 8052 is one of the microbial species which has the ability to utilize xylose degraded from 'waste' biomass to produce useful products such as ABE (acetone, butanol and ethanol). It has been shown that the complete xylose metabolism pathway existed in the genome of *Clostridium beijerinckii* NCIMB 8052[2]. This is also confirmed by our previous studies. In reality, not many species have been found to be able to use xylose as the sole carbon source. To elucidate how *Clostridium beijerinckii* NCIMB 8052 gained its ability to ferment xylose, we carry out an evolutionary study.

## 2 Results

We first annotated the genome of *Clostridium beijerinckii* NCIMB 8052, and identified xylose isomerase or its components. The result is shown in the Table 1.

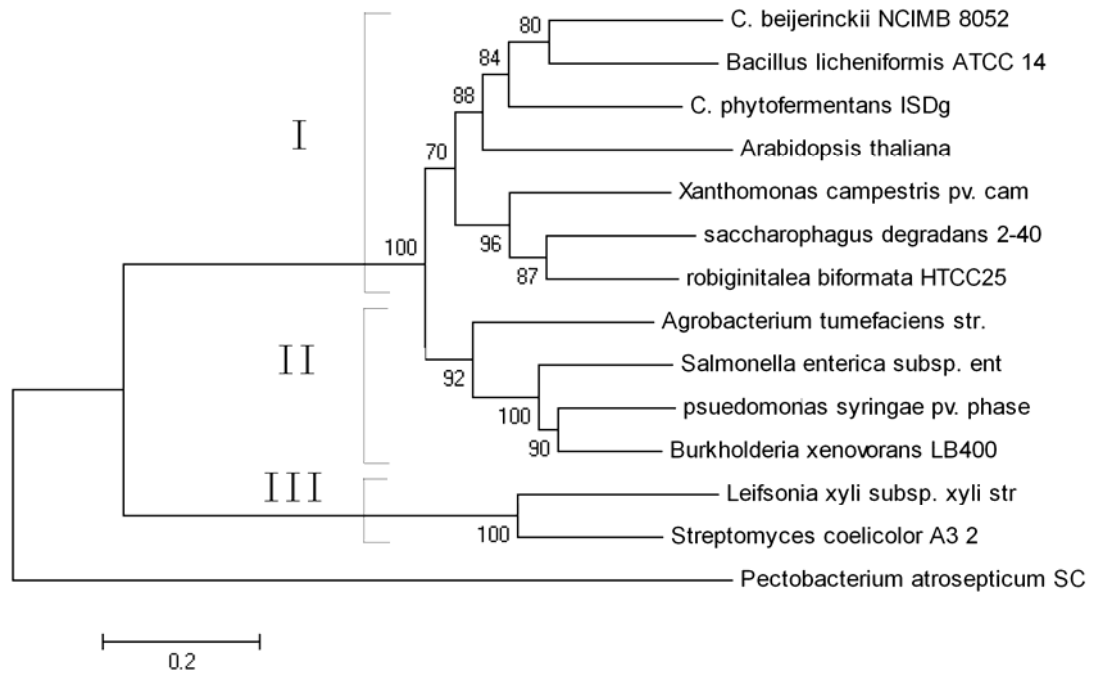
**Table 1.** The Xylose Isomerase Genes Components Identified in the *Clostridium beijerinckii* NCIMB 8052 Genome.

<i>Cbei No.</i>	<i>start</i>	<i>end</i>	<i>Strand</i>	<i>Length</i>	<i>product</i>
0450	547974	548810	+	278	xylose isomerase domain protein TIM barrel
2383	2749301	2750626	+	441	xylose isomerase
4546	5263986	5264882	-	298	xylose isomerase domain protein TIM barrel
4649	5385101	5385874	-	257	xylose isomerase domain protein TIM barrel
4842	5663859	5664830	-	323	xylose isomerase domain protein TIM barrel

Then, we use xylose isomerase genes from some representing species, and construct a phylogenetic tree (Fig. 1) using the neighbor-join method[3]. While this part is similar to a work published by other researchers two years ago[4], we believe that our result makes better biological sense. In contrast to the reference, the xylose isomerase genes from *Xanthomonas compestris*, *Saccharophagus degradans* and *Clostridium* (*C. phytofermantans* and *C. beijerinckii*) are in the same cluster. We know *Xanthomonas compestris* is a plant pathogen. And similar to *C. phytofermantans* and *C. beijerinckii*, *Saccharophagus degradans* can degrade lignocellulose and use xylose to ferment ethanol. So it is sound that they are grouped together. This is further confirmed by the fact that our tree has higher bootstrap values at the several key branch points which lead to the major differences from the two studies.

It also makes sense that the xylose isomerase gene from *C. beijerinckii* is evolutionarily close to *C. phytofermantans* and *Bacillus licheniformis*, since they are all firmicutes. The reason that *Bacillus licheniformis* is closer to *C. beijerinckii*

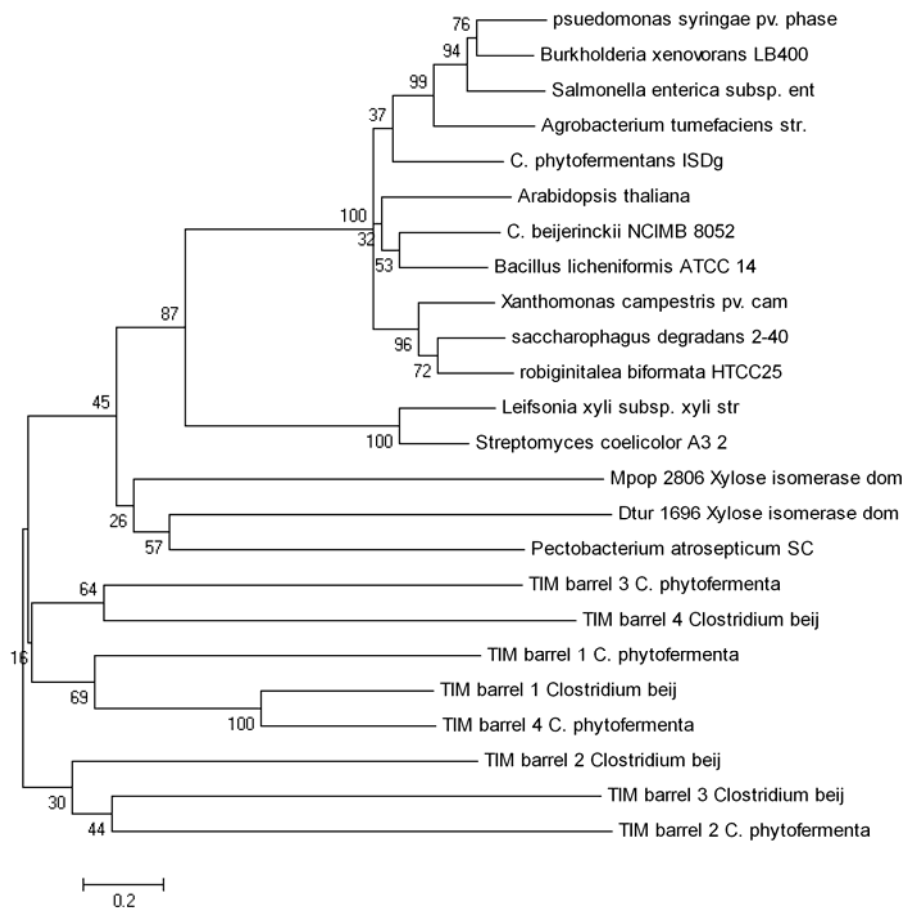
probably is another evidence that *C. beijerinckii* gained large number of outside genes through horizon gene transfer, as we suggested in earlier report[5].



**Fig. 1.** The Phylogenetic Tree Constructed using the xylose isomerase sequences from more than a dozen tested species.

Based on the figure 1, we add the sequences of the xylose isomerase TIM barrel domain proteins in *C. beijerinckii* and *C. phytofermantans*, coincidentally, 4 each, to construct a new phylogenetic tree. We found that these TIM barrel proteins all have very low homology to any of the 3 clusters of the xylose isomerase family. A lot of works are still needed to be done to solve the mystery of the origination of the TIM barrels.





**Fig. 2.** New Phylogenetic Tree with the TIM Barrel sequences from *Clostridium beijerinckii* NCIMB 8052 and *C. phytofermentans* ISDg.

## References

1. Qureshi, N., Saha, B.C., Cotta, M.A.: Butanol production from wheat straw hydrolysate using *Clostridium beijerinckii*. *Bioprocess Biosyst Eng* (2007)
2. Gu, Y., Ding, Y., Ren, C., Sun, Z., Rodionov, D.A., Zhang, W., Yang, S., Yang, C., Jiang, W.: Reconstruction of xylose utilization pathway and regulons in Firmicutes. *BMC Genomics* **11** (2010) 255
3. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4** (1987) 406-425
4. Brat, D., Boles, E., Wiedemann, B.: Functional expression of a bacterial xylose isomerase in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* **75** (2009) 2304-2311
5. Shi, Y., Li, Y.X., Li, Y.Y.: Large number of phosphotransferase genes in the *Clostridium beijerinckii* NCIMB 8052 genome and the study on their evolution. *BMC Bioinformatics* **11 Suppl 11** (2010) S9

## Viral Quasispecies Reconstruction Based on Unassembled Frequency Estimation

Serghei Mangul<sup>a</sup>, Irina Astrovskaya<sup>a</sup>, Bassam Tork<sup>a</sup>, Ion Mandoiu<sup>b</sup>, and  
Alex Zelikovsky<sup>a</sup>

<sup>a</sup>Department of Computer Science, Georgia State University, Atlanta, GA 30303  
{serghei, iraa, btork, alexz}@cs.gsu.edu

<sup>b</sup>Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269  
mandoiu@cse.uconn.edu

**Key words:** 454 pyrosequencing, expectation maximization, viral quasispecies, haplotype assembling, haplotype discovery

### 1 Introduction

The genomic diversity of RNA viruses (such as Hepatitis C virus (HCV), Human immunodeficiency virus (HIV), SARS and influenza) is a subject of the great interest since it is a plausible cause of vaccines failures and virus resistance to existing therapies. RNA lacks ability to detect and repair mistakes during replication, many mutations are well tolerated and passed down to descendants producing a family of co-existing related variants of the original viral genome referred to as *quasispecies* [4, 14, 11]. Knowing the sequences of the most virulent variants can help in the design of effective drugs [3, 13] and vaccines [7, 5] by targeting particular viral genome *in vivo*. This paper is devoted to the following problem.

**Quasispecies Spectrum Reconstruction (QSR) Problem.** *Given a collection of 454 pyrosequencing reads taken from a sample quasispecies population, reconstruct the quasispecies spectrum, i.e., the set of sequences and the relative frequency of each sequence in the sample population.*

The QSR problem has been first addressed directly in [6, 15]. Eriksson et al. [6] proposed a multi-step approach consisting of genotyping error correction via clustering, haplotype reconstruction via chain decomposition, and haplotype frequency estimation via EM method with validation on HIV data. In Westbrook et al. [15], the focus is on haplotype reconstruction via transitive reduction, overlap probability estimation and network flows with application to simulated HCV data. Recently the results of applications of the software tool ShoRAH [16] to HIV virus have been published in [17]. A novel combinatorial method have been also applied to HIV and HBV data with similar to ShoRAH results [12]. Finally, in [10] we have proposed a novel algorithm **Viral Spectrum Assembler (ViSpA)**.

Our contributions include (1) a novel **Haplotype Discovery** algorithm HapDis which adds to set of candidate strings a virtual string which emits all reads that do not fit well to candidate strings, (2) combining ViSpA with HapDis allowing ViSpA preferably assemble reads attributed by HapDis to the virtual string.

## 2 Haplotype Discovery

### 2.1 Maximum Likelihood Model

Maximum likelihood model includes a panel and an instance of sequencing machine run consisting of read spectrum i.e. the set of reads and the relative frequency of each read.

Let us define panel to be consisting of (1) a set of candidate strings (e.g. obtained from existing databases or assembled from reads) that are believed to emit the reads and (2) a weighted match between reads and strings, where weight is calculated based on the mapping of the reads to the strings.

The possible gaps in the maximum likelihood model include (a) erroneous reads (caused by genotyping errors), (b) an incorrect list of candidate strings (absence of candidates caused by gaps in current databases and presence of chimeric candidates), (c) an inaccurate read-to-string match and, finally, (d) a non-uniform emitting of reads by strings. Since the genotyping quality is improving we focus on the incompleteness of the panel, i.e. list of candidate strings.

**Haplotype Discovery Problem.** *Given read spectrum and a panel, i.e. set of candidate strings, weighted match between reads and strings, find strings missing from the panel.*

We measure the model quality by the deviation between expected and observed read frequencies as follows:

$$D = \frac{\sum_j |o_j - e_j|}{|R|},$$

where  $o_j$  is observed read frequencies,  $e_j$  - expected read frequencies and  $R$  is number of reads.

Expected read frequencies are calculated based on maximum likelihood frequencies estimations of strings and weighted match between reads and strings as follows:

$$e_j = \sum_i \frac{h_{i,j}}{\sum_l h_{i,l}} f_j^{ML},$$

where  $h_{i,j}$  is weighted match based on mapping of read  $r_j$  to string  $s_i$ ,  $f_j^{ML}$  - maximum-likelihood frequency of candidate string.

### 2.2 ML estimates of string frequencies

Maximum-likelihood estimates of string frequencies are calculated by the Expectation Maximization algorithm.

First, we create a bipartite graph  $G = \{S \cup R, E\}$  such that each candidate string is represented as a vertex  $s \in Q$ , and each read is represented as a vertex  $r \in R$ . With each vertex  $s \in Q$ , we associate unknown frequency  $f_s$  of the candidate string. And with each vertex  $r \in R$ , we associate observed read frequency  $o_r$ . Then for each pair  $s_i, r_j$ , we add an edge  $(s_i, r_j)$  weighted by probability of string  $s_i$  to emit read  $r_j$  with  $m$  genotyping errors:

$$h_{s_i,r_j} = \binom{l}{m} (1 - \epsilon)^{l-m} \epsilon^m,$$

where  $l$  is length of read sequence, and  $\epsilon$  is the genotyping error rate.

EM algorithm starts with the set of  $N$  strings. For each string we denote by  $f_s$  its(unknown) frequency. After initializing frequencies  $f_{s_{q \in Q}}$  at random, the algorithm repeatedly performs the next two steps until convergence:

- E-step: Compute the expected number  $n(j)$  of reads that come from string  $i$  under the assumption that string frequencies  $f(j)$  are correct, based on weights  $h_{i,j}$
- M-step: For each  $i$ , set the new value of  $f_s$  to the portion of reads being originated by string  $s$  among all observed reads in the sample

### 2.3 HapDis Algorithm

The main idea of the algorithm is to add to set of candidate strings a virtual string which virtually emits reads that do not fit well to assembled sequences.

Initially all reads are connected to the virtual string with weight  $h_{i,j} = 0$ . The first iteration finds the ML frequency estimations of candidates strings, ML frequency estimations of virtual string will be equal to 0, since all edges between virtual string and reads  $h_{v_s,j} = 0$ . Then these estimation are used to compute expected frequency of the reads according to formula Section 2.1. If the expected read frequency is less than the observed one (under-estimated), then the lack of the read expression is added to the weight of the read connection to the virtual string. For over-estimated reads, the excess of read expression is subtracted from the corresponding weight (but keeping it non-negative). The iterations are continued while the deviation between expected and observed read frequencies is decreasing by more than  $\epsilon$ .

---

#### Algorithm 1 HapDis algorithm

---

```

 $h_{i,j} = \binom{l}{m} (1 - \epsilon)^{l-m} \epsilon^m,$ 
add virtual string  $v_s$  to the set of candidate strings
initialize weights  $h_{v_s,j} = 0$ 
while D change  $i \in \epsilon$  do
  calculate  $f_j^{ML}$  by EM algorithm
   $e_j = \sum_i \frac{h_{i,j}}{\sum_l h_{i,l}} f_j^{ML}$ 
   $D = \frac{\sum_j |o_j - e_j|}{|R|}$ 
   $\delta = o_j - e_j$ 
  if  $\delta > 0$  then
     $h_{v_s,j} += \delta$ 
  else
     $h_{v_s,j} = \max\{0, h_{v_s,j} + \delta\}$ 
  end if
end while

```

---

Based on weight between virtual string and all reads it is possible to find set of reads that were not emitted by candidate strings. From this set of reads become possible to reconstruct set of strings missing from the panel. Based on the frequency of virtual string it is possible to decide if the panel is likely to be incomplete, i.e. if the virtual string frequency is larger then certain threshold then it is likely that some strings are missing from the panel. The total frequency of missing strings is estimated by frequency of virtual string.

### 3 HapDis Enhancement of VISPA

Below is the flowchart for the proposed enhancement of ViSpA. The weights on read-to-virtual-string connection obtained by HAPDIS estimate the probability of a read to be emitted by an unassembled sequence. These probabilities are fed back to ViSpA and reads with low probability (to belong to an unassembled sequence) will be assigned high weight so that s-t-paths will try to avoid using them unless s-t-connection is cut. So ViSpA will be modified accordingly. Newly assembled quasispecies (Qsps) are added to the original library of candidates and HapDis will estimate the frequency of unassembled sequences as well as estimate new read weights. The iterations of the big loop will be repeated until certain stopping condition is satisfied, e.g., there are no new quasispecies sequences or the virtual string has too small estimated frequency. Then final EM will estimate ML frequencies and output the resulted viral spectrum.

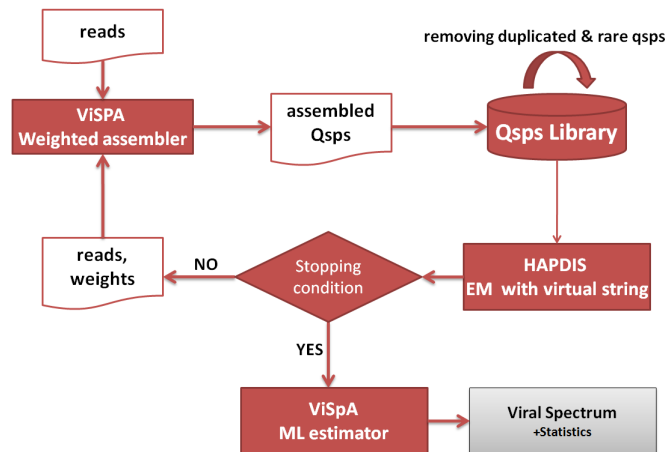


Fig. 1. Enhancement of ViSpA

## References

1. National center for biotechnology information, <http://www.ncbi.nlm.nih.gov>.
2. S. Balsler, K. Malde, A. Lanzen, A. Sharma, and I. Jonassen. Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics*, 26:i420–5, 2010.
3. N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenfuehrer, and K. Roomp et al. Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, 21:3943–3950, 2005.
4. Martinez-Salas E. Sobrino F. de la Torre J.C. Portela A. Ortin J. Lopez-Galindez C. Perez-Brena P. Villanueva N. Najera R. Domingo, E. The quasispecies (extremely heterogeneous) nature of viral rna genome populations: biological relevance a review. *Gene*, 40, pages 1–8, 1985.
5. D.C. Douek, P.D. Kwong, and G.J. Nabel. The rational design of an AIDS vaccine. *Cell*, 124:677–681, 2006.
6. N. Eriksson, L. Pachter, Y. Mitsuya, S.Y. Rhee, and C. Wang et al. Viral population estimation using pyrosequencing. *PLoS Comput Biol*, 4:e1000074, 2008.
7. B. Gaschen, J. Taylor, K. Yusim, B. Foley, and F. Gao et al. Diversity considerations in HIV-1 vaccine selection. *Science*, 296:2354–2360, 2002.
8. T. Von Hahn, J.C. Yoon, H. Alter, C.M. Rice, B. Rehermann, P. Balfe, and J.A. Mckeating. Hepatitis c virus continuously escapes from neutralizing antibody and t-cell responses during chronic infection in vivo. *Gastroenterology*, 132:667–678, 2007.
9. Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M. Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, 09 2009.
10. Astrovskaya I., B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe, and Zelikovsky A. Inferring viral spectrum from 454 pyrosequencing reads. *BMC Bioinformatics (submitted)*.
11. M. Eigen M, J. McCaskill, and P. Schuster. The molecular quasi-species. *Adv Chem Phys*, 75:149–263, 1989.
12. Mattia Prosperi, Luciano Prosperi, Alessandro Bruselles, Isabella Abbate, Gabriella Rozera, Donatella Vincenti, Maria Solmone, Maria Capobianchi, and Giovanni Ulivi. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics*, 12(1):5+, 2011.
13. S-Y. Rhee, T.F. Liu, S.P. Holmes, and R.W. Shafer. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput Biol*, 3:e87, 2007.
14. Holland J.J. Steinhauer, D.A. Rapid evolution of rna viruses. *Annual Review of Microbiology*, 41, pages 409–433, 1987.
15. K. Westbrooks, I. Astrovskaya, D. Campo, Y. Khudyakov, P. Berman, and A. Zelikovsky. HCV quasispecies assembly using network flows. In *Proc. ISBRA*, pages 159–170, 2008.
16. Osvaldo Zagordi, Lukas Geyrhofer, Volker Roth, and Niko Beerenwinkel. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *Journal of computational biology : a journal of computational molecular cell biology*, 17(3):417–428, March 2010.
17. Osvaldo Zagordi, Rolf Klein, Martin Dumer, and Niko Beerenwinkel. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Research*, 38(21):7400–7409, 2010.

## Comprehensive analysis of promoter features related to tissue-specific genes in rice

Wen-Chi Chang<sup>1,¶</sup> and Ying-Chi Wen<sup>1</sup>

<sup>1</sup> *Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan*

¶To whom correspondence should be addressed:

Wen-Chi, Chang, Ph.D., College of Biosciences and Biotechnology, Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan, ROC

Phone: +886-6-2757575 Ext. 65670; E-mail: [sarah321@mail.ncku.edu.tw](mailto:sarah321@mail.ncku.edu.tw)

### ABSTRACT

A set of genes in the same tissue might have similar expression profiles and participate in similar functions during different plant developmental stages. These tissue-specific genes are believed to be regulated by a similar set of transcription factors, and supposed that they contain similar regulatory patterns in their promoters. Consequently, we combined computational methods with experimental data to comprehensively analyze promoter features related to tissue-specific genes in rice. Firstly, microarray data from different rice developmental stages are used to identify tissue-specific expression genes. As expression data we used the microarray expression data of GSE19024 from NCBI, which includes data for 39 tissues of the rice plant from two varieties, Zhenshan 97 and Minghui 63. According to transcriptome analysis by Wang et al., we group organs into 10 clusters. We defined genes as 'tissue-specific' when the Z-score exceeded the threshold value of 2.5. Then, several motif search methods are employed to discover tissue-specific structures in those gene promoters. Based on our results, several tissue(organ)-specific motif were identified in each tissues. Many of the structural patterns are corresponding to important transcription factors. Furthermore, numerous unknown motifs could be found in the promoters of tissue(organ)-specific genes. These tissue-specific motifs might be a novel transcription factor binding sites and play critical roles during rice development.

**Keywords:** tissue-specific promoters, gene regulation, transcription factors

### INTRODUCTION

The regulation of gene expression is dynamic under different developmental stages in plants. Some genes are expressed in a special time, specific tissue, and particular condition. Therefore, identifying a suit of genes expressed with temporal and spatial is an important issue for understanding the specification of morphology and physiology in the tissue or organ systems. DNA microarray high-throughput technology has been widely used to study transcriptional expression pattern at whole-genome scale in the past decade. Several studies used this approach to investigate dynamic gene expression profile in various developmental processes in plants [1-3]. Those provide valuable information about which gene group play critical rules in which developmental stages or tissues. For example, Wang et al identify that 2667 probe sets differentially expressed among four stages of panicle development and indicate RFL and LAX play an essential role for determining of rice inflorescence architecture [3]. Furthermore, it is well known that investigation into transcription factors (TFs) and their corresponding *cis*-acting elements in promoters have attracted much attention from researchers of gene regulation. Therefore, understanding of transcription factors (TFs) and their binding sites in promoters is a key point to study the regulation of transcription. In order to response to different physiological environments during developmental phases, the induction or repression of particular genes is primarily controlled by recognition and binding of TFs to *cis*-regulatory elements in the gene promoter regions [4]. A number of studies also indicate a group of co-expressed genes under specific status are regulated by particular transcription factors [5-7]. As Skirycz et al demonstrate DOF transcription factor OBP1 controlling numerous genes in cell cycle during *Arabidopsis* development [7]. Although a

large number of studies have been made on identification of co-occurrence regulatory motifs in co-expressed gene promoters, little is known about the tissue-specific structure patterns in plant promoters. Furthermore, researchers have focused primarily on *Arabidopsis*, very little attention is given to other plants. Rice is one of the most important food crops worldwide and also a model for genomic research in cereals. Surprisingly, discussion of this kind of issue has never been examined in rice. Therefore, the purpose of this paper is to discover the tissue-specific structure patterns in rice promoters.

## IMPLEMENTATION

Figure 1 shows the system flow of this research. We combined computational methods with experimental data to comprehensively analyze promoter features related to tissue-specific genes in rice. Firstly, microarray data from different rice developmental stages are used to identify tissue-specific expression genes. As expression data we used the microarray expression data of GSE19024 from NCBI (<http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE19024>), which includes data for 39 tissues of the rice plant from two varieties, Zhenshan 97 and Minghui 63. According to transcriptome analysis by Wang et al. [3], we group organs into 10 clusters, it contains callus/germination seed, panicle, stamen, endosperm, plumule, stem, flower, root, leaf and seedling. For each gene we processed raw expression data to Z-scores, using the mean and the standard deviation of gene expression values over all tissues. We defined genes as ‘tissue-specific’ when the Z-score exceeded the threshold value of 2.5. Then, promoter regions from -2000 to +200 were extracted from TIGR V.5 database [8]. TRANSFAC version 11.0 was used to identify known transcription factor binding sites [9]. MEME [10] was employed to discover conserve sequences among tissue-specific promoters. Consequently, removing redundant position weight matrix (PWMs) by TOMTOM [11]. The tissue-specific structural patterns were then being determined.

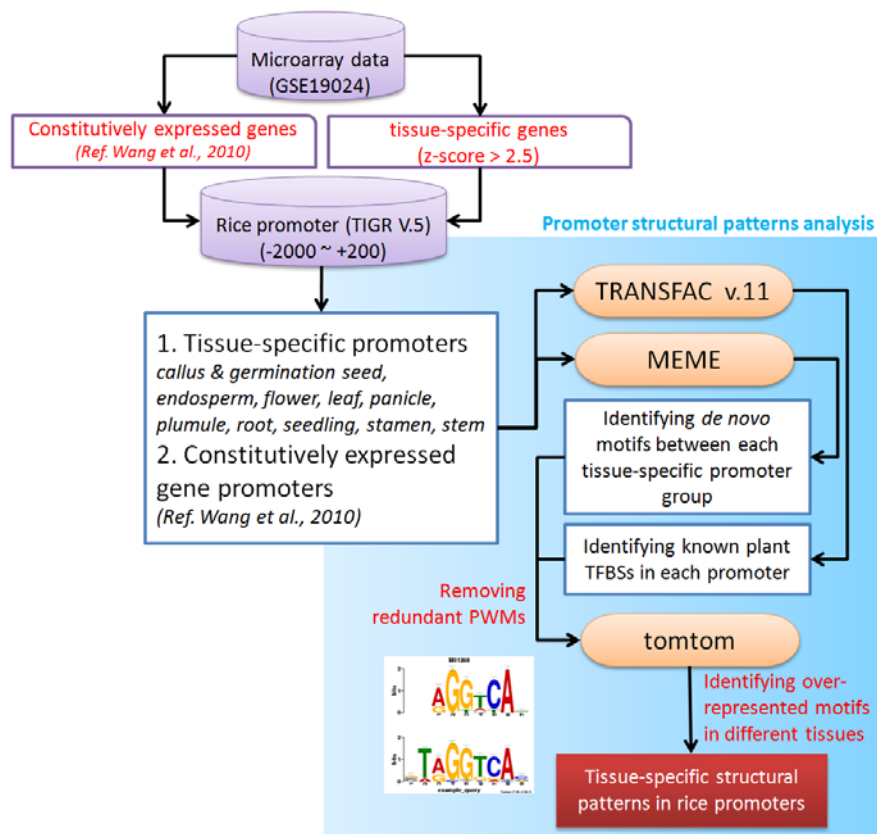


Figure 1. The system flow of this research.

## RESULTS AND DISCUSSION





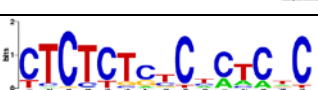

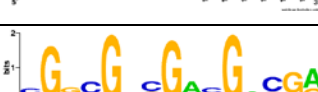



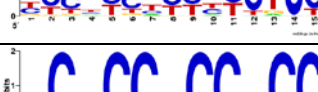



The number of tissue-specific genes in callus & germination seeds, endosperm, flower, leaf, panicle, plumule, root,



## ISBRA 2011 Short Abstracts

seedling, stamen, and stem are 109, 264, 42, 140, 138, 18, 129, 44, 1212, and 26, respectively. Table 1 shows several tissue-specific motifs identified in various tissues. It indicates numerous unknown motifs could be found in endosperm, leaf, panicle, and stamen, separately. Compare to mammalian cell, less plant specific motifs are recognized. Base on our results, those unknown motif might be candidates of transcription factor binding sites and play important roles in gene regulation.

Table 1 The motif logos of the tissue-specific structure present in different tissues

Tissue	Motif logo	Consensus sequence	Comment (E-value)
endosperm		CG[CT][CG]G[CG][CG][GTA]C[GC]C[ACG][GC][CG][GC]	CDC5, (2.80E-11)
endosperm		CTTTCCA[TC]CACATC	Unknown, (2.20E-09)
endosperm		GC[GCA]GC[GC]A[CT]G[GA][CT][GA][CAG][GC][GC]	Unknown, (4.60E-09)
endosperm		[AG][AG][AG]A[CT]GGAGG[GT]AGTA	Unknown, (1.10E-04)
leaf		[CT]TCTCT[CG][TC]C[TCA][CA][TA]C[TA]C	Unknown, (5.40E-11)
leaf		[AT][GA][CA][GA][TG][CG]A[AG]C[GA][GA]C[GA][AG]C	Unknown, (1.40E-02)
panicle		[CG]G[GCA][CA]G[AG][CG]G[AG][CAG]G[AC][CG][GC][AG]	Unknown, (6.00E-16)
panicle		C[GA]AAT[GA]TTTG[GA]ACAC	Unknown, (4.90E-04)
panicle		GTTTG[GA][AG]AA[GA]C[GA]TGC	BZR1, (5.60E-04)
stamen		[TC][CT][CT][TC][CT][CT][TC][CT][CT][TC][CT]CTCC	Alfin1, (1.3E-323)
stamen		[CG]C[TGA]CC[TA]CC[TAG]CC	Unknown, (1.30E-83)
stamen		ATGTTACTGTAGCA	Unknown, (2.20E-70)
stamen		[CG]GATCGAT[CG]GA	Unknown, (3.70E-62)
stamen		AAA[CG][AT]TT[TC]GATGTGA	Unknown, (8.70E-62)

## ACKNOWLEDGEMENTS

The authors sincerely appreciate the National Science Council of the Republic of China for financially supporting this research under Contract Numbers of NSC 99-2621-B-006 -001 -MY2 and NSC 99-2628-B-006 -016 -MY3.

## REFERENCES

- [1] V. A. Benedito, *et al.*, "A gene expression atlas of the model legume *Medicago truncatula*," *Plant J*, vol. 55, pp. 504-13, Aug 2008.
- [2] M. Schmid, *et al.*, "A gene expression map of *Arabidopsis thaliana* development," *Nat Genet*, vol. 37, pp. 501-6, May 2005.
- [3] L. Wang, *et al.*, "A dynamic gene expression atlas covering the entire life cycle of rice," *Plant J*, vol. 61, pp. 752-66, Mar 2010.
- [4] D. Walther, *et al.*, "The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*," *PLoS Genet*, vol. 3, p. e11, Feb 9 2007.
- [5] A. Chawade, *et al.*, "Putative cold acclimation pathways in *Arabidopsis thaliana* identified by a combined analysis of mRNA co-expression patterns, promoter motifs and transcription factors," *BMC Genomics*, vol. 8, p. 304, 2007.
- [6] D. W. Kim, *et al.*, "Functional conservation of a root hair cell-specific cis-element in angiosperms with different root hair distribution patterns," *Plant Cell*, vol. 18, pp. 2958-70, Nov 2006.
- [7] A. Skirycz, *et al.*, "The DOF transcription factor OBP1 is involved in cell cycle regulation in *Arabidopsis thaliana*," *Plant J*, vol. 56, pp. 779-92, Dec 2008.
- [8] S. Ouyang, *et al.*, "The TIGR Rice Genome Annotation Resource: improvements and new features," *Nucleic Acids Res*, vol. 35, pp. D883-7, Jan 2007.
- [9] V. Matys, *et al.*, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Res*, vol. 34, pp. D108-10, Jan 1 2006.
- [10] T. L. Bailey, *et al.*, "MEME: discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Res*, vol. 34, pp. W369-73, Jul 1 2006.
- [11] T. L. Bailey, *et al.*, "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Res*, vol. 37, pp. W202-8, Jul 1 2009.

## Extended and robust protein sequence annotation over conservative non hierarchical clusters.

Damiano Piovesan<sup>1</sup>, Pier Luigi Martelli<sup>1</sup>, Piero Fariselli<sup>1</sup>, Andrea Zauli<sup>2</sup>,  
Ivan Rossi<sup>2</sup> and Rita Casadio<sup>1</sup>.

<sup>1</sup> Bologna Biocomputing Group, Bologna Computational Biology Network,  
University of Bologna, Italy; <sup>2</sup>BioDec srl, Bologna, Italy  
{piovesan, gigi, piero,casadio}@biocomp.unibo.it; {andrea, ivan}@biodec.com

**Abstract.** Genome annotation is one of the most important issues in the genomic era. The exponential grow rate of newly sequenced genomes and proteomes urges the development of fast and reliable annotation methods, suited to exploit all the information available in curated data bases of protein sequences and structures. To this aim we developed BAR+, the Bologna Annotation Resource (available at <http://bar.biocomp.unibo.it/bar2.0>). The basic notion is that sequences with high identity value to a counterpart can inherit the same function/s and structure, if available. As a case study we describe how the ATP-binding domain of the ABC transporters can be found and modeled in over 30,000 new sequences not annotated before.

**Keywords:** protein functional annotation; protein structural annotation; cross genome comparison; distantly related homolog; profile HMMs; ATP-binding domain; ABC transporters.

### 1 Introduction

As a result of large sequencing projects, data banks of protein sequences and structures are growing rapidly. The number of sequences is however orders of magnitude larger than the number of structures known at atomic level and this is so in spite of the efforts in accelerating processes aiming at the resolution of protein structure. Tools have been developed in order to bridge the gap between sequence and protein 3D structure, based on the notion that information is to be retrieved from the data bases and that knowledge-based methods can help in approaching a solution of the protein folding problem [1]. The problem of computing the protein 3D structure starting from sequence is presently classified as easy to be solved, difficult albeit with a putative solution, “ab initio” and therefore very difficult. It depends on the level of sequence identity that the target sequence has with proteins already solved with atomic details in the Protein Data Bank (PDB). When a template with a high level of sequence identity to the target at hand exists, then the protein folding problem can be routinely solved by assigning with different optimization procedures the atomic coordinates of the template to the target. However when sequence identity falls in the

twilight region ( $\leq 30\%$  of sequence identity), then different heuristic procedures may help in finding putative folds for the target. The process may or may not lead to a successful solution, depending on different assumptions and strategies, including alignments among predicted features [2]. Finally, “ab initio methods” (based on first principles) are still under developments and far from being useful when searching for a putative model [1].

This work describes a recent non-hierarchical clustering procedure that was implemented with the specific purpose of fully exploiting the present knowledge in the data bases of sequences, structures and functions. This procedure largely increases the number of sequences that can be annotated by annotation transfer in a set of 988 genomes [3]. When in a given cluster distantly related sequences from different genomes coexist, the procedure allows a safe transfer of annotation both for structure and function, independently of the level of sequence identity. As a case study we analyze the cluster that includes the largest number of sequences (87,893) mainly from Prokaryotes. Some 30,000 sequences without annotation inherit validated functional and structural annotation from the cluster. The cluster includes the ATP-binding domain of the ABC transporters.

## 2 The Bologna Annotation Resource

A previous version of our method was already described and validated [3] and here we use a complementary and independent annotation resource recently developed (BAR+). Similarly to BAR, BAR+ is also a non hierarchical clustering method relying on a comparative large-scale genome analysis. The method is based on a non hierarchical clustering procedure characterized by a stringent metric that ensures a reliable transfer of features within clusters. The basic notion is that sequences with high identity value to a counterpart can inherit the same function/s and structure, if available. What is totally new in our analysis is to cluster sequences with the constraint that sequence identity should be equal or higher than 40% on at least 90% of the pair wise alignment length. By this sequences are clustered in sets that can be annotated in terms of function and structure depending on the annotation type of the sequences within the cluster. Our method starts with an all-against-all BLAST alignment [4] of all the sequences in a GRID environment (within Comput-Er; <http://www.comput-er.it/>). The alignments are then regarded as an undirected graph; after the clustering procedure that constrains both the sequence identity value and the alignment length, all the connected nodes (proteins) collapse into a single group (cluster). A cluster that incorporates a UniProt entry inherits its annotations (GO terms, PDB structures, SCOP classifications, Pfam families, when available). GO and Pfam features in the clusters are validated by computing a P-value. Clusters can contain distantly related proteins that by this can be annotated with high confidence, after the statistical validation procedure of GO and Pfam terms (P-value  $< 0.01$ , [3]). Ultimately the method analyses a total of over 13 million protein sequences including 988 genomes and UniProt release 2010\_05.

### 3 Cluster analysis

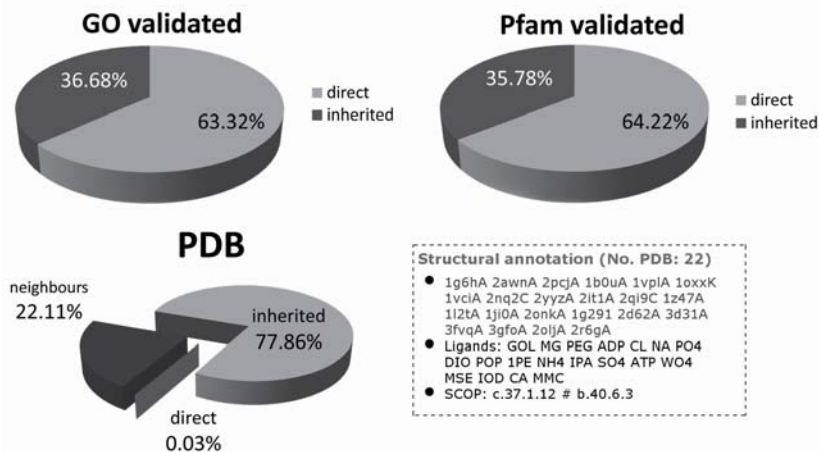
70% of the whole data set of sequences falls into 913,962 clusters. The remaining 30% originate singletons (containing just one sequence). Well annotated sequences are characterized by all the functional and structural annotations derived from UniProt entries. Ligands are also listed when present in their PDB file/s. When a well annotated sequence falls into a cluster, their annotation characterizes the cluster given our stringent criteria of cluster generation. With this procedure, when hypothetical and/or putative proteins fall into an annotated and validated cluster, they can safely inherit GO terms and Pfam domain/s even in the case of very low sequence identity with the well annotated proteins. By this they can be labeled as distantly related homologs and inherit function and structure in a validated manner. We previously discussed that this procedure can increase the level of annotation when compared to that of UniProt [3]. When PDB templates are present within a cluster (with or without their SCOP classification), profile HMMs are computed on the basis of sequence to structure alignment [5] and are cluster-associated (Cluster-HMM). Our system is endowed with a library of 10,858 HMMs for aligning even distantly related sequences to a given PDB template/s. BAR+ allows three main categories of annotation: PDB [with or without SCOP (\*)] and GO+/Pfam; PDB (\*) without GO+/Pfam; GO+/Pfam without PDB (\*) and no annotation. Each category can further comprise clusters where GO and Pfam functional annotations are or are not validated.

Our BAR+ contains 207,371 clusters that allow transfer of validated annotation terms by aligning new target sequences. Inheritance is possible provided that the target is 40% identical on at least 90% of the pair wise alignment length to any of the sequences in the validated clusters.

### 3 A case study: the ATP-binding domain of ABC transporters

The most populated cluster of BAR contains 87,893 protein sequences with only 69 sequences from Eukaryotes (average protein length  $(281 \pm 16\%)$  residues). The cluster contains 22 PDB structures from Prokaryotes (the Root Mean Square Deviation (RMSD) of the backbone of all structures is 0.189 nm with 0.039 nm Standard Deviation (SD)). Some 56,448 sequences, including only 44 from Eukaryotes (Plantae and algae) are endowed with 292 GO terms and 11 Pfam. After statistical validation the systems lists 73 GO terms (55 Molecular Function; 14 Biological Processes; and 6 Pfam terms) ( $P$  value  $< 0.01$ ). The most frequent and validated Pfam term (carried along by the largest number of sequences) is ABC\_tran (PF00005), corresponding to the ATP binding domain of the ABC transporters. ABC transporters belong to the ATP-Binding Cassette superfamily, involved in the export and import of a wide variety of substrates ranging from small ions to macromolecules. With our procedure the remaining 31,445 sequences of the cluster inherit by transfer all the validated GO and Pfam terms. These comprise 25 sequences from Eukaryotes not annotated before, including 22 from *Xenopus tropicalis*, the only animal in the cluster. The cluster-HMM is a model based on the structural [6] and multi-alignment of sequences with 40% sequence identity to the templates. This can be adopted to align distantly related

sequences to the templates for structural model building. Within the cluster, some 50,000 sequences are less than 30% identical to the templates and with this procedure they can be endowed with a structural model. Cluster annotation details are shown in Fig.1.



**Fig. 1.** Annotation by inheritance of the ATP binding domain of the ABC transporters. The cluster, containing 87,893 protein sequences is endowed with 22 PDB templates, 6 Pfam and 73 GO validated terms. The percentage of sequences with a direct UniProt annotation is shown together with that inheriting the validated annotation within the cluster. Over 50,000 sequences with low homology ( $\leq 30\%$ ) to the templates can be modeled with the Cluster HMM (see text for details). The inset lists PDB codes of the templates, ligands and SCOP classification.

**Acknowledgments.** The authors would like to thank INFN (Istituto Nazionale di Fisica Nucleare) and CNAF (Centro Nazionale per la Ricerca e Sviluppo delle Tecnologie Informatiche e Telematiche) for support in GRID computing.

## References

1. Lesk A. Introduction to protein science. Oxford University Press, London (2010)
2. Fariselli, P., Rossi, I., Capriotti, E., Casadio, R. The WWWH of remote homolog detection: the state of the art. *Brief Bioinform.* 2, 78--87 (2007)
3. Bartoli, L., Montanucci, L., Fronza, R., Martelli, P.L., Fariselli, P., Carota, L., Donvito, G., Maggi, G., Casadio, R. The Bologna Annotation Resource: a non-hierarchical method for the functional and structural annotation of protein sequences relying on a comparative large-scale genome analysis. *J Proteome Res.* 8, 4362--4371 (2009)
4. McGinnis, S., Madden, T.L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, 32(Web Server issue),W20--5 (2004)
5. Eddy, S.R. Profile Hidden Markov Models. *Bioinformatics* 14, 755--763 (1998)
6. Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.L. MUSTANG: A multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics* 64,559--574 (2006)

## Designing Reusable User-Interfaces for Querying a Collection of Neuroscience Ontologies

Akshaye Dhawan and Alison L. Nolan

Ursinus College, Department of Computer Science, PA 19426  
{adhawan, alnolan}@ursinus.edu

**Abstract.** This paper examines the problem of generating effective, reusable web interfaces for searching and browsing neuroscience data represented in the Web Ontology Language (OWL). We briefly explain our design of a collection of ontologies on the nervous systems of different mollusk. We then motivate our goal to design interfaces that are reusable across these different ontologies. In order to achieve this re-usability, we view the underlying semantic data in the ontology as a graph and are currently exploring the use of different graph properties to infer the structure of the class hierarchy in the ontology. The interface allows a user to query the underlying ontology without the use of a query language like SPARQL.

### 1 Introduction

In this paper we examine techniques for designing reusable user-interfaces for browsing and searching a collection of ontologies representing information on neurons, neural networks and their properties for different mollusks. These ontologies represent the data collected as part of the NeuronBank Project [1, 2]. We have represented this information using the Web Ontology Language (OWL) [3].

In order to make these ontologies accessible to the end users, the information represented in them must be readily available through intuitive web interfaces for browsing and searching this data. However, since these ontologies all differ slightly, it would be much more efficient to design these interfaces in a reusable manner such that the interface is generated dynamically from the underlying data-model.

The remainder of this paper is as follows. In Section 2 we briefly explain the design and structure of our collection of ontologies. In Section 3 we motivate the need for reusable interfaces that can query an ontology and explain our approach to developing these interfaces. Finally, we conclude in Section 4.

### 2 Ontology Design

Across all species, neurons are described by sets of attributes (e.g. neurotransmitter, spike shape etc.) and identified by delineating the subset of attributes

necessary and sufficient to reliably identify that neuron across different specimens. There is variation, however, in the attributes used to describe neurons in different species. We create ontologies for data on two different species of mollusks: *Tritonia diomedea* and *Melibe leonina*. We include information on 45 identified types of neurons and their interconnections for *Tritonia* and 4 different neurons for *Melibe* in our ontologies. The data used for this was taken from NeuronBank [4] and represented in OWL (NeuronBank uses Protege-Frames [5] to represent its data). We reuse a number of classes from other upper ontologies including the Basic Formal Ontology [6] and ontologies includes as part of the Open Biomedical Ontologies Project [7] including the Common Anatomy Reference Ontology (CARO) [8] .

Due to space constraints, we do not explain our ontologies in detail. However, we create classes derived and organized under the aforementioned upper ontologies for the major components of the nervous system. Hence, some of our classes include Neurite (with Axon and Dendrite as subclasses), Synapse and Neuron. The Neuron class is further subclasses into types of neurons. For example, for *Tritonia*, some subclasses under Neuron include Pleural\_Neuron (for all the neurons below the Pleural Ganglion of *Tritonia*), Cerebral\_Neuron (neuron in the cerebral ganglion), Pedal\_Neuron (neurons in the pedal ganglion) etc. Notice that the classification for this species is based on brain region.

## 3 User-Interface Design

### 3.1 Motivation

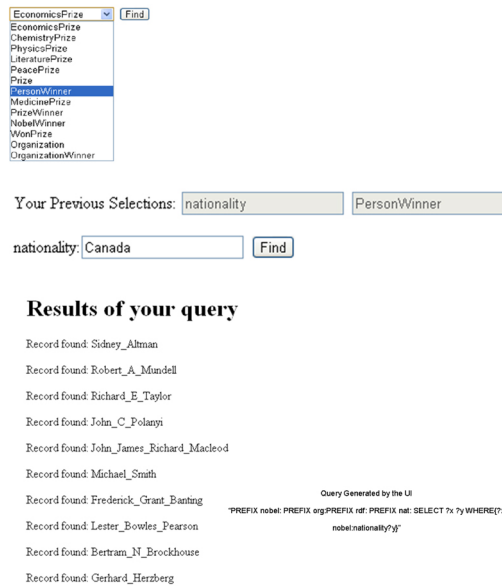
While much research effort has gone into developing standards (like RDF and OWL) for representing information on the Semantic Web and for developing languages to query this information (SPARQL), little has been done to make this information accessible to an end-user not familiar with query languages. For our application, we developed a number of ontologies on closely related mollusk nervous systems. The intended end-users of these ontologies are neuroscientists working on these species and they do not have the training to write queries in SPARQL.

Additionally, despite having a number of concepts in common, each of these ontologies differed in species specific ways. Also, it is envisioned that as new lab techniques to identify this neurons are introduced, the underlying ontologies may evolve over time. This led us to our goal of developing a simple web interface to access the ontology that is capable of generating SPARQL queries for an end-user and presenting the results to these queries. Additionally, the interface should be dynamically generated from the underlying ontology, thereby requiring no changes when the ontology is modified and allowing reuse across different ontologies. The work presented in this paper extends our previous work presented as a poster [9].



## 3.2 Approach

Our approach to generating an interface is based on presenting the user with an upper-level menu that list the most relevant classes. As explained shortly, we infer the most relevant classes based on some properties of the data model graph. Once the user selects a class, the next menu is populated based on retrieving the properties of the class the user selected. As the user interacts with the ontology and makes selections, we build behind the scenes a SPARQL query representing the users choices as filters on the ontology. An example of our method and the generated query is shown for a much simpler ontology - that of all Nobel Prize winners.



**Fig. 1.** The query interface for the Nobel Prize Ontology

Generating an interface from an underlying ontology is a challenging problem because ontologies are viewed as graphs with no explicit hierarchies. However, most users perceive the ontology as a hierarchical data representation containing some upper-level classes that represent a natural entry point for a user searching or browsing this ontology. For example, for the ontologies of the different nervous systems of the mollusks that we are working with, a logical starting point to begin browsing the ontology could be a neuron, a connection, or even a reference. Our goal in this research was to design a means of inferring this hierarchy that underlies most ontology. This would allow for the design of interfaces that are completely reusable since no class structure needs to tie the interface to a given ontology.

The basis of our approach consists of examining the graph of all the data triples in the ontology. A triple relates a subject to an object using a predicate. We build from this graph a data-model graph that consists of a node for every class in the ontology. We associate with every class a count of how many instances of that class exist in the ontology. We then join two nodes, representing two different classes by an edge if they are related in the original ontology by a given predicate. For each edge, we compute a weight that denotes the number of times this predicate links the given subject class to the given object class. We are currently examining a variety of heuristics that take into account the degree of the class in the data-model graph to infer if it is an upper-level class.

## 4 Conclusion

In this paper we motivate an application driven need for designing reusable user interfaces that can query a collection of neuroscience ontologies. The interface is capable of generating SPARQL queries based on the users selections and is dynamic in that it adjusts to changes in the class hierarchy of the underlying ontology. The interface also infers relevant entry point classes to the ontology based on properties of the data model for the given ontology. As part of our future work, we are exploring other algorithms to infer the class hierarchy and designing experiments to evaluate the effectiveness of the interface.

## References

1. Calin-Jageman, R., Dhawan, A., Yang, H., Wang, H.C., Tian, H., Phoungphol, P., Frederick, C., Balasooriya, J., Chen, Y., Prasad, S., Sunderraman, R., Zhu, Y., Katz, P.: Development of neuronbank: A federation of customizable knowledge bases of neuronal circuitry. In: Services, 2007 IEEE Congress on. (2007) 114–121
2. Katz, P.S., Calin-Jageman, R., Dhawan, A., Frederick, C., Guo, S., Dissanayaka, R., Hiremath, N., Ma, W., Shen, X., Wang, H.C., Yang, H., Prasad, S., Sunderraman, R., Zhu, Y.: Neuronbank: a tool for cataloging neuronal circuitry. *Frontiers in Systems Neuroscience* **4**(0) (2010)
3. (W3C), W.O.G.: Owl web ontology language, <http://www.w3.org/tr/owl-features/> (February 2004)
4. : Neuronbank, <http://neuronbank.org/>
5. Noy, N.F., Fergerson, R.W., Musen, M.A.: The knowledge model of protege-2000: combining interoperability and flexibility, Springer (2001) 17–32
6. Grenon, P., Smith, B., Goldberg, L.: Biodynamic ontology: Applying bfo in the biomedical domain
7. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* **25**(11) (November 2007) 1251–1255
8. Neuhaus, D.S.: Preprint caro: The common anatomy reference ontology
9. Dhawan, A., Nolan, A.: Poster: Designing reusable user-interfaces for browsing a collection of neuroscience ontologies. In: Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on. (2011) 240

## A new approach to the analysis of genome-wide association study (GWAS) data

Kwangsoo Kim<sup>1</sup> and Chol Shin<sup>2</sup> and Hong Seo Ryoo<sup>1\*</sup>

<sup>1</sup>School of Industrial Management Engineering,  
Korea University, Seoul, Korea

<sup>2</sup>Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine,  
Korea University Ansan Hospital, Gyeonggi-do, Ansan, Korea

**Abstract.** This paper presents a new approach to the analysis of genome-wide association study (GWAS) data. To date, most of the methods used in the analysis of GWAS data are based on the statistical approaches which deal with each single nucleotide polymorphisms (SNPs) without considering their interactions. In situations like this, we designed a new measurement for quantifying the importance of combinations of SNPs and developed a methodology for efficiently investigating combinations of SNPs and identifying significant SNPs and their combinations. On the GWAS data of the Ansung and Ansan population-based cohorts in Korea, we performed association studies of height and showed remarkable results of the proposed approach compared to those of regression analysis.

**Keywords:** genome-wide association study, single-nucleotide polymorphism, statistical analysis, case-control study, pattern generation, machine learning

### 1 Introduction

Recently advances in high-throughput genotyping technology lead to various genome-wide association studies (GWASs)(e.g. [1]) to identify and explain the association between the traits of interest and the common genetic variations. In those studies, most of the researchers applied simple statistical methods, such as linear regression or logistic regression, to analyze the large-scale GWAS data. The regression-based analysis methods evaluate the significance of each SNPs without considering the interactions between the SNPs and hence they are not suitable for dealing with complex traits influenced by various genetic and environmental risk factors. In this paper, we present a new measurement for quantifying the importance of combinations of SNPs and an approach for efficiently investigating combinations of SNPs and identifying significant SNPs and their combinations.

---

\* Corresponding author

## 2 Materials and Methods

### 2.1 Data Information

Here is a brief description of the GWAS data studied in this paper. The GWAS data originally contained 500,568 SNPs genotyped by Affymetrix Genome-Wide Human SNP array 5.0 in 10,004 genomic DNA samples from the Ansung and Ansan population-based cohorts in Korea. After quality control steps, we included 8842 DNA samples and 334,546 SNP markers in the association analysis of the height. Detailed informations are described in [2].

For the case-control association study, we defined tall height cases ( $n=1,426$ ) as height greater than 1 standard deviation from the mean and small height controls ( $n=1,373$ ) as less than 1 standard deviation from the mean.

### 2.2 Comparison with Statistical Analysis

We performed statistical analyses via SAS program (version 9.2). As stated in the previous research [3], we assessed height associations by using linear regression analysis adjusted for sex and age under additive, dominant and recessive models for all individuals ( $n=8,842$ ). After selecting 28 SNPs by p-value, we examined the SNPs in a height case-control study by using logistic regression analysis for 1,426 cases and 1,373 controls.

### 2.3 Method Development

It is natural that a SNP is considered as a risk factor when the SNP occurs frequently in the patients while its occurrences are rare in the normal. This can be extended easily from the case of single SNP to the case of combinations of SNPs. Based on this simple concept, we first scanned all single SNPs for counting their occurrences in the two groups of case and control data. After selecting 100 SNPs by the differences of occurrences between the two groups, we make all possible combinations with three SNPs and get their occurrences and differences as the single case. Finally we selected 30 combinations of SNPs with biggest difference values and applied them to the regression analysis.

## 3 Results and Conclusions

In the tables below, we proposed the results of linear regression for association and logistic regression for case-control. As we mentioned earlier, we selected 28 SNPs reported in [3] by linear regression and applied the SNPs into logistic regression analysis as a replication study. Without depending on the linear regression, we selected 30 combinations of SNPs via difference measure and applied them into linear and logistic analyses as replication studies. With these results, our proposed method showed that it identifies the combinations of SNPs whose p-values are much smaller than those of single SNPs selected by regression. The method is meritorious in that it do not need to solve the regression problem and takes less computational time.

**Table 1.** Results obtained by regression analysis

RS ID	Subjects		
	Ansung cohort	Ansan cohort	Combined cohorts
Association study	Linear Regression p-value		
Additive model			
rs6918981	1.28E-04	1.23E-04	3.11E-08
rs10513137	4.51E-05	3.90E-04	1.65E-07
rs6440003	4.06E-05	1.52E-03	5.08E-07
Dominant model			
rs10513137	3.14E-04	3.44E-04	9.13E-07
rs1520223	8.42E-04	9.38E-04	1.03E-06
rs6918981	1.03E-03	5.86E-04	1.21E-06
Recessive model			
rs11989122	3.51E-03	2.06E-05	7.00E-07
rs7032940	1.85E-03	4.58E-04	2.26E-06
rs10816937	2.75E-03	4.69E-04	3.51E-06
Case-control study*	Logistic Regression p-value		
Additive model			
rs6918981	1.45E-03	4.51E-03	2.31E-05
rs3791675	2.17E-02	1.15E-03	7.96E-05
rs7313075	1.67E-02	8.34E-03	2.34E-04
Dominant model			
rs6918981	2.68E-03	9.75E-03	1.04E-04
rs12426318	2.42E-02	4.43E-03	1.20E-04
rs7313075	1.22E-02	1.34E-02	1.85E-04
Recessive model			
rs4811971	1.47E-04	9.64E-03	7.27E-06
rs7032940	2.49E-02	7.67E-04	4.64E-05
rs7036157	4.25E-02	7.67E-04	9.20E-05

\*: replication study with the SNPs selected by the association study

**Table 2.** Results obtained by proposed method

RS ID	Subjects		
	Ansung cohort	Ansan cohort	Combined cohorts
Association study	Linear Regression p-value		
rs11880550, rs2820072, rs12507269	8.43E-08	8.02E-08	3.79E-14
rs11880550, rs1151808, rs4922981	7.29E-09	1.03E-05	7.56E-13
rs6869605, rs10766922, rs6026584	2.91E-10	2.48E-04	1.13E-12
Case-control study	Logistic Regression p-value		
rs1338638, rs5915065, rs2043183	7.13E-07	2.39E-07	6.04E-13
rs1338638, rs5915065, rs2786152	1.25E-06	2.93E-07	1.37E-12
rs1338638, rs1415757, rs2043183	8.88E-07	4.96E-07	1.75E-12

4 K. Kim and C. Shin and H.S. Ryoo

## References

1. Hindorf, L.A., Junkins, H., Hall, P., Mehta, J., Manolio, T.: A catalog of published genome-wide association studies Available at: [www.genome.gov/gwastudies](http://www.genome.gov/gwastudies).
2. Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M., Cha, S.H., Kim, J.W., Han, B.G., Min, H., Ahn, Y., Park, M.S., Han, H.R., Jang, H.Y., Cho, E.Y., Lee, J.E., Cho, N.H., Shin, C., Park, T., Park, J.W., Lee, J.K., Cardon, L., Clarke, G., McCarthy, M.I., Lee, J.Y., Lee, J.K., Oh, B., Kim, H.L.: A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genetics* **41**(5) (2009) 527–534
3. Kim, J.J., Lee, H.I., Park, T., Kim, K., Lee, J.E., Cho, N.H., Shin, C., Cho, Y.S., Lee, J.Y., Han, B.G., Yoo, H.W., Lee, J.K.: Identification of 15 loci influencing height in a korean population. *Journal of Human Genetics* **55** (2010) 27–31

## Tav4SB: grid environment for analysis of kinetic models of biological systems\*

Mikołaj Rybiński<sup>1,2</sup>, Michał Lula<sup>1</sup>, Sławomir Lasota<sup>1</sup> and Anna Gambin<sup>1,2</sup>

<sup>1</sup> Institute of Informatics, University of Warsaw

<sup>2</sup> Mossakowski Medical Research Centre, Polish Academy of Sciences

**Abstract.** Taverna Workbench eases integration of software tools for life science research in experiments expressed as workflows. The Taverna services for Systems Biology (Tav4SB) project provides a set of new Web service operations which extend the functionality of Taverna Workbench in the systems biology domain. Tav4SB operations allow to perform numerical simulations or model checking of, respectively, deterministic or stochastic semantics of biological models. To visualize the results of model analysis a flexible plotting operation is provided as well. Tav4SB operations are executed in a grid environment, integrating heterogeneous software such as Mathematica, PRISM and SBML ODE Solver. User guide, contact information and full documentation of available Web service operations, exemplary workflows and other, additional resources can be found at the Tav4SB project's Web page: <http://bioputer.mimuw.edu.pl/tav4sb/>.

**Introduction.** The Taverna Workbench [11] is a tool which facilitates the design and execution of the *in silico* experiments. Experiments are constructed as workflows which can be stored and executed when needed. The building blocks of a workflow are services, called processors. Technically, workflow is a set of processors, together with connections between their inputs and outputs. The remote processors are implemented as Web service (WS) operations. Scattered physically throughout computational resources of numerous scientific facilities, combined WSs allow to perform highly complex analyses, surpassing power of a standard workstation.

Taverna services come from a diverse set of life sciences domains. In the field of computational biology, Taverna mainly provides services related to sequence annotation and analysis. Here, we present remote processors that extend Taverna's functionality in the systems biology domain, specifically, in the analysis of kinetic models of biological systems. Our hardware base offers computational resources sufficient for computationally demanding experiments, such as multiple invocations of the model-checking procedure. Essentially, Taverna Workbench provides a convenient user interface for our WS operations. Analysis of the behavior of cellular systems under various conditions can be conducted without the need of programming own WS client.

**Main features of Tav4SB.** Mathematical framework determines the structure of the kinetic formulation for a given biochemical network model. The most common repre-

---

\* This work was partially supported by the Polish government grant N N206 356036 and by the Biocentrum Ochota project (POIG.02.03.00-00-003/09). The first author is a scholar within the Human Capital Operational Programme financed by European Social Fund and state budget.

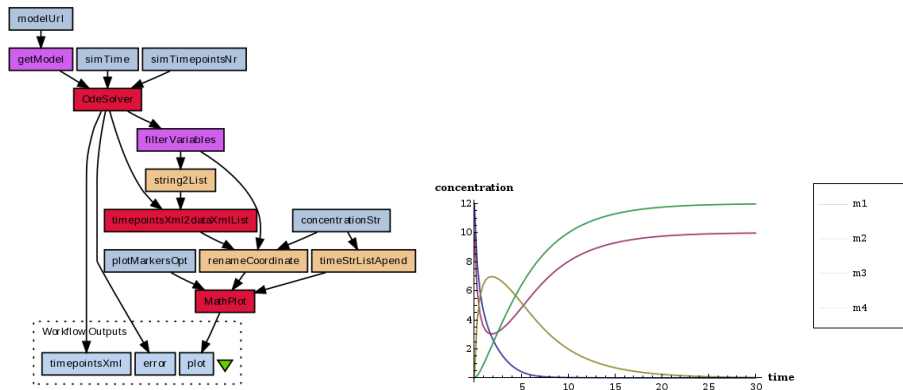
representations are ordinary differential equations (ODEs) for the deterministic framework and continuous-time Markov chain (CTMC) for the stochastic framework [1,14].

Operations provided by our Web server allow to perform: (1) numerical simulations for the deterministic formulation of the network model with use of the SBML ODE Solver library[13], (2) probabilistic model checking of Continuous Stochastic Logic (CSL) [2] formula over a CTMC with use of the PRISM tool[9], and (3) visualization of data series such as ODEs trajectories, or values of parametrized CSL properties, with use of the Mathematica tool (Wolfram Research, Inc., 2008,Version 7.0).

The SBML ODE Solver library enables numerical analysis of models encoded directly in the Systems Biology Markup Language (SBML) [10], the standard of our choice. The library employs libSBML [4] to automatically derive ODEs plus their Jacobian and higher derivatives as well as CVODES package — state of the art numerical integration library from SUNDIALS [8].

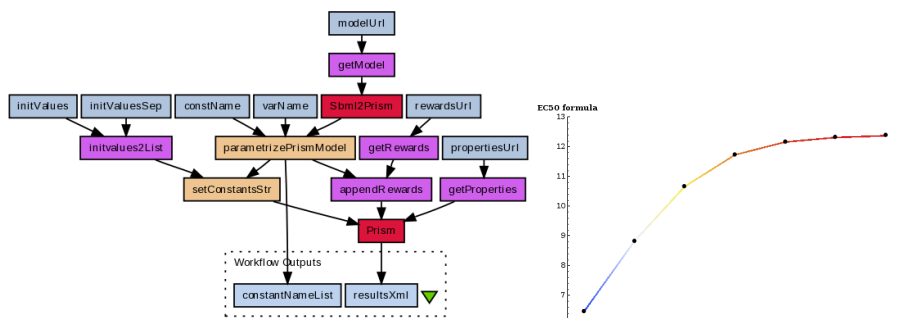
PRISM is one of the leading tools implementing probabilistic model checking, a technique of formal verification of systems that exhibit a stochastic behavior. A system to be analyzed is modeled as a Markov chain, and a correctness property is expressed in a suitable probabilistic temporal logic. Some recent works, see e.g. [7,12], demonstrate applicability of PRISM to analysis of models of biological systems. Case studies include models of cell cycle control, fibroblast growth factor signaling, and MAPK cascade. For biological applications, a CTMC (continuous-time Markov chain) is typically chosen as an underlying model, and the properties are specified in a continuous time logic, for instance in CSL. The approach seems promising as it often can yield a better understanding of the dynamics of systems to be analyzed.

Wolfram’s Mathematica has one of the most advanced graphics capabilities among computer mathematics tools. Tav4SB provides Mathematica’s two- and three-dimensional list plots together with a versatile set of options for customizing their display.



**Fig. 1.** The “Simulate SBML-derived ODEs” workflow and resulting trajectories plot for the enzymatic reaction model of [5]. The red boxes represent nested workflows, corresponding to Tav4SB WS operation wrappers and a helper. See text for more details.





**Fig. 2.** The computational part of the “Probabilistic model checking of the SBML stochastic model” workflow and the resulting plot for the stochastic version of the enzymatic reaction model. The red boxes represent nested workflows, corresponding to Tav4SB WS operation wrappers. See text for more details.

**Exemplary use cases.** We constructed a set of exemplary workflows. Their main purpose is to demonstrate usage of the Tav4SB WS operations from the Taverna Workbench client. There are two kinds of exemplary workflows: Tav4SB WS operation wrappers and exemplary *in silico* experiments.

The wrapper workflows illustrate the direct usage of Tav4SB operations in Taverna. Their purpose is to be re-used as a nested workflows, as demonstrated in two exemplary experiments described below. Additionally, we built a number of helper Taverna processors, used for interacting with XML-formatted inputs and outputs of the WS operations. Those helpers are standard Taverna’s XML splitters and local services as well as additional BeanShell scripts.

In our two exemplary experiments we have used an enzymatic reaction model with species names and parameters values from the [5]. The first workflow numerically simulates models’ ODEs and plots the results. ODEs are derived automatically from the SBML model file, based on the rate laws for described reactions. The enzymatic reaction deterministic model contains mass-action kinetics rates. As a result of running this simple experiment one gets the trajectories of the species (ODEs) variables, together with their plot. Figure 1 depicts the simulation workflow and the resulting plot for all model’s species variables over a 30 seconds time period.

The second experiment runs a probabilistic model checking for the stochastic version of the enzymatic reaction model of [5], also encoded in the SBML format. The reward-based CSL formula, which is being checked, is

$$R_{\#r1=?} [F(p > 0.5 * \lim_{t \rightarrow \infty} p(t))].$$

Roughly speaking, the formula answers the following question: how many times, on average, the enzyme-substrate complex association reaction *r1* has to occur before the amount of product *p* reaches 50% of its maximum? This corresponds to the half maximal effective concentration ( $EC_{50}$  coefficient). The formula is evaluated for different

enzyme initial amounts to find its optimal efficiency. As this is a time consuming task, and plotting usually requires many runs to fine-tune the plot parameters, the experiment is divided into two separate parts: a computational part and a plotting part. Figure 2 depicts the computational part of the workflow and the resulting plot.

The plot can be read as: if  $E(0)$  is equal to 1 then on average, before product reaches half of its maximum, each enzyme has to convert slightly more than 6 substrates. To no surprise, when  $E(0)$  is equal to 12 — the initial amount of substrate, each enzyme has to convert at most one substrate. The total, parallel enzymatic reaction system's efficiency doesn't improve significantly from that point as not much more than 12 complex formation reactions  $r1$  are needed to achieve half of the maximum product amount.

**Availability.** The definition of the operations provided by the Tav4SB WS plus exemplary workflows files, together with installation and execution instructions are available from the project's Web page: <http://bioputer.mimuw.edu.pl/tav4sb/>. Documentation of the Tav4SB WS can be found in the BioCatalogue [3], a curated catalogue of life sciences Web services. Wrappers and experiments workflows are also available from the myExperiment repository [6], together with the workflow figures.

## References

1. Aldridge, B. B. et al.: Physicochemical modelling of cell signalling pathways. *Nature Cell Biology* 8(11), 1195–1203 (2006)
2. Aziz, A. et al.: Verifying continuous time markov chains. In: Proc. 8th International Conference on Computer Aided Verification (CAV'96). LNCS, vol. 1102, pp. 269–276 (1996)
3. Bhagat, J. et al.: BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Research* 38(Web Server issue), W689–694 (2010)
4. Bornstein, B. J. et al.: LibSBML: an API library for SBML. *Bioinformatics (Oxford, England)* 24(6), 880–1 (2008)
5. Cho, K.-H. et al.: Experimental design in systems biology, based on parameter sensitivity analysis using a monte carlo method: A case study for the TNF $\alpha$ -mediated NF- $\kappa$ B signal transduction pathway. *SIMULATION* 79, 726–739 (2003)
6. Goble, C. A. et al.: myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research* 38(Web Server issue), W677–682 (2010)
7. Heath, J. et al.: Probabilistic model checking of complex biological pathways. *Theor Comput Sci* 391(3), 239–257 (2008)
8. Hindmarsh, A. C. et al.: SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software* 31(3), 363–396 (2005)
9. Hinton, A. et al.: PRISM: A tool for automatic verification of probabilistic systems. *Lect Notes Comput Sc* 3920, 441–444 (2006)
10. Hucka, M. et al.: The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531 (2003)
11. Hull, D. et al.: Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 34(Web Server issue), W729–732 (2006)
12. Kwiatkowska, M. et al.: Using probabilistic model checking in systems biology. *ACM SIGMETRICS Performance Evaluation Review* 35(4), 14–21 (2008)
13. Machné, R. et al.: The SBML ODE solver library: a native API. for symbolic and fast numerical analysis of reaction networks. *Bioinformatics* 22, 1406–1407 (2006)
14. Wolkenhauer, O. et al.: Modeling and simulation of intracellular dynamics: choosing an appropriate framework. *IEEE Transactions on Nanobioscience* 3(3), 200–207 (2004)

## A Normalized Weighted RMSD Measure for Protein Structure Superposition

Xueyi Wang

Department of Mathematics and Computer Science  
Northwest Nazarene University, Nampa, ID 83686 USA  
xwang@nnu.edu

### 1 Introduction

Protein structure superposition is useful for evaluating the quality of predicted protein models, assessing the precision of NMR ensembles, and identifying structurally conserved or flexible regions. Root-mean-squared deviation (RMSD) is the most widely used measure for comparing protein structures. For a pair of superimposed structures, we measure the average distance of all point pairs and for multiple superimposed structures, we measure the average distance of point pairs in all structure pairs. One deficiency with the RMSD is its sensitivity to outliers, in which case a single outlier may cause a significant increase of the RMSD.

We introduce a new measure called normalized weighted RMSD (nwRMSD), which is extended from weighted RMSD with position weights [6], to minimize the structure superposition. We show that with normalization, the nwRMSD becomes a natural extension of RMSD and we can compare nwRMSD values of different superimposed structures. We also show that nwRMSD can be regarded as a function space, where many existing measures are special cases. Furthermore, we present an efficient iterative algorithm to minimize the nwRMSD given any convergent weight function and propose a new weight function for measuring superimposed structures. For pairwise structure superposition, we test on the structure superposition of predicted protein structure models and experimentally determined targets in free modeling category of CASP7 [4] and new folds category of CASP8 [1] and compare nwRMSD to measures used by CASP (GDT\_TS [7], AL0\_P [7], and MAMMOTH Z-score [5]) and the Gaussian-weighted RMSD [3]. For multiple structure superposition, we test on NMR ensembles in CASP8 and compare to the ensembles optimized by the standard RMSD and those used by the Protein Data Bank (PDB). The results show in general the nwRMSD performs better than other standard CASP scores in measuring the similarity between predicted protein structure models and experimentally determined targets and performs better than the standard RMSD and the original PDB ensembles in displaying structurally conserved or flexible regions of NMR ensembles.

### 2 Normalized Weighted RMSD

We assume there are  $n$  structures each having  $m$  points (atoms) and each structure  $S_i$  for  $(1 \leq i \leq n)$  has points  $p_{i1}, p_{i2}, \dots, p_{im}$ . We assign a position weight  $w_k \geq 0$  to each superimposed position  $k$  that  $\sum_{k=1}^m w_k > 0$  and define a normalized position weight  $\hat{w}_k = m w_k / \sum_{k=1}^m w_k$  (note that  $\sum_{k=1}^m \hat{w}_k = m$ ). We define a weighted average structure  $\bar{S}$  to have points  $\bar{p}_k = \sum_{i=1}^n \hat{w}_k p_{ik} / \sum_{i=1}^n \hat{w}_k$  for  $(1 \leq k \leq m)$ . We define a normalized

weighted root mean squared deviation (nwRMSD)

$$\sqrt{\frac{\sum_{i=2}^n \sum_{j=1}^{i-1} \sum_{k=1}^m \hat{w}_k \|p_{ik} - p_{jk}\|^2}{\left(\frac{n(n-1)}{2} \sum_{k=1}^m \hat{w}_k\right)}} = \sqrt{\frac{2}{mm(n-1)} \sum_{i=2}^n \sum_{j=1}^{i-1} \sum_{k=1}^m \hat{w}_k \|p_{ik} - p_{jk}\|^2}.$$

The normalization of position weights makes it possible to directly compare nwRMSD values when using different weight assignments. We have the following two theorems establish how the nwRMSD changes when weights change.

**Theorem 1.** Given two weight functions  $w_k$  and  $w_k'$  ( $1 \leq k \leq m$ ) for  $n$  structures and two nwRMSD measures  $nwRMSD$  and  $nwRMSD'$ , if  $w_k' = sw_k$ , where  $s > 0$  is a scalar, then  $nwRMSD = nwRMSD'$ .

**Theorem 2.** Given two weight functions  $w_k$  and  $w_k'$  ( $1 \leq k \leq m$ ) for  $n$  structures, RMSD for each position  $k$ :  $RMSD_k = \sqrt{2 \sum_{i=2}^n \sum_{j=1}^{i-1} \|p_{ik} - p_{jk}\|^2 / n(n-1)}$ , and two nwRMSD measures  $nwRMSD$  and  $nwRMSD'$ , if  $w_k' = w_k$  when  $k \neq c$  and  $w_k' = w_k + \Delta w_k$  when  $k = c$  ( $1 \leq c \leq m$ ), then we have:

(a) if  $\Delta w_k > 0$ ,  $nwRMSD \geq nwRMSD'$  iff  $RMSD_c \leq nwRMSD$  and  $nwRMSD \leq nwRMSD'$  iff  $RMSD_c \geq nwRMSD$

(b) if  $\Delta w_k < 0$ ,  $nwRMSD \geq nwRMSD'$  iff  $RMSD_c \geq nwRMSD$  and  $nwRMSD \leq nwRMSD'$  iff  $RMSD_c \leq nwRMSD$

Next we list three theorems that establish the relations of the nwRMSD and the average structure.

**Theorem 3.** The normalized weighted sum of squared distances for all pairs equals the normalized weighted sum of squared distances to the average structure:  $\sum_{i=2}^n \sum_{j=1}^{i-1} \sum_{k=1}^m \hat{w}_k \|p_{ik} - p_{jk}\|^2 = n \sum_{i=1}^n \sum_{k=1}^m \hat{w}_k \|p_{ik} - \bar{p}_k\|^2$ .

**Theorem 4.** The average structure  $\bar{S}$  minimizes the weighted sum of squared distances from all the structures, i.e. for any structure  $Q$  with points  $q_1, q_2, \dots, q_m$ ,  $\sum_{i=1}^n \sum_{k=1}^m \hat{w}_k \|p_{ik} - q_k\|^2 \geq \sum_{i=1}^n \sum_{k=1}^m \hat{w}_k \|p_{ik} - \bar{p}_k\|^2$  and equality holds if and only if  $q_k = \bar{p}_k$  for all positions with  $w_k > 0$ .

**Theorem 5.** The structure in a set  $Q_1, \dots, Q_t$  that minimizes the weighted sum of squared distances to all structures  $S_i$  for ( $1 \leq i \leq n$ ) is the one whose nwRMSD is closest to  $\bar{S}$ .

Next we present an iterative algorithm that takes  $O(nm)$  operations in each iteration to minimize the nwRMSD for multiple structures.

**Algorithm 1.** Given  $n$  structures with  $m$  points (atoms) each and weights  $w_k$  at each position, minimize nwRMSD to within a threshold value  $\varepsilon$  (e.g.  $\varepsilon = 1.0 \times 10^{-5}$ ).

1. Translate a weighted centroid of each structure  $S_i$  for ( $1 \leq i \leq n$ ) to the origin.
2. Calculate the average structure  $\bar{S}$  and deviation  $SD = \sum_{i=1}^n \sum_{k=1}^m \hat{w}_k \|p_{ik} - \bar{p}_k\|^2$ .
3. For each  $S_i$  ( $1 \leq i \leq n$ ), superimpose it to  $\bar{S}$  using Horn's method that minimizes  $\sum_{k=1}^m \hat{w}_k \|R_i p_{ik} - \bar{p}_k\|^2$  with an optimum rotation matrix  $R_i$ . Replace  $S_i^{\text{new}} = R_i \times S_i$ .
4. Calculate a new average  $\bar{S}^{\text{new}}$  and deviation  $SD = \sum_{i=1}^n \sum_{k=1}^m \hat{w}_k \|p_{ik}^{\text{new}} - \bar{p}_k^{\text{new}}\|^2$ .
5. If  $SD - SD^{\text{new}} < \varepsilon$ , exits; otherwise, set  $SD = SD^{\text{new}}$  and  $\bar{S} = \bar{S}^{\text{new}}$  and go to step 3.

Algorithm 1 minimizes nwRMSD if all position weights are fixed. If we already know a weight function  $f(k)$  for  $(1 \leq k \leq m)$  that assigns higher weights to better superimposed positions and lower weights to outliers, then we could use the following heuristic algorithm to optimize structure superposition.

**Algorithm 2.** Given  $n$  structures with  $m$  points (atoms) each, optimize structure superposition based on weight function  $f(k)$  for  $(1 \leq k \leq m)$ .

1. Set all  $w_k = 1$  for  $(1 \leq k \leq m)$  and minimize  $SD$  of the protein structures using the Algorithm 1.

2. For each aligned position  $k$ , calculate and set  $w_k^{\text{new}} = f(k)$  and minimize  $SD^{\text{new}}$  using Algorithm 1.

3. If  $SD - SD^{\text{new}} < \varepsilon$  (e.g.  $\varepsilon = 1.0 \times 10^{-5}$ ), then the algorithm terminates; otherwise set  $SD = SD^{\text{new}}$  and go to step 2.

Since  $SD \geq 0$  and  $SD$  decreases in steps 2 and 3, Algorithm 2 will eventually stop.

### 3 Results and Discussion

We can regard the nwRMSD as a function space, where many existing distance-cutoff, number-cutoff, or position weighted RMSD based measures can be regarded as special cases of the nwRMSD measure. For example, we can map a distance-cutoff method to

the nwRMSD by using a weight function  $w_k = \begin{cases} 1 & d_k \leq t \\ 0 & d_k > t \end{cases}$ , so  $\hat{w}_k = \begin{cases} m/m_1 & d_k \leq t \\ 0 & d_k > t \end{cases}$  and

$$\text{nwRMSD} = \sqrt{\sum_{k=1}^m \hat{w}_k d_k^2 / m} = \sqrt{\sum_{k=1}^{m_1} m d_k^2 / m_1 / m} = \sqrt{\sum_{k=1}^{m_1} d_k^2 / m_1} = \text{RMSD}.$$

Then we can use the same method to minimize nwRMSD, use the weight function to choose a new set of  $m_2$  point pairs, and repeat these steps until the RMSD converges or certain criterion satisfies.

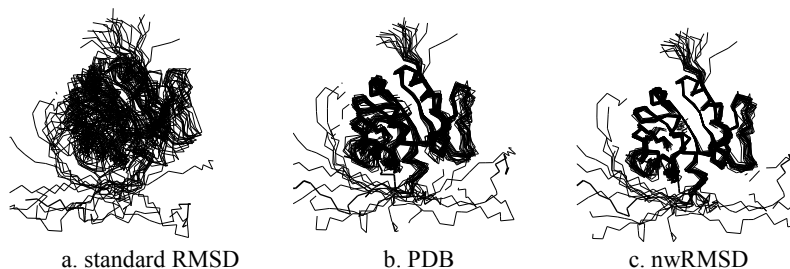
We choose a weight function  $w_k = 1/(\log(d_k^3 + 1) + c)$  and Algorithm 2, where  $c$  is a non-negative constant, to optimize both pairwise and multiple structure superpositions.

For pairwise structure superposition, we test the weight function on 17 protein targets in the free modeling category in CASP7 experiment [4] and 8 protein targets in the new fold category in CASP8 experiment [1]. In both experiments, evaluators (experts) vote on a few best models and second best models among all models using the scores from existing measures as references. Table 1 shows the top models chosen by experts and show the rankings by nwRMSD, GDT\_TS [7], AL0\_P [7], MAMMOTH Z-score [5], and the Gaussian-weighted RMSD [3] for CASP8 targets. The performance of the nwRMSD is comparable to GDT\_TS and AL0\_P in general. Considering the fact that the expert rankings are highly influenced by the GDT\_TS and AL0\_P scores [2], the comparable performance shows that the nwRMSD measure can be considered as an alternative measure for the CASP model evaluation.

For multiple structure superposition, we test the weight function on 14 NMR structure targets in CASP8 and compare the results to both the superposition by standard RMSD and the one used by PDB. Figure 1 shows the NMR ensemble of T0472 (2K4M). We can see that the ensembles optimized by PDB and nwRMSD are significantly better than optimized by RMSD and the ensemble by the nwRMSD is slightly better than the one by PDB.

**Table 1.** Comparison the rankings of expert, nwRMSD, and other measures for the predicted structure models for 8 targets in the new fold category in CASP8.

Target	Predicted structure	Rankings					
		Expert	nwRMSD	GDT_TS	AL0_P	MAMMOTH	Gaussian
T0397-D1	TS093_2	1	31	38	56	54	38
	TS020_5	2	66	16	1	56	33
T0405-D1	TS489_1	1	1	1	1	1	51
	TS371_5	2	2	3	2	6	24
T0443-D1	TS149_3	1	2	1	7	11	70
	TS149_5	2	1	3	19	12	60
T0460-D1	TS489_3	1	1	1	1	1	73
	TS387_1	2	2	2	4	2	90
T0476-D1	TS489_1, TS404_2	1, 2	1, 11	1, 2	1, 2	8, 12	47, 100
T0482-D1	TS489_3	1	1	1	1	1	100
	TS081_3	2	6	2	4	10	65
T0510-D3	TS404_4_2	1	1	1	1	2	35
	TS340_3, TS385_4	2, 3	5, 6	2, 3	2, 3	3, 4	81, 82
T0513-D2	top 28 of GDT_TS	1-28	1-27, 29	1-28	1-28	1, 4-22, 24-30	24 ~ 78



**Fig. 1.** The superposition of NMR structure target T0472 (2K4M).

**Acknowledgments.** We thank Prof. Jane S. Richardson at Duke University for helpful discussion. This work is supported by NIH grant #P20 RR0116454.

## References

1. Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J.L., Levy, Y.: Assessment of CASP8 structure predictions for template free targets. *Proteins*, S9, pp. 50--65 (2009)
2. Cozzetto, D., Kryshchovych, A., Fidelis, K., Moulton, J., Rost, B., Tramontano, A.: Evaluation of template-based models in CASP8 with standard measures. *Proteins*, S9, pp. 18--28 (2009)
3. Damm, K.L., Carlson, H.A.: Gaussian-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structure. *Biophysical Journal*, 90, pp. 4558--4573 (2006)
4. Jauch, R., Yeo, H.C., Kolatkar, P.R., Clarke, N.D.: Assessment of CASP7 structure predictions for template free targets. *Proteins*, 69(S8), pp. 57--67 (2007)
5. Ortiz, A.R., Strauss, C.E., Olmea, O.: MAMMOTH: An automated method for model comparison. *Protein Sci.* 11(11), 2606--2021 (2002)
6. Wang X., Snoeyink, J.: Multiple Structure Alignment by Optimal RMSD Implies that the Average Structure is a Consensus. In *Proceedings on 2006 LSS Computational Systems Bioinformatics Conference (CSB)*, pp. 79--87 (2006)
7. Zemla, A.: LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13), pp. 3370--3374 (2003)

## Accelerated Microbial Evolution in Complex Dynamic Environments through Step-wise Adaptation

Vadim Mozhayskiy<sup>1</sup> and Ilias Tagkopoulos<sup>1</sup>

<sup>1</sup> Department of Computer Science and Genome Center  
University of California Davis  
One Shields Avenue, Davis, CA 95616, USA  
itagkopoulos@ucdavis.edu

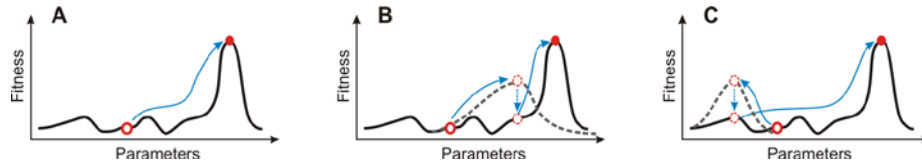
**Abstract.** During their lifetime, organisms as simple as bacteria are exposed to a variety of environments, each with its distinct spatio-temporal dynamics. Microbial communities display a remarkable degree of phenotypic plasticity, and organisms with high fitness emerge quite rapidly during evolution in novel environments. However, while adaptation occurs rapidly in certain environmental transitions, in others organisms struggle to adapt. Here, we investigate the hypothesis that the rate of evolution can both increase or decrease, depending on the similarity and complexity of the intermediate and final environments. Our results show that the rate of evolution can be accelerated by evolving cell populations in sequential combinations of environments that are increasingly more complex. To quantify environmental complexity, we evaluate various information-theoretic metrics, and we show that multivariate mutual information of environmental signals correlates well with the rate of evolution measured in our simulations. We find that strong positive and negative correlations between the intermediate and final environments lead to the increase of evolutionary rates, when the environmental complexity increases. . Elucidating such dependencies is the first step towards controlling the rate and direction of evolution, which is of interest to bioengineering and biotechnological applications.

**Keywords:** Microbial Evolution, Biological Networks, Simulation, Multi-scale Modeling.

### 1 Introduction and Methods

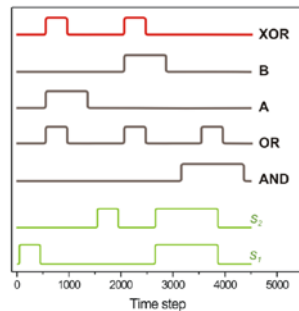
All life forms, from microbes to higher vertebrates, are constantly subjected to evolutionary processes that lead to adaptation and phenotypic variation. Whether evolutionary forces lead to new and rapidly evolving species, as it is in the case of adaptive radiation, or are responsible for phenotypic divergence within a species, the underlying mechanism by which complex behavior arises remains the same: gradual accumulation of selected genetic mutations and epigenetic changes gives rise to a myriad of anatomical, physiological and behavioral expressions. A challenging task is to identify the environmental and organism-specific characteristics that allow the rapid adaptation from past to new environments. Computer simulations with logic

gates and RNA secondary structures have demonstrated that facilitated variation spontaneously emerges during evolution [1], while theoretical models provide support that varying environments alter evolution [2] and gives rise to modular structures [3].



**Fig. 1.** Step-wise evolution: Adaptation to a complex environment (A) can be accelerated (B) or decelerated (C) if guided through intermediate steps of a lesser complexity. Fitness profile for a population evolving in a target complex environment (solid black curve) is a multidimensional surface with multiple local maxima. Adaptation to intermediate environments (dashed grey fitness profiles) can direct evolution towards to (or away from) the global fitness maximum of the target environment.

In complex environments, organisms have to explore a large parameter space before settling in stable fitness points (Fig. 1). Local minima and discontinuities may lead to sub-optimal fitness peaks, from where it may be difficult, or even infeasible, to escape. In addition, it has been shown that phenotypes that occupy flatter regions of the fitness surface are more robust to mutations, a phenomenon that was coined as “survival of the flattest”[4]. To further investigate the hypothesis that intermediate environments can accelerate or decelerate the evolution, we first define metrics to quantify environmental complexity, and then proceed to measure the rate of evolution in five environments by performing multi-scale simulations of evolving microbial populations in fluctuating environments.



**Fig. 2.** Environments: Environmental signals (green) and nutrient abundance for five environments (bottom to top: AND, OR, A, B, XOR) shown as a function of time steps within one epoch. Nutrient presence is a delayed function of the two signals.

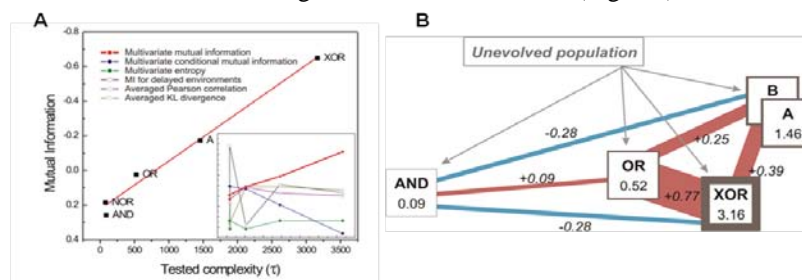
We used the EVE simulator (Evolution in Variable Environments) to perform simulations of microbial populations in six environments with distinct temporal dynamics. The EVE simulator employs abstract, multi-scale models of basic sub-cellular phenomena related to expression (transcription, translation, protein modification, degradation, basal expression), evolution (mutation, gene duplication, gene deletion), network regulation and other evolutionary processes (natural selection). It has been used successfully to generate hypotheses in nutrient-limited microbial communities [5], and it has been documented elsewhere [6,7]. In our simulations, organisms evolving in a fixed sized population cannot directly sense the



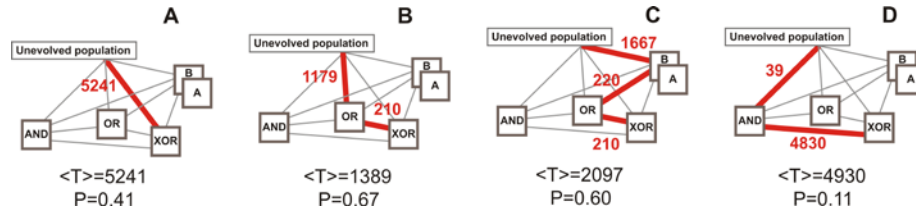
presence of resources, but can infer their future presence, if they couple to various environmental signals through biochemical and regulatory interactions. For our simulations here, we used environments where two signals,  $s_1$  and  $s_2$ , correlate to the presence of nutrients in the environment (Fig. 2). For example the I/O characteristic of environment A is given by the logic *Nutrients Presence [A] = Delayed ( $s_1$  AND NOT( $s_2$ ))*. XOR environment with the most complex correlation structure, due to the fact that the XOR gate is not linearly separable.

## 2 Results and Discussion

It is expected that during evolution, the time it takes until a fit phenotype emerges is inversely proportional to the complexity of the environment and information processing that the fit network model should be capable of. It is an open question, however, how we can capture the environmental complexity and link it to the time it will take to have a partial or full solution – in other words, to calculate the “rate of evolution”. To address this question, we evolved initially random populations in five environments (Fig. 2), and then measured the rate of evolution over 32 simulation runs. As expected, the rate of evolution (measured as the time constant in an exponential fit) was slower for XOR than any other environment tested. We found that multivariate mutual information,  $I(s_1; s_2; \text{nutrients})$ , correlates well with the rate of evolution for the environments that we studied (Fig. 3A). Next, we calculated the environmental similarity between any two environments by measuring the pair-wise Pearson correlation of nutrient presence. Fig. 3B depicts the final “environmental network” that reflects the two measured quantities, the environmental quantity and similarity. We continued by testing whether evolution can be accelerated or decelerated by using different paths within the environmental network (Fig. 4). We found that exposing organisms to intermediate correlated environments of increasing complexity leads to a higher rate of evolution. As shown in Fig. 4, adaptation to a XOR environment becomes 74% faster, if cells first evolve in an intermediate OR environment. This phenomenon holds even for more than one intermediate environment (Fig. 4C). Interestingly, this acceleration is also observed in the case where the intermediate environment is strongly anti-correlated to the final one, as in the case of XOR evolution through the AND environment (Fig. 4D).



**Fig. 3.** (A) Multivariate mutual information was found to correlate with the evolutionary rate measured in simulations. Insert depicts this relationship with other informational measures used. (B) Relative complexity (node values) and similarity (link values) of AND, OR, A, B, and XOR environments.



**Fig. 4.** Single-step and multi-step evolution towards a XOR phenotype: The highlighted edges of each network correspond to the environmental transitions made during evolution (evolutionary path). The value of each edge corresponds to the average time to evolve a fit phenotype, in number of epochs.  $\langle T \rangle$  and  $P$  are the average total time it takes to evolve the XOR phenotype and the success probability (ratio of successful over total experiments). (A) Direct evolution to an XOR environment is slow and has a low probability of success; (B,C) Initial evolution to one or more intermediate environments of lower complexity and subsequent evolution to the final environment accelerates adaptation, and boosts the fraction of successful simulations; (D) Intermediate environment, which is strongly anti-correlated to the final one: evolution is accelerated as well, but the fraction of successful simulations is significantly lower.

In this paper, we show that the rate of evolution can be both accelerated and decelerated by exposing a cell population to a series of environments. The implications of this work span many areas of biological research. First, we extend our current understanding of evolution by deriving a set of rules that explain the directional change of evolutionary rates when populations are exposed to a series of environments. By quantifying the environmental complexity, we made possible a more rigorous assessment of the evolutionary potential of microbial communities in complex environments. In the context of synthetic biology and bioengineering, our work can provide a first glimpse on how to control evolutionary rates by systematically exposing microbial cultures to correlated environments, a method that may prove a powerful tool for the fine-tuning of genetic constructs and for the engineering of desired phenotypes.

## References

1. Gerhart J, Kirschner M (2007) The theory of facilitated variation. *Proceedings of the National Academy of Sciences of the United States of America* 104: 8582-8589.
2. Kashtan N, Noor E, Alon U (2007) Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences of the United States of America* 104: 13711-13716.
3. Kashtan N, Mayo AE, Kalisky T, Alon U (2009) An Analytically Solvable Model for Rapid Evolution of Modular Structure. *Plos Computational Biology* 5: 14.
4. Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412: 331-333.
5. Tagkopoulos I, Liu YC, Tavazoie S (2008) Predictive behavior within microbial genetic networks. *Science* 320: 1313-1317.
6. Tagkopoulos I (2008) Emergence of Predictive Capacity within Microbial Genetic Networks, PhD Thesis: Princeton University.
7. Mozhayskiy V, Tagkopoulos I (2011) In Silico Evolution of Multi-Scale Microbial Systems in the Presence of Mobile Genetic Elements and Horizontal Gene Transfer. ISBRA. CSU, China.

## Computational Framework to Support Data Storage and Analysis in Translational Medicine

Newton Shydeo Brandão Miyoshi<sup>1</sup>, Daniel Guariz Pinheiro<sup>2</sup>, Wilson Araújo da Silva Junior<sup>2</sup>, Joaquim Cezar Felipe<sup>1</sup>,

<sup>1</sup>Department of Computing and Mathematics, FFCLRP

<sup>2</sup>Department of Genetics, FMRP

<sup>1,2</sup>University of São Paulo, Ribeirão Preto, Brazil

newton.sbm@usp.br, dgpinheiro@gmail.com, wilsonjr@usp.br, jfelipe@ffclrp.usp.br

**Abstract.** Using knowledge generated in science to promote human health is the main goal of translational medicine. To make this possible we need computational methods to handle the large amount and heterogeneity of information that arise from bench to bedside. A computational barrier to be overcome is the integrations of clinical, socio-demographic and biological data. In these effort ontologies plays an essential role by being a powerful artifact for knowledge representation. Chado is a modular database model ontology-oriented that gained popularity for being a robust, flexible and generic platform to store biological data but it lacks supporting representation of clinical and social demographic information. This paper presents a framework to support translational research by integrating clinical and socio-demographic data with biomolecular information coming from different “omics” technologies. For this we extended Chado to allow the representation of clinical and socio-demographic information.

**Keywords:** Translational Medicine; Biological Database; Data Integration; Chado

### 1 Introduction

Translational research seeks to reduce the gap that exists between the bench and bedside. This is a great challenge that has many barriers to overcome. One of the most important is related to the nature of the data. The nature of clinical data is very different from molecular data although they are often closely related. To conduct a deep investigation regarding complex mechanisms responsible for the onset of pathological processes, a global analysis concerning different levels of information is most necessary. To make this possible, two aspects of data handling must be well defined: storage and analysis. It is necessary to provide a computational platform and a data model able to store, represent and integrate clinical and biomolecular information consistently. From a well formalized and structured model it is possible to design novel methods of computational analysis.

In the area of genomics there are several models of biological databases such as AceDB and Ensembl. These models serve as the basis for construction of computational tools for genomic analysis in an organism-independent way. A model

of biological databases which has gained popularity is Chado. It is a robust, flexible and generic platform that can be adapted to support research in different organisms.

In this context the present work aims at the definition of an integrative computational platform for translational science. We will use and extend Chado as the underline database model to be able to aggregate, in a consistent way, clinical and molecular data, enabling the development of computational analysis to be applied in the field of translational medicine. To guarantee the standardization and enable further development of generic analysis tools we propose the design and use of a common reference ontology. Through this framework it will be possible to integrate sequence data, gene expression data from microarray, microRNA and disease association data with the clinical and socio-demographic features.

## 2 Proposed Framework

The proposed framework is compound by the proposed Chado clinical module, a migration methodology to be applied in legacy clinical research databases and an ontological mapping that allows data standardization, integration and development of generic analysis tools.

### 2.1 Chado

Mungal, Emmert and the Flybase group proposed a modular design based on ontology to represent biological information called Chado. Chado is a relational database schema that can be used as a basis for any group of genomic research. Chado is part of GMOD project[1] (Generic Model Organism Database) and is currently used by several research groups such as Xendb, ParameciumDB, AphidBase among others.

One hallmark of Chado in relation to other generic databases models is that it makes intensive use of ontologies. Ontology plays a central role in Chado, because all stored information must be related to some ontology or controlled vocabulary. Some ontologies are already incorporated into but it is possible to incorporate new ontologies.

There are computational tools that are compatible with Chado database. These tools are mostly provided by the GMOD group. We can cite the genome browser Gbrowse[2] and the annotation tool Apollo[3]. Chado also allows incorporation of other tools through the creation of Bridge Layers which consist of built views to make Chado similar to other databases and act as layers of compatibility with other tools.

### 2.2 Proposed Model

The proposed Clinical Module is compound mainly by five tables: patient, patient\_project, clinicosocialdata, clinicosocialdata\_relationship and extrainfo. The patient table is self-described, this is where are stored the patients. Each patient can participate in various projects, this information is represented in table patient\_project. The clinicosocialdata table is where the most of the information are stored. This table was designed to represent, in a flexible way, any kind of clinical or socio-demographic information. The table clinicosocialdata\_relationship is used when it is necessary to

represent complex relationships between clinical or socio-demographic data. In extrainfo table are stored data that are patient-independent such as cities names and codes. The semantics of the clinical data stored in this module is typed by an ontology stored in the Controlled Vocabulary module of Chado.

## 2.3 Proposed Migration Methodology

We developed a methodology to migrate data from a legacy databases to the Chado Clinical Module. The methodology consists in four steps:

1. Create an ontology to represent the clinical database;
2. Store the clinical database ontology in the Controlled Vocabulary Module of Chado;
3. Store the data in Clinical Module according to the clinical database ontology;
4. Create a set of views in Chado to act as a bridge layer of the clinical database;

## 2.4 Ontological Mapping

The proposed structure does not define the meaning of information stored. These information could be represented using specific ontologies that capture the meaning of that data in the particular database. But to get the most out of this generic model, enabling the development of analysis tools that could be applied in different instances of Chado with data descending from different clinical databases it is necessary to define a common semantic. This can be done by adopting a reference ontology, so the analysis tools could be designed to obtain semantic information from the reference ontology. The work then consists of ontological mapping between the ontology that describes the clinical database and the reference ontology.

In this work we proposed the SNOMED CT[4] as the common reference ontology. The advantage of using SNOMED as reference ontology is because it covers many independent domains. It is composed by more than 300.000 terms and the domains coverage vary from body structures, diseases, pharmaceutical products to geographic locations, social contexts and physical objects.

## 3 Results

To test the proposed framework we have implemented an instance of Chado using the Database Management System PostgreSQL 8.4[5]. We also implemented the proposed Clinical Module.

We tested the approach with success in data from the project "Oncogenomics Applied to Therapy of Head and Neck Carcinoma" from GENOPROT Network (CNPq) where the information was stored in the database of Clinical Genomics Project which is part of the Ludwig/FAPESP Human Cancer Genome Project.

Clinical and demographic data were obtained from patients with tumors of head and neck through the Service of Head and Neck Surgery in School Hospital of Faculty of Medicine (SH-FM) of University of São Paulo, at Ribeirao Preto, Brazil.

A Chado instance was installed on the relational DBMS PostgreSQL. The clinical database has about 20 tables with some table containing up to 120 columns. The main table stores information about the patient like age, sex, weight and height. The clinical information was stored in a MySQL[6] relational DBMS.

## 4 Conclusion

Turning knowledge generated by science in a real benefit to enhance human health is a difficult task. This is one of main goals of translational research, more specifically research in translational medicine. To make this real, a computing infrastructure is required to support storage, management, integration and analysis of both biological and clinical information.

This work aims to take a step toward this infrastructure proposing a data model that enables the representation of clinical, socio-demographic and biological information in a single database. The biodatabase model Chado was extended to include a module for representing clinical information

Through the proposed clinical information module several clinical databases can be adapted. The real benefit of adopting a generic model for information representation becomes concrete with the emergence of various applications and analysis tools that are constructed and maintained by the community that adopts this model. It also facilitates the integration of applications and the exchange of data between research groups and also for research groups that do not adopt Chado and may be wearing it after the proposed extension.

The adoption of Chado as the basic model of biological database allows the reuse of the existing tools built from Chado or adapted to it through bridge layers for analysis and visualization of molecular data. With the proposal of the clinical module this solution becomes a robust computational framework to support research in translational medicine.

**Acknowledges:** The authors thank CAPES for the financial support.

## References

1. GMOD. *GMOD*. 2009. Available at: [http://gmod.org/wiki/Main\\_Page](http://gmod.org/wiki/Main_Page).
2. L.D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva, and S. Lewis, "The generic genome browser: a building block for a model organism system database.," *Genome research*, vol. 12, Oct. 2002, pp. 1599-610.
3. S. Lewis, S. Searle, N. Harris, M. Gibson, V. Iyer, J. Richter, C. Wiel, L. Bayraktaroglu, E. Birney, M. Crosby, J. Kaminker, B. Matthews, S. Prochnik, C. Smith, J. Tupy, G. Rubin, S. Misra, C. Mungall, and M. Clamp, "Apollo: a sequence annotation editor," *Genome Biology*, vol. 3, 2002, pp. research0082.1-0082.14.
4. IHTSDO. IHTSDO: SNOMED CT. 2010. Available at: <http://www.ihtsdo.org/snomed-ct/>
5. PostgreSQL Global Development Group. PostgreSQL. 2010. Available at: <http://www.postgresql.org/>.
6. Oracle. MySQL :: The world's most popular open source database. 2010. Available at: <http://www.mysql.com/>

## Bayesian Elastic Net for Multi-Class Classification and Survival Analysis

Lingling Zheng<sup>1,2</sup>, Minhua Chen<sup>3</sup>, Joseph Lucas<sup>1</sup>, Lawrence Carin<sup>3</sup>,

<sup>1</sup> Program of Computational Biology & Bioinformatics, Duke University, Durham, NC 27710

<sup>2</sup> Institute for Genome Sciences & Policy, Duke University, Durham, NC 27710

<sup>3</sup> Department of Electrical & Computer Engineering, Duke University, Durham, NC 27710

High-throughput technologies have emerged over the last few years as important tools for genomic research. Variable selection, such as gene selection using microarray expression data, becomes fundamental to high-dimensional data analysis. It enables us to identify the most significant genes associated with a certain kind of disease. However, the large-scale data measured usually comes with a much higher dimension of genes than the number of observed samples. Thus, it poses unique challenges for data mining. Additionally, highly correlated features are also frequently encountered in variable selection. Therefore, standard statistical techniques are often inadequate. We address the problem by introducing Bayesian Elastic Net to achieve a grouped and sparse solution. Here, we mainly extend this method in two directions. First, since genomic or proteomic data usually comes with more than two types of samples, we develop a multinomial probit regression model of Bayesian Elastic Net for classification problem. It is capable of identifying classifiers to infer a set of important and highly correlated predictors, e.g. genes or peptides. Additionally, in order to identify genomic biomarkers for clinical application, we present an adaptation of Bayesian Elastic Net under censored exponential regression model. Although sampling from the full posterior distribution using Gibbs sampler can be straightforward and accurate, it is computationally expensive if given large sample size or predictor numbers. Therefore, we adopt a variational Bayesian (VB) inference method, which has advantage in fast convergence and computational efficiency. Those two models are validated by first performing simulation on toy datasets, and then on case studies. We apply the multinomial probit regression model to classify and predict sepsis patients' disease status. The dataset in this study, collected by National Center for Genome Resources (NCGR), is drawn from measurements of clinic indices of sepsis patients at different status, including NIS (non-infected SIR), UCS (uncomplicated sepsis), SS/Shock (severe sepsis or septic shock) and SD (sepsis death). A goal of this study is to help diagnose patients' prognosis based on the biomarker indices. Additionally, we assess Bayesian survival model on lung cancer patients' microarray data and failure time. The genetic signatures selected from the model separate the good and poor prognosis patient groups at a significant level, using the Kaplan-Meier curves. Results from Gibbs sampling are compared with the VB approximation. It shows that Gibbs sampler has better accuracy in classification, while VB tends to achieve better sparseness and efficiency. Thus, Bayesian Elastic Net strategy is a practical and informative method for variable selection. Additionally, the proposed multinomial probit regression model offers consistent and reliable classification across multi-categorized samples. Finally our survival regression model is a powerful technique for studying genomic effects on the duration of survival.

**Keywords:** Bayesian Elastic Net, multinomial probit regression, survival analysis, variational Bayesian, variable selection.

**GWAS-GMDR: a program package for genome-wide scan of gene-gene interactions with covariate adjustment based on multifactor dimensionality reduction**

Min-Seok Kwon<sup>1†</sup>, Kyunga Kim<sup>2†</sup>, Sungyoung Lee<sup>1</sup>, Wonil Chung<sup>2</sup>, Sung-Gon Yi<sup>2</sup>, Junghyun Namkung<sup>1</sup> and Taesung Park<sup>1,2\*</sup>

<sup>1</sup>Interdisciplinary Program in Bioinformatics, <sup>2</sup>Department of Statistics, Seoul National University, Seoul 151-742, Korea

<sup>†</sup>The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

**Abstract.** Multifactor dimensionality reduction (MDR) has been successfully applied to identification of gene-gene interactions that are well recognized as playing an important role in understanding complex traits. Generalized MDR (GMDR) was its extension that allows adjustment for covariates. The current GMDR software mainly focuses on candidate gene association studies with a relatively small number of genetic markers and has some limitations to be extended to genome-wide association studies (GWAS) with a large number of genetic markers. We developed GWAS-GMDR, an effective parallel computing program package with special features for GWAS using distributed job scheduling method and/or CUDA-enabled high-performance graphic processing units (GPU). First, it implemented an effective memory handling algorithm and efficient procedures for GMDR to make joint analysis of multiple genes feasible for GWAS. Second, a weighted version of cross-validation consistency based on 'top- $K$  selection' ( $WCVC^K$ ) was proposed to report multiple candidates for causal gene-gene interactions. Our simulations indicated that  $WCVC^K$  has better performance in its ability to identify epistatic loci, compared with selecting models that just minimized the prediction error. Third, various performance measures were implemented to evaluate MDR classifiers, including balanced accuracy, tau-b, likelihood ratio and normalized mutual information. Fourth, three popular methods for handling missing genotypes were implemented complete, available and missing category. Finally, our applications support both CPU-based and GPU-based parallel computing system. We applied our applications using a real genome wide data set from WTCCC hypertension dataset to identify two-way interaction models in genome-wide scale.

**Keywords:** Gene-gene interaction; Multifactor dimensionality reduction(MDR); Genome-wide association study(GWAS); General purpose graphic processing unit (GPGPU);



## 1 Introduction

Multifactor dimensionality reduction (MDR) is a popular method to identify gene-gene (G×G) interactions that affect disease susceptibility simultaneously (Ritchie et al., 2001). MDR classifies samples into high and low risk groups based on relative risks at their genotype combinations of genetic variables. An evaluation measure (e.g., balanced accuracy) is used to assess classification and prediction performances and select the best MDR classifier. Finally, the single best MDR classifier is suggested via a voting algorithm based on cross-validation (CV), such as cross-validation consistency (CVC; Ritchie et al., 2001).

In genetic association studies, there often exist covariates that would interfere with correct inference on genetic effects. Thus the adjustment for such covariates is necessary. Generalized MDR (GMDR) was extended from the MDR to permit adjustment for covariates and implemented in Java GMDR (Lou et al., 2007). GMDR computes score statistics using a generalized linear model that contains covariates as well as genetic interactions, and constructs GMDR classifiers based on score statistics.

Since the genotype data with up to ~1 million single nucleotide polymorphisms (SNPs) became common, there is a growing need for more efficient GMDR program that enables genome-wide analysis of G×G interactions. However, Java GMDR was developed mainly for candidate gene studies with a small number of genetic variables, and no GMDR software is available for large-scale genotype data. For the original MDR analysis, some high performance software have been proposed, such as parallel MDR (pMDR) (Bush et al., 2006) that employs parallel computing environment, and MDRgpu (Sinnott-Armstrong et al., 2009) that requires graphics processing units. Unfortunately, none of these MDR software have capability of adjusting covariates.

We developed GWAS-GMDR, the first parallel computing software for GMDR analysis of GWAS. In addition, GWAS-GMDR has implemented some special features for GWAS. For example, it can report multiple best SNP combinations with similar G×G effects, instead of reporting one single best SNP combination.

## 2 Methods & Materials

### 2.1 MDR implementation

The GWAS-GMDR was modularized with four main components: data-processing, score-calculation, analysis, and result-reporting modules. The data-processing includes data-loading, handling of missing data, and data-partitioning for CV. In the score-calculation module, score statistics are computed for covariate adjustment. The analysis module consists of five sub-modules: generating combinations of genetic variables to be analyzed; calculating multilocus-genotype counts or average scores; constructing MDR classifiers; evaluating the classifiers; and storing results. In the result-reporting module, all results are aggregated and summarized to report final results. When there is no covariate, GWAS-GMDR skips the score-calculation module and performs the same MDR analysis.

When high-order interactions are exhaustively searched in GWAS, it may be infeasible to populate all combinations of genetic variables due to physical limitation

of system (e.g., machine memory). To resolve this problem, we developed an efficient indexing algorithm that partitions the indices for all possible combinations into equal-size subsets. Corresponding to each index-subset, MDR classifiers are constructed and evaluated. Users can specify the size of index-subsets to accommodate a memory condition of their machines as well as to optimize the performance of the software.

Besides balanced accuracy, we implemented eight additional evaluation measures, including tau-b, likelihood ratio, and normalized mutual information, three of which are known to improve the MDR performance (Namkung et al, 2009a). Also, three popular approaches were implemented to handle missing genotypes, such as ‘complete’, ‘available’ and ‘missing category’ (Namkung et al, 2009b).

## 2.2 Weighted cross-validation contingency

In GWAS-GMDR, we proposed an extension of CVC, called as  $WCVC^K$ , which is a weighted voting algorithm based on top- $K$  selection. This procedure selects a user-specified number ( $K$ ) of the best MDR classifiers (i.e., top  $K$  classifiers), rather than the single best classifier, at each CV step. For each selected MDR classifier, its weighted vote  $WCVC^K$  is calculated as below to indicate how many of the training-test sets support the selected classifiers as the  $K$  best classifiers in  $m$ -fold CV (e.g.,  $m = 10$ ):

$$WCVC_j^K = \frac{1}{m} \sum_{l=1}^m w_{jl} I_{jl} \quad \text{where } I_{jl} = \begin{cases} 1 & \text{if } j^{\text{th}} \text{ MDR classifier is identified as one} \\ & \text{of top } K \text{ classifiers at } l^{\text{th}} \text{ CV dataset} \\ 0 & \text{otherwise} \end{cases}$$

and  $w_{jl}$  = weight for  $j^{\text{th}}$  MDR classifier at  $l^{\text{th}}$  CV dataset

Here, the weights are used to take their performances at each CV step into account when top  $K$  classifiers are selected. Examples of a weight system are the inverse rank of top  $K$  classifiers within each CV step. Based on  $WCVC^K$  (e.g., all combinations with  $WCVC^K > 0.8$ ), multiple candidates of causal G×G interactions can be reported along with their performance measures (e.g., predictability for training and test datasets). Note that  $WCVC^1$  is equal to CVC.

## 2.3 Simulation and genome-wide real data analysis

Simulation was conducted to compare the performance of  $WCVC^K$  with general selecting model that just minimized the prediction error. This simulation was assumed case-control study using similar two-locus epistasis models considered by Namkung *et al.* (2009a). Using WTCCC hypertension dataset, we demonstrated genome-wide two-way interaction analysis. After quality control process, WTCCC hypertension dataset has 327,632 SNPs and 6,417 samples.

## 3 Result & Discussion

We developed GWAS-GMDR for G×G interaction analysis with covariate adjustment in genome-wide scale. GWAS-GMDR is flexible for various kinds of

computing environments. Under a parallel computing environment, users can use GWAS-GMDR via a job scheduling system including portable batch system which most cluster systems employ. When parallel computing is unavailable, GWAS-GMDR is executable with a single processor. And, users can run GWAS-GMDR (GPU version) on CUDA-enabled computing system with GPU devices with high-performance.

In order to demonstrate the performance, we installed GWAS-GMDR and the current Java GMDR on 2-GHz Dual Core AMD Opteron<sup>(tm)</sup> Processor and a workstation with three NVIDIA GeForce GTX285 graphic cards. We tested them for scanning all pairwise interactions under various settings (i.e., no or three covariates, sample sizes of 500~10000, and 100~30000 SNPs). With a single processor, GWAS-GMDR performed much faster (2~7 times) than the Java GMDR across all the test settings. In testing with larger sample sizes and a large number of SNPs, GWAS-GMDR showed further reduction in processing time in a remarkable degree under parallel computing. When 1000 SNPs are pre-screened for interaction analysis, GWAS-GMDR can evaluate all pairwise and three-way interactions in less than 1 minute even with 20 processors. Additionally, we found that datasets with no covariates can be analyzed 2~3 times faster by GWAS-GMDR than other CPU-based MDR software, such as libMDR and pMDR. With a GPU system, GWAS-GMDR completed two-way interactions in ~4.1 days with 327,632 SNPs and 6,417 samples.

The modularization feature makes GWAS-GMDR pliable to future modifications on individual modules. The new indexing algorithm can optimize the performance according to machine memory condition; eliminate limitations on sample sizes, the number of genetic variables, and the order of gene-gene interactions; and hence make it practically feasible to investigate G×G interactions for genome-wide scale datasets.

Finally, the GWAS-GMDR provides users with various options for missing handling and for evaluation measures. The proposed weighted voting algorithm (WCVC<sup>k</sup>) enables users to produce a list of candidate causal G×G interactions in the order of importance.

## References

1. Bush, W.S. et al. (2006) Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics*, 22, 2173–2174.
2. Hahn, L.W. et al. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19, 376–382.
3. Lou, X.Y., et al. (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am J Hum Genet*, 80, 1125–1137.
3. Namkung J. et al. (2009a) New evaluation measures for multifactor dimensionality reduction classifiers in gene-gene interaction analysis. *Bioinformatics*, 25, 338–345.
4. Namkung J. et al. (2009b) Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. *Genet. Epidemiol.*, [Epub ahead of print].
5. Ritchie, M.D. et al. (2001) Multifactor dimensionality reduction reveals high-order interaction among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, 69, 138–147.
6. Sinnott-Armstrong, N.A. et al. (2009) Accelerating epistasis analysis in human genetics with consumer graphics hardware. *BMC Res. Notes*, 2, 149.

## SNP-PRAGE: SNP-based Parametric Robust Analysis of Gene set Enrichment

Jaehoon Lee<sup>1</sup>, Soyeon Ahn<sup>1</sup>, Sohee Oh<sup>1</sup>, Taesung Park<sup>1</sup>,

<sup>1</sup>Department of Statistics, Seoul National University, Seoul, San 56-1, Shilim-dong, Korea

**Abstract.** GWA (Genome-wide association) analysis has been successful to detect significant SNPs and genes affecting common diseases. However, the single SNP approach in GWA can miss the joint effect from multiple genetic risks. Considering the multiple SNPs jointly under the prior biological knowledge can increase power of GWA analysis. When multiple SNPs are considered, the corresponding SNP-level association measures can be correlated due to LD among SNPs. We proposed SNP-PRAGE, a SNP-based parametric robust analysis of gene-set enrichment, which handles correlation adequately among association measures of SNPs by the parametric assumption. We summarized SNP-level association measures based on the gene size and the LD structure of nearby genes. We conducted the simulation study and applied SNP-PRAGE to hypertension data of 7,551 samples from KARE (Korea Association Resource) cohorts recruited in Korea. Our results shows SNP-PRAGE can reduce many false positives and requires much less computational efforts than the previous permutation-based gene set approaches.

**Keywords:** Genome-wide association analysis, gene set analysis, linkage disequilibrium (LD), parametric

### 1 Introduction

The GWA (genome-wide association) analysis has been successful to investigate genetic variant of individuals associated with some targeted phenotypes. Most GWA tests only consider association of a single SNP and list the most significant SNPs or genes. However, complex diseases often result from joint action of multiple risk factors and therefore the single-SNP-based approach may miss the genetic variants that jointly have significant risk effect but individually make only a small contribution. To address these issues, we consider the effect of multiple SNPs jointly and use prior biological knowledge for the SNPs.

GSEA [1], the pioneering gene set analysis method, was extended to GWA data by Wang *et al.* [2]. They repeated the permutation of sample label and calculation of gene set statistics 1,000 times for the test of gene set statistic. This permutation-based testing can preserve a correlation among the SNP-level measures due to LD among SNPs, but this is very computationally expensive process in genome-wide scale.

In order to reduce computing time, some parametric methods based on a specific distribution have been used. For example, GLOSSI method developed by Chai *et al.* [3] used Fisher's combination test under the assumption of correlated p-values. Nam *et al.* [4] proposed the Z-statistic method that compares a specific gene set to other sets. The Z-statistic method is the extension of PAGE [5], which is the parametric and competitive GSA for microarray data. However, the Z-statistic method assumes no correlation among SNP-level p-values.

We propose SNP-PRAGE, a SNP-based parametric robust analysis of gene-set enrichment, which is based on the simple normality assumption. We consider the correlation among SNP-level p-values without taking permutation step. We compare our method to Wang's method, GLOSSI and Z-statistic method via the simulation study in terms of size, power and computing time and also show the result based on hypertension phenotype of 7,551 samples from KARE cohorts.

## 2 Methods

### 2.1 Z-statistic method

Nam *et al.* [4] implemented the Z statistic method in their software, GSA-SNP. They used a negative logarithm of  $m^{\text{th}}$  best p-value ( $p_{(m)}$ ) within each gene as the gene summary measure. Using this gene summary measure, they calculated a Z-score as gene-set level summary. This Z-score is expected to follow a normal distribution based on the central limit theorem. In order to meet a normal distribution assumption, they assumed the gene level order statistic summary has identical and independent distribution (i.i.d.). However, the  $m^{\text{th}}$  best p-value is not identical over the gene size because a gene with many SNPs will have a lower  $m^{\text{th}}$  best p-value than genes with a few SNPs. They also assumed the gene-level summary measure has a homogeneous variance over the gene sets. However, the variance of their summary measure depends on the gene size. When the gene size is large, the variance of the summary measure of the gene will be small.

### 2.2 SNP-PRAGE

To address the issues of the Z-statistic method we mentioned above, we can multiply the gene size to the  $m^{\text{th}}$  best p-value so that it has approximately identical distribution over the gene size. The moment generating function does not depend on the gene size when p-values are independent of each other and the gene size is large enough. Gene-level measure may have an independent distribution but SNP-level p-values are not independent of each other because of the LD structure. We propose using the effective gene size instead of gene size so that gene-level summary measure has approximately identical distribution over the gene size irrespectively of the LD structure between p-values.

$$t_{ij} = (n_{ij}^* + 1)p_{(m)},$$

where  $t_{ij}$ ,  $n_{ij}^*$  are gene level summary measure and effective gene size (1), respectively, for the  $j$ th gene in the  $i$ th set.

Our empirical research shows that gene level measure does not have common variance over gene set especially with the small gene set size. So we assumed that the gene level measure has a heterogeneous variance over the gene sets. We compute the sample variance distinctly over gene set and apply Welch's t-statistic for the test.

### 3 Real Data analysis

We used canonical pathways from MsigDB database (<http://www.broadinstitute.org/gsea/msigdb/index.jsp>). We applied SNP-PRAGE to KARE (Korea Association REsource) GWA data. Participants of population-based cohorts recruited in Korea are genotyped with the Affymetrix Genome-Wide Human SNP array 5.0. We analyzed the hypertension data from 7,551 unrelated individuals. The logistic regression analysis with an additive model is conducted after adjustment for age, sex, and recruitment center. We obtained 5 significant gene sets from the result of SNP-PRAGE based on q-value 0.1 as cut off. These gene sets are known to be related to hypertension, directly or indirectly. [6-9]

### 4 Simulation study

We generated simulation data based on KARE data for the various gene size and SNP effect size. We compared the performance of SNP-PRAGE, modified GSEA method [1], GLOSSI [3], and Z-statistic method [4].

We found type 1 error and power of the Z-statistic method depend largely on gene size. When causal gene set consists of genes with large gene size, Z-statistic method led very high type 1 error and power, and gave larger power as  $m$  is larger. So the results from Z-statistic method can have false positive especially when gene set has larger genes.

However, SNP-PRAGE gave the consistent results irrespective of gene size. As  $m$  is larger, SNP-PRAGE has a little larger power. Based on the results, SNP-PRAGE has comparable performance to GLOSSI and GSEA in terms of power and size.

Z-statistic method has shortest computing time for analysis. It is because they do not consider LD structure between SNPs. Among methods which consider LD between SNPs, SNP-PRAGE has shortest computing time and the nonparametric GSEA method takes 18.5 times computational efforts than SNP-PRAGE based on our simulation results. When considering the computing time for the single SNP analysis

in KARE data is more than 300 times than one in simulation data, GSEA method will take very long period of computing time.

## 5 Discussion

We compared the performance of three parametric test-based methods (Z-statistic method, GLOSSI, SNP-PRAGE) and one nonparametric test-based method (GSEA) for the test of the gene set. The Z-statistic method does not consider LD and reduce much computing time but may have lots of false positive results because of overestimated gene set statistics when the gene set has many large genes. We found that considering LD block of SNPs helps us to deal with the correlation between p-values appropriately for estimating the effective gene size. Multiplying the effective gene size to minimum p-value for the gene-level summary of SNP-PRAGE can reduce the false positive results from large gene size. SNP-PRAGE has comparable performance to GLOSSI and GSEA despite not undergoing permutation step which requires a lot of computing time

## References

1. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genomewide expression profiles. *Proc. Nat. Acad. Sci.* 102(43): 15545–15550 (2005)
2. Wang, K., Li, M. and Bucan, M.: Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet* 81(6): 1278–1283 (2007)
3. Chai, HS, Sicotte, H., Bailey, K., Turner, S., Asmann Y., Kocher, J.: GLOSSI: a method to assess the association of genetic loci-set with complex diseases. *BMC Bioinformatics* 10: 102 (2009)
4. Nam, D. Kim, J., Kim, S., Kim, S.: GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucl. Acids. Res.* 39(6) (2010)
5. Kim, S. and Volsky, D.: PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics* 6(1): 144 (2005)
6. Li, C., Sumou, I., Ya-guang, D., Ying, L., Jian-quang, Q., Chao-shu, T., Jun-bao, D.: Imbalance of endogenous homocysteine and hydrogen sulfide metabolic pathway in essential hypertensive children. *Chin. Med. J.* 120(5): 389-393 (2007)
7. Hassanain, HH, Gregg, D, Marcelo, ML, Zweier, JL, Souza, HP, Selvakumar, B, Ma, Q, Moustafa-Bayoumi, M, Binkley, PF, Flavahan, NA, Morris, M, Dong, C, Goldschmidt-Clermont, PJ.: Hypertension caused by transgenic overexpression of *rac1*. *Antioxid. Redox. Signal.* 9: 91-100 (2007)
8. Fortuno, MA, Ravassa, S, Fortuno, A, Zalba, G, Diez, J.: Cardiomyocyte Apoptotic Cell Death in Arterial Hypertension. *Hypertension.* 38: 1406-1412 (2001)
9. Sharkey, L., Kirchain, S., McCune, S., Simpson, G., Archambault, E., Boatright, N., Hicks, E., Fray, J.: Progesterone increases blood pressure in spontaneous gestational hypertension in rats. *Am. J. Hypertension.* 18(1): 36-43 (2005)

## P.R.E.S.S. – An R-package for Exploring Residual-Level Protein Structural Statistics

Yuanyuan Huang<sup>1,3</sup>, Steve Bonett<sup>2</sup>, and Zhijun Wu<sup>1,3\*\*</sup>

<sup>1</sup> Program on Bioinformatics and Computational Biology

<sup>2</sup> Summer REU Program on Computational Systems Biology

<sup>3</sup> Department of Mathematics

Iowa State University, Ames, IA 50014, U.S.A.

{sunnyuan,zhijun}@iastate.edu, sbonett@gmail.com

**Abstract.** P.R.E.S.S. is an R package developed to allow researchers to get access to and manipulate on a large set of statistical data on protein residue-level structural properties such as residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. A large set of high-resolution protein structures are downloaded and surveyed. Their residue-level structural properties are calculated and documented. The statistical distributions and correlations of these properties can be queried and displayed. Tools are also provided for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, new tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. P.R.E.S.S. will be released in R as an open source software package, with a user-friendly GUI interface, accessible and executable by a public user in any R environment.

**Key words:** Protein structure analysis, protein residual-level structural properties, structural bioinformatics, statistical potentials, structural correlation plots

### 1 Introduction

The atomic-level structural properties of proteins, such as bond lengths, bond angles, and torsion angles, have been well studied and understood based on either chemistry knowledge or statistical analysis. Similar properties on the residue-level, such as the distances between two residues and the angles formed by short sequences of residues, can be equally important for structural analysis and modeling, but they have not been examined and documented on a similar scale. While these properties are difficult to measure experimentally, they can be statistically estimated in meaningful ways based on their distributions in known protein structures.

In our recent work [1], we have downloaded a large number of high-resolution X-ray structures from PDB Data Bank [2], and collected and analyzed several

---

\*\* Corresponding author



important residue-level structural properties including the distances between two neighboring residues; the angles formed by three residues in sequence; and the torsion angles of four residues in sequence. We call them, respectively, the residue level virtual bond lengths, virtual bond angles, and virtual torsion angles. We have examined the statistical distributions of these virtual bonds and virtual angles in known protein structures. In a four-residue sequence, there are two virtual bond angles and one torsion angle in between. We name them, according to their order in the sequence, the  $\alpha$ -angle,  $\tau$ -angle, and  $\beta$ -angle, where  $\tau$  is the torsion angle (Fig. 1a). In a five-residue sequence, there are three virtual bond angles and two torsion angles. We name them, according to their order in the sequence, the  $\alpha$ -angle,  $\tau_1$ -angle,  $\beta$ -angle,  $\tau_2$ -angle,  $\gamma$ -angle, where  $\tau_1$  and  $\tau_2$  are torsion angles (Fig. 1b). For these sequences, we have investigated the correlations among some of associated angles and in particular, the  $\alpha$ - $\tau$ - $\beta$  correlations for four-residue sequences and  $\tau_1$ - $\beta$ - $\tau_2$  correlations for five-residue sequences. We have shown that the distributions of residue distances and angles may vary with varying residue sequences, but in most cases, are concentrated in some high probability ranges, corresponding to their frequent occurrences in either  $\alpha$ -helices or  $\beta$ -sheets in proteins. We have shown that between  $\alpha$  and  $\tau$  angles and  $\tau$  and  $\beta$  angles, there exist strong correlations, which suggests that proteins follow certain rules to form their residue level angles as well, just like those for their atomic level  $\phi$ - $\psi$  angles. To the authors knowledge, these properties have not been discovered and documented before, but can be very valuable in applications [1]. In this paper, we describe a related piece of work with [1] on developing a software package called P.R.E.S.S. for direct access to the statistical data on the residue-level structural properties we have collected and analyzed. The software is developed in R [3] and will be released as an open source package, with a user-friendly GUI interface, accessible and executable by a public user in any R environment. With this software, the distributions and correlations of given types of residue distances or angles can all be retrieved and displayed. Tools are also provided in P.R.E.S.S. for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. We describe the organization of the software, the data source, the computational methods, and all the functional modules. We provide examples to demonstrate the use of the software.

## 2 Sytem Organization and Interface

P.R.E.S.S. can be divided into two ends, front end and back end. The back end includes the parts for downloading structural data, calculating residue distances and angles, and saving the distances and angles. The front end is responsible for providing all data retrieving and analysis functions using the distance and angle data calculated and saved in the back end. In the back end, there are three major components: 1). Download the structural data from PDB Data

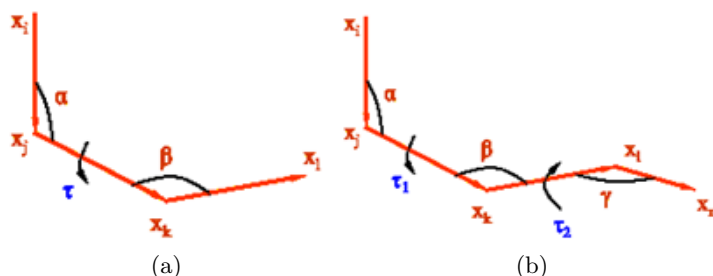


Fig. 1: **Residue distances and angles.** (a) The  $\alpha$ - $\tau$ - $\beta$  angle triplet in a four-residue sequence. (b) The  $\alpha$ - $\tau_1$ - $\beta$ - $\tau_2$ - $\gamma$  angle quadruple in a five-residue sequence. The residues are assumed to be located at  $x_i, x_j, x_k, x_l, x_m$ .

Bank, which can be down automatically or manually, but currently only semi-automatically. 2). Compute and collect residue-level distances and angles. 3). Save the distances and angles into five databases. They are database for virtual bond lengths, database for virtual bond angles, database for virtual torsion angles, database for virtual angle sequences of four-residue sequences, and database for virtual angle sequences of five-residue sequences, named B, A, T, ATA, and TAT databases, respectively. More specific information on the content of each of these databases is given below.

**B-database:** Stores the virtual bond lengths for all the neighboring pairs of residues for each downloaded structure. Each record in the database contains the following information:

Protein ID	Residue 1	Residue 2	1-2- distance
------------	-----------	-----------	---------------

**A-database:** Stores the virtual bond angles formed by all the connected triplet of residues of each downloaded structure. All residue-level 1-3-distances are also saved. Each record in the database contains the following information:

Protein ID	Residue 1	Residue 2	Residue 3	Bond Angle	1-3- distance
------------	-----------	-----------	-----------	------------	---------------

**T-database:** Stores the virtual torsion angles ( $\tau$ ) formed by all the connected quadruplets of residues of each downloaded structure. All residue-level 1-4-distances are also saved. Each record in the database contains the following information:

Protein ID	Residue 1	Residue 2	Residue 3	Residue 4	$\tau$	1-4 distance
------------	-----------	-----------	-----------	-----------	--------	--------------

**ATA-database:** Stores the  $\alpha$ - $\tau$ - $\beta$  angle sequences for all the four-residue sequences in each downloaded structure. Each record in the database contains the following information:

Protein ID	Residue 1	Residue 2	Residue 3	Residue 4	$\alpha$	$\tau$	$\beta$
------------	-----------	-----------	-----------	-----------	----------	--------	---------

4

**TAT-database:** Stores the  $\alpha$ - $\tau_2$ - $\beta$ - $\tau_2$ - $\gamma$  angle sequences for all the five-residue sequences in each downloaded structure. Each record in the database contains the following information.

Protein ID	Residue 1	Residue 2	Residue 3	Residue 4	Protein 5	$\alpha$	$\tau_1$	$\beta$	$\tau_2$	$\gamma$
------------	-----------	-----------	-----------	-----------	-----------	----------	----------	---------	----------	----------

In the front end, there are two major components: the GUI interface and the computational unit. The GUI interface takes a query from the user and passes it to the computational unit. The computational unit has a collection of routines, responsible for various computational tasks. It retrieves the data from the databases in the back end, performs certain calculations, and returns the results to the GUI interface. The interface then displays the results. More specifically, the GUI interface shows a window of six functional panels (Fig. 2), each accepting a specific type of queries: 1). Queries on virtual bond lengths for two residues. 2). Queries on virtual bond angles for three residues. 3). Queries on virtual torsion angles and ATA correlations for four residues. 4. Queries on TAT correlations for five residues. 5. Structural analysis and evaluation. 6. Help information. The overall system organization of P.R.E.S.S. is shown in Fig. 3. We

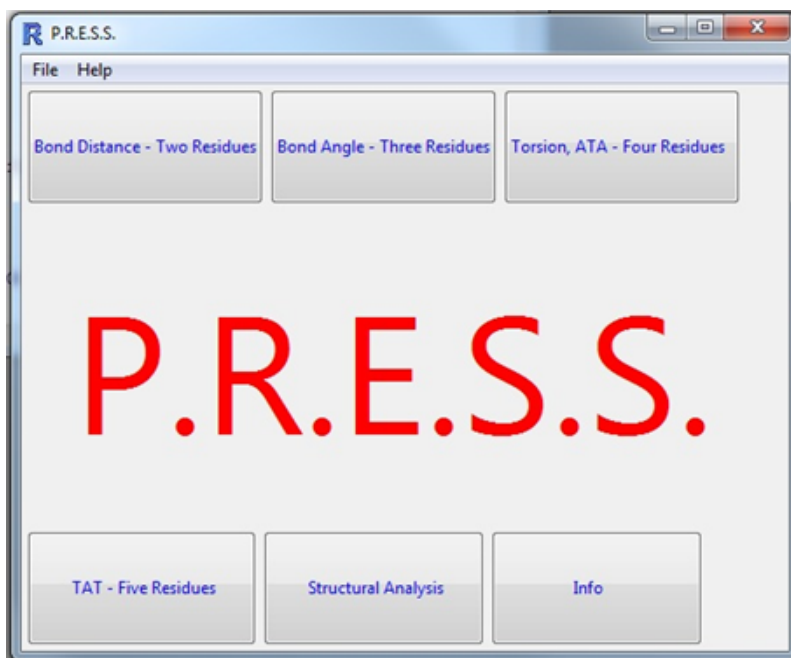


Fig. 2: **P.R.E.S.S. graphics interface.** PRESS has a graphics interface with six functional panels corresponding six functional routines, each providing a specific structural computing or analysis function.

describe the data source and computational methods used in P.R.E.S.S. in the following section.

### 3 Functional Modules

#### 3.1 Distribution of Virtual Bond Lengths

One of functions of P.R.E.S.S. is to retrieve the virtual bond lengths for a given pair of residues and find the distribution of the particular bond length over a certain distance range. The found distribution can be displayed in a graph as shown in Fig. 4. The residue pair to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the bond lengths.

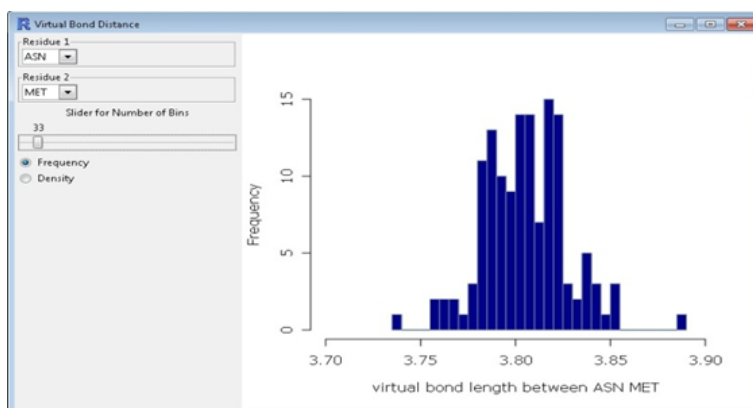


Fig. 3: **Distribution of virtual bond lengths.** This snapshot shows the distribution graph for the virtual bond lengths between ASN and MET. The users can not only move the slider to adjust the bin size of the histogram, but also switch between frequency and density displays.

#### 3.2 Distribution of Virtual Bond Angles

One of functions of P.R.E.S.S. is to retrieve the virtual bond angles for a given sequence of three residues and find the distribution of the particular bond angle over a certain angle range. The found distribution can be displayed in a graph as shown in Fig. 5. The residue triplet to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the bond angles.

6

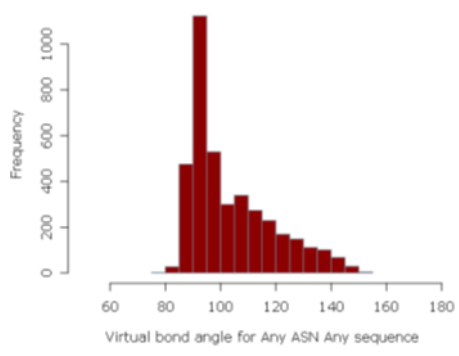


Fig. 4: **Distribution of virtual bond angles formed by Any, ASN, and Any.** This snapshot shows the distribution graph for the virtual bond angles formed by a residue sequence Any, ASN, and Any.

### 3.3 Angle-Distance Correlations

When the distribution of a virtual bond angle is queried, an option is available for displaying the correlation between the bond angle and the corresponding residue 1-3-distance. This correlation can be requested for any sequence of three residues, as shown in Fig. 6.

### 3.4 Distribution of Virtual Torsion Angles

The virtual torsion angles for a given sequence of four residues can be retrieved. The distribution of the particular torsion angle can be displayed over a certain angle range. The residue quadruplet to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the torsion angles. The graph can be displayed along with the distributions of the neighboring virtual bond angles ( $\alpha, \beta$ ), as shown in Fig. 7. The density distribution of the angle sequence  $\alpha\text{-}\tau\text{-}\beta$  can be displayed as a 3D plot in  $\alpha\text{-}\tau\text{-}\beta$  space, as shown Fig. 8. The correlation between the virtual torsion angle and the corresponding residue 1-4 distance for a given sequence of four residues can also be displayed.

### 3.5 Correlation of Virtual Torsion Angles

The angle sequence  $\alpha\text{-}\tau_1\text{-}\beta\text{-}\tau_2\text{-}\gamma$  for a given sequence of five residues can be retrieved. The density distributions of the virtual bond angle  $\beta$  and its neighboring two virtual torsion angles can be displayed. The graphs can be displayed in a matrix of plots, as shown in Fig. 9. The density distribution of  $\tau_1\text{-}\beta\text{-}\tau_2$  can also be displayed as a 3D plot in the  $\tau_1\text{-}\beta\text{-}\tau_2$  space as shown Fig. 10.

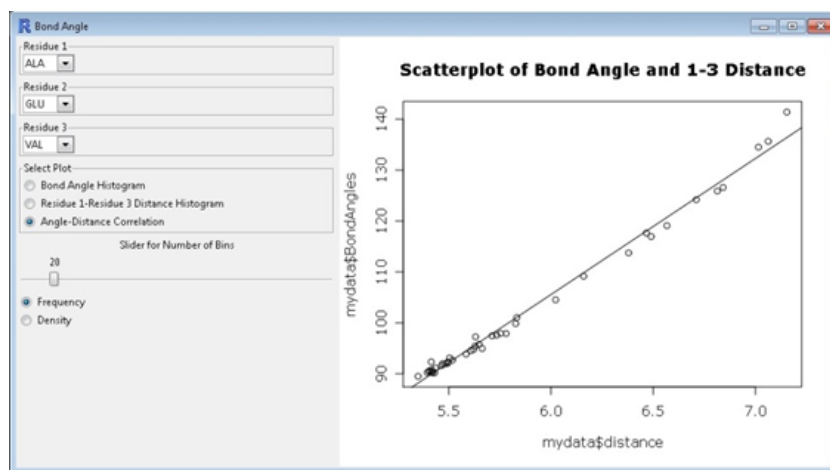


Fig. 5: Scattered plot of the virtual bond angles against their residue 1-3 distances. This snapshot shows the distribution graph for the angle-distance pairs for residues ALA, GLU, and VAL.

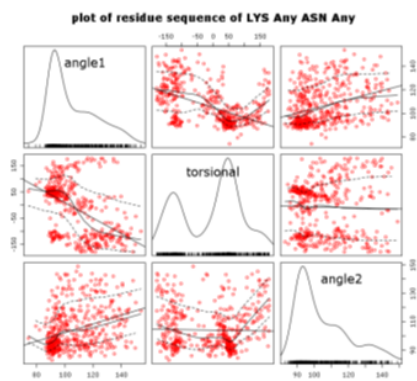


Fig. 6: A matrix of scattered plots of the distributions and correlations of the virtual torsion angle and its two neighboring virtual bond angles. The plots show the distribution and correlation graphs for the virtual torsion angle and its two neighboring virtual bond angles for residues LYS, Any, ASN, Any. The matrix of plots is 3 by 3. The graph in each is defined as follows: Square(1,1) = distribution of virtual bond angle 1; Square(2,2) = distribution of the virtual torsion angle; Square(3,3) = distribution of virtual bond angle 2; Square(1,2) = correlation between virtual bond angle 1 and the virtual torsion angle; Square(1,3) = correlation between the bond angle 1 and virtual bond angle 2; Square(2,3) = correlation between the virtual torsion angle and virtual bond angle 2.

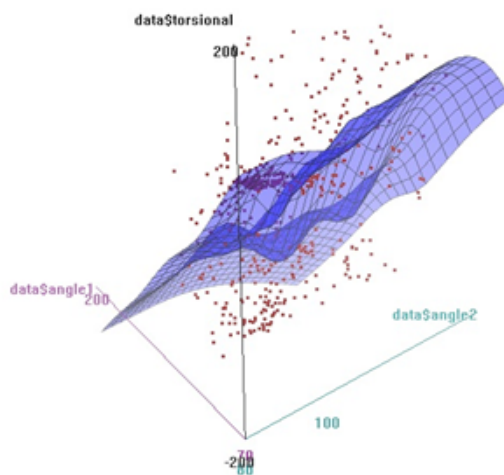


Fig. 7: **3D scattered plot for the virtual torsion angle and its two neighboring virtual bond angles.** This snapshot shows the density distribution of the  $\alpha$ - $\tau$ - $\beta$  angle triplets for residue sequence LYS, Any, ASN, and Any in the  $\alpha$ - $\tau$ - $\beta$  space. A lowess approximation to the distribution is also plotted.

### 3.6 Structural Analysis – Computation of Statistical Potentials

One of important functions of P.R.E.S.S. is to evaluate the statistical potentials on the virtual bonds or virtual bond angles for a given structure (Fig. 11). The potentials are defined in terms of the statistical distributions of the virtual bond lengths and virtual bond angles. The virtual bond length potential can be evaluated for every neighboring pair of residues of the given structure. Therefore, the distribution of the potential energy along the residue sequence of the structure can be obtained and displayed to show how flexible the virtual bonds are along the sequence. The higher the potential energy is for a specific bond, the lower the probability of the bond length is in the distribution of the bond length in known proteins, and hence the more deviated it must be from its average value the bond length (Fig. 12). The virtual bond angle potential can be evaluated for every sequence of three residues of the given structure as well. The distribution of the potential energy along the residue sequence of the structure can also be obtained and displayed to show how flexible the virtual bond angles are along the sequence. It has the same property as that for the bond length energy for structural evaluation (Fig. 13).

### 3.7 Structural Analysis – Residue Angle-Angle Correlation Plots

One of the most important functions of P.R.E.S.S. is that it can evaluate the correlations of the virtual bond and torsion angles and display a residue-level

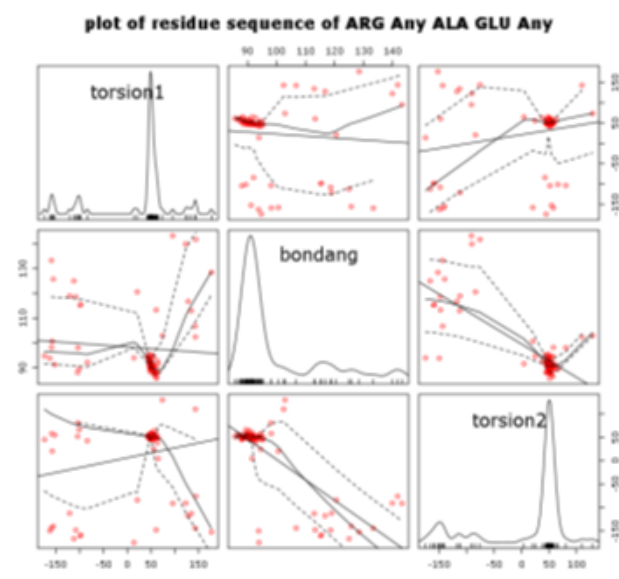


Fig. 8: A matrix of scattered plots for the distributions and correlations of virtual bond and torsion angles  $\tau_1$ - $\beta$ - $\tau_2$ . The plots show the distribution and correlation graphs for the virtual bond and torsion angles  $\tau_1$ - $\beta$ - $\tau_2$  for residues ARG, Any, ALA, GLU, and Any. The matrix of plots is 3 by 3. The graph in each is defined as follows: Square(1,1) = distribution of virtual torsion angle  $\tau_1$ ; Square(2,2) = distribution of virtual bond angle  $\beta$ ; Square(3,3) = distribution of virtual torsion angle  $\tau_2$ ; Square(1,2) = correlation between  $\tau_1$  and  $\beta$ ; Square(1,3) = correlation between  $\tau_1$  and  $\tau_2$ ; Square(2,3) = correlation between  $\beta$  and  $\tau_2$ .



10

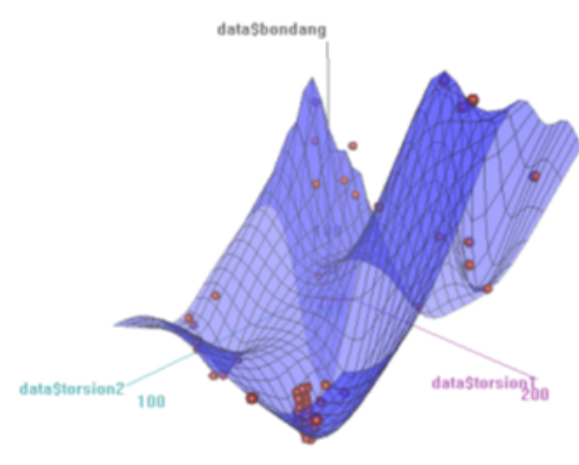


Fig. 9: **scattered plot of virtual bond and torsion angles.** This snapshot shows the density distribution of the  $\tau_1$ - $\beta$ - $\tau_2$  angle sequence in a  $\tau_1$ - $\beta$ - $\tau_2$  space for a residue sequence ARG, Any, ALA, GLU and Any. A loess approximation to the distribution is also plotted.

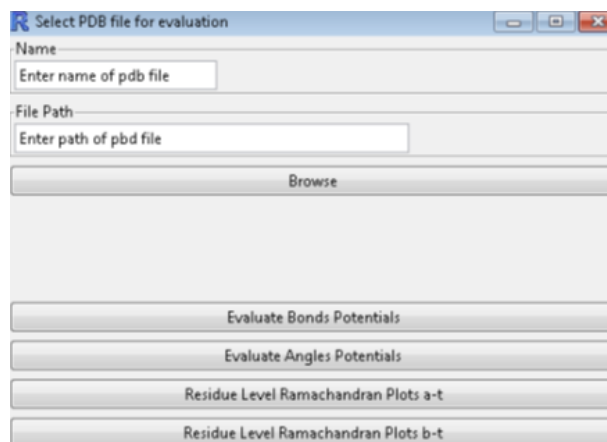


Fig. 10: **Virtual bond length potentials.** A window is popped out for the user to upload the structural file. The system can then evaluate the virtual bond length potential for each neighboring pair of residues in the protein sequence and display the distribution of the potential energy over the residue sequence.

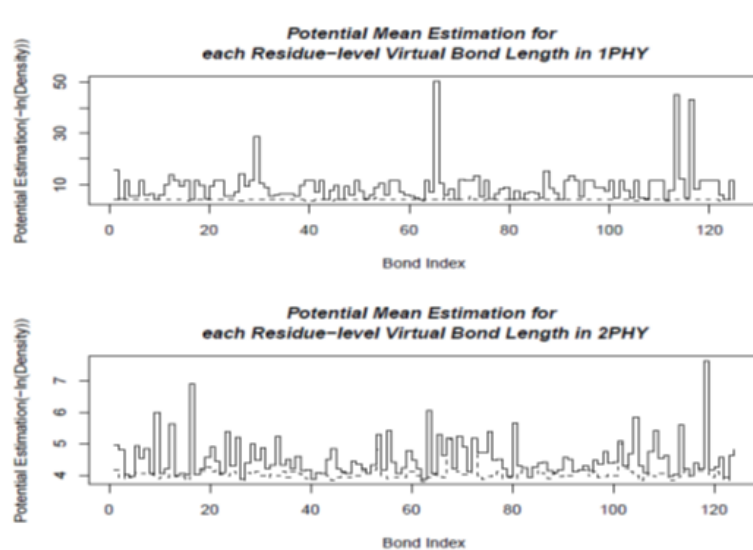


Fig. 11: **Distributions of virtual bond energies for 1PHY (2.4Å) and 2PHY (1.4Å).** The energy levels of the virtual bond lengths of two structures 1PHY and 2PHY are shown in solid lines. The minimal possible energies are plotted as the dashed line. If there is no distribution data for some virtual bond, such as the bond at index 98, the potential function is not defined, and there is a gap in the energy plot for that bond. These two structures are determined with different resolutions for the same protein. The better-resolved structure (2PHY) has lower potential energies in average than the poorly determined one (1PHY).

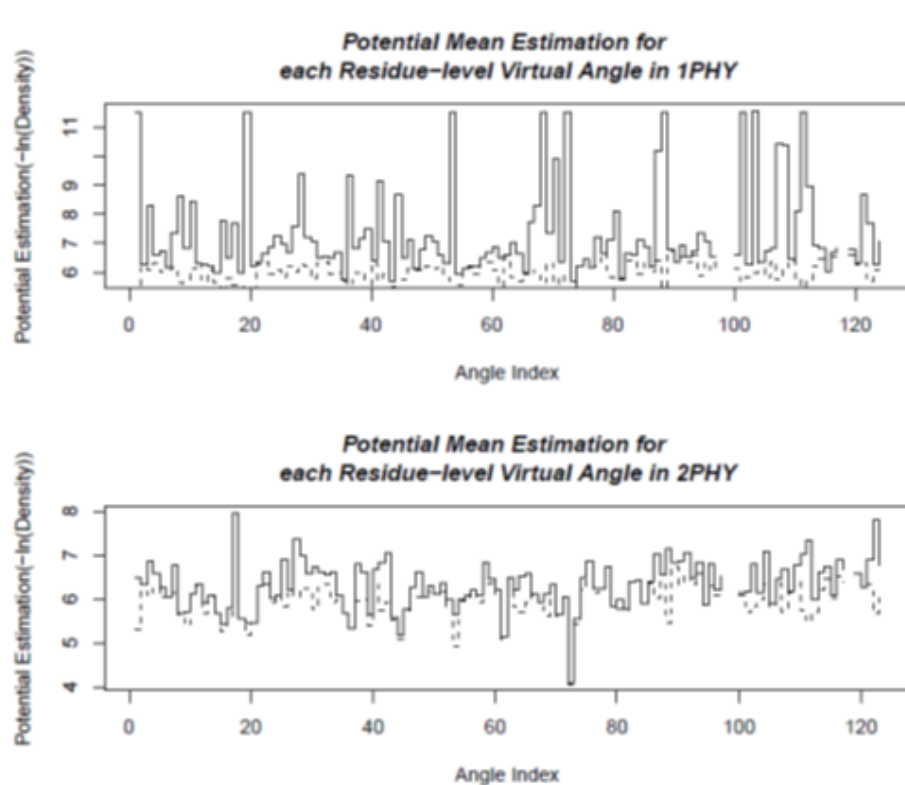


Fig. 12: **Distribution of virtual bond length and bond angle energies for 1PHY (2.4Å) and 2PHY (1.4Å).** The energy levels of the virtual bond angles of two structures 1PHY and 2PHY are plotted in solid lines. The minimal possible energies are shown as the dashed line. These two structures are determined with different resolutions for the same protein. The better-resolved structure (2PHY) has lower potential energies in average than the poorly determined one (1PHY).

Ramachandran-like plot for a given structure. Two of the angle-angle correlation plots are proven to be especially valuable. One is the  $\alpha$ - $\tau$  correlation plot or the AT-plot for short. Another one is the  $\tau$ - $\beta$  correlation plot or the TB-plot for short. Given a protein structure, the  $\alpha$ - $\tau$  or  $\tau$ - $\beta$  angle pairs can be computed along the residue sequence for the structure. Each angle pair can be plotted as a dot in the  $\alpha$ - $\tau$  or  $\tau$ - $\beta$  space. The distribution of the dots over the contour of the general  $\alpha$ - $\tau$  or  $\tau$ - $\beta$  density distribution can then be evaluated to show how the angle pairs in the given structure correlated against to their average correlations in known proteins. These plots can be used effectively to differentiate high-quality structures from low-quality structures at the residue level as the Ramachandran plots for structural evaluation at the atomic level, as shown in Fig. 13 and 14.

### 3.8 Display of Residue Angle Correlations of a Structure

The contours of density distributions of  $\alpha$ - $\tau$  and  $\tau$ - $\beta$  angle pairs can be plotted in 2D  $\alpha$ - $\tau$  and  $\tau$ - $\beta$  angle planes. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured, Favoured, and Allowed, corresponding to regions of high 50%, 75%, and 90%, respectively. The  $\alpha$ - $\tau$  or  $\tau$ - $\beta$  angle pairs for every sequence of four residues of a given structure can be computed and plotted in the  $\alpha$ - $\tau$  or  $\tau$ - $\beta$  plane, on top of the contour of the general  $\alpha$ - $\tau$  or  $\tau$ - $\beta$  density distribution function. The structure is considered to be well formed in terms of its virtual bond angles and virtual torsion angles if most of the plotted dots are in the high-density regions of the  $\alpha$ - $\tau$  or  $\tau$ - $\beta$  density distribution contour.

## 4 Summary and Discussion

In this paper, we have reported our recent work for the development of an R package, called P.R.E.S.S., which allows researchers to get access to and manipulate on a large set of statistical data on protein residue-level structural properties such as residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. We have downloaded and surveyed a large set of high-resolution protein structures, and calculated and documented an important set of their residue-level structural properties in P.R.E.S.S. With P.R.E.S.S., the statistical distributions and correlations of these properties can be queried and displayed. Tools are also provided for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, new tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. We have discussed the principle for the development of P.R.E.S.S. for statistical analysis on protein structures. We have described the system organization and interface of the software, and provided detailed information on how the structural data was collected and documented in P.R.E.S.S., and how all the statistical results were calculated. We have described the major computational and analysis functions of

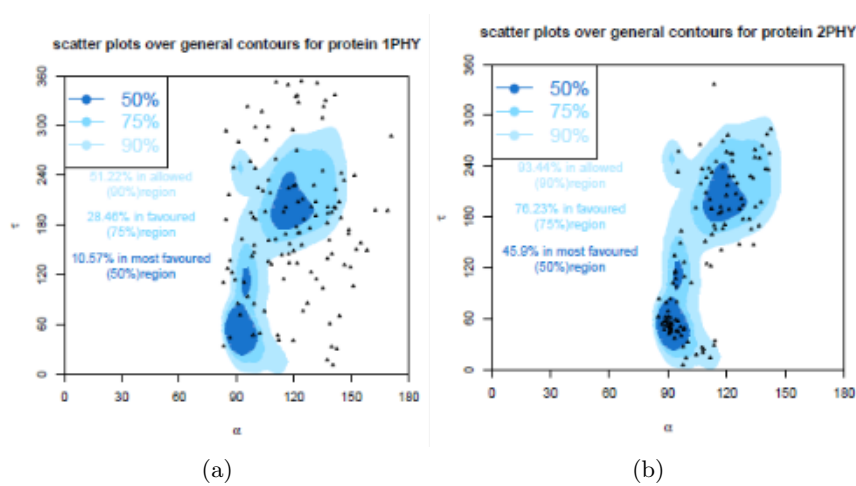


Fig. 13: **The  $\alpha$ - $\tau$  correlation plots for a protein at two different resolutions.** The photoreactive yellow protein in the dark state, 1PHY (2.7Å), shown in (a) compared with the DD-peptidase, 2PHY (1.4Å), shown in (b). The background contours are generated from the general density distributions of the  $\alpha$ - $\tau$  angle pairs in known proteins. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured (high 50% density), Favoured (high 75% density), and Allowed (high 90% density) regions. The scattered triangles correspond to the  $\alpha$ - $\tau$  angle pairs in the given protein structures. The lines in (a) indicate that there are 51.22% of the triangles of the  $\alpha$ - $\tau$  angles pairs in 1PHY falling in the 90% region, 28.46% of triangles falling in the 75% region, and only 10.57% of the triangles falling in the 50% region. On the other hand, In (b), there are 93.44% of the triangles of the  $\alpha$ - $\tau$  angles pairs in 2PHY falling in the 90% region, 76.23% of triangles falling in the 75% region, and 45.9% of the triangles falling in the 50% region.

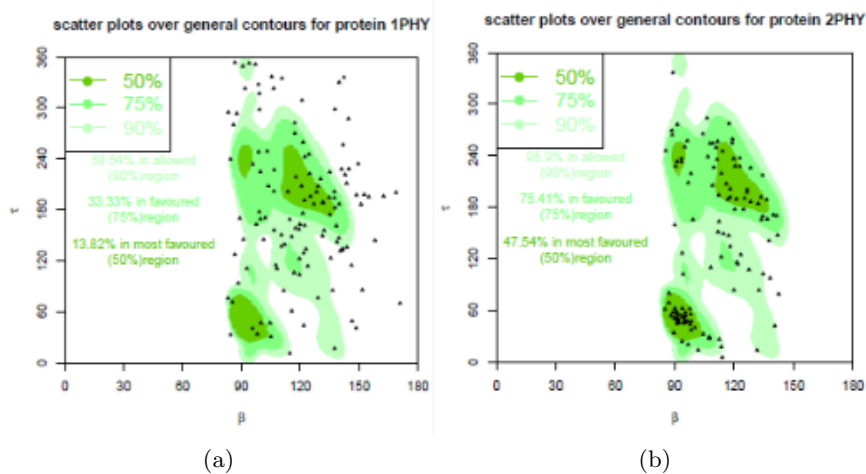


Fig. 14: **The  $\tau$ - $\beta$  correlation plots for a protein at different resolution.** The photoreactive yellow protein in the dark state, 1PHY (2.7Å), shown in (a) compared with the DD-peptidase, 2PHY (1.4Å), shown in (b). The background contours are generated from the general density distributions of the  $\tau$ - $\beta$  angle pairs in known proteins. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured (high 50% density), Favoured (high 75% density), and Allowed (high 90% density) regions. The scattered triangles correspond to the  $\tau$ - $\beta$  angle pairs in the given protein structures. The lines in (a) indicate that there are 58.54% of the triangles of the  $\tau$ - $\beta$  angle pairs in 1PHY falling in the 90% region, 33.33% of the triangles falling in the 75% region, and only 13.82% of the triangles falling in the 50% region. On the other hand, in (b), there are 95.9% of the triangles of the  $\tau$ - $\beta$  angle pairs in 2PHY falling in the 90% region, 75.41% of the triangles falling in the 75% region, and 47.54% of the triangles falling in the 50% region.

P.R.E.S.S. and demonstrated them in many examples. P.R.E.S.S. will be released in R as an open source software package, with a user-friendly GUI interface, accessible and executable by a public user in any R environment. The statistical distributions of residue-level distances and angles in known protein structures provide a valuable source of information for estimating these residue level structural properties of proteins, which are not otherwise accessible experimentally. However, these statistical measures rely upon the quality as well as quantity of the sampled known structures. We have downloaded around one thousand high-quality structures from the PDB Data Bank, which should be sufficient to obtain reliable statistical estimates of the distributions of virtual bond lengths, virtual bond angles, virtual torsion angles, and some of their correlations, but of course there is the possibility that for some cases of specific residue sequences, the values might deviate from the overall characteristic distributions. In P.R.E.S.S., we have provided information about the size of the data set for each estimate. The useful tool from this study is a residue-level Ramachandran-type of plot for correlations between pairs of neighboring virtual bond angles and virtual torsion angles. Several examples have been given in the present paper, but these differ from the atomic-level Ramachandran Plot in an important way, because the density distribution contours of these residue-level angles show relatively larger deviations. Thus their use requires specifying more precisely what density regions should be permitted for high-quality structures. Further evaluations are needed to decide generally what these evaluation criteria should be.

## 5 Acknowledgements

This work is partially supported by the NIH/ /NIGMS grant R01GM081680 and by the NSF/ /DMS grant DMS0914354.

## References

1. Huang, Y., Bonett, S., Kloczkowski, A., Jernigan, R., and Wu, Z., Statistical measures on protein residue-level structural properties, *J. Struct and Funct Genomics* (2011), DOI: 10.1007/s10969-011-9104-4
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., The Protein Data Bank, *Nucleic Acids Research* 28, 235-242, 2000.
3. Bio3D: An R package for the comparative analysis of protein structures. Grant, Rodrigues, ElSawy, McCammon, Caves, (2006) *Bioinformatics* 22, 2695-2696
4. Doreleijers, J. F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Makley, J. L., and Ulrich, E. L., BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank, *J. Biomol. NMR* 26, 139-146, 2003.
5. Bourne, P. E. and Weissig, H. *Structural Bioinformatics*. John Wiley & Sons, Inc., 2003.

## Efficient Error Correction for Deep Sequencing of Viral Amplicons

Pavel Skums<sup>1</sup>, Zoya Dimitrova<sup>1</sup>, David Campo<sup>1</sup>, Gilberto Vaughan<sup>1</sup>, Livia Rossi<sup>1</sup>, Joseph Forbi<sup>1</sup>, Jonny Yokosawa<sup>2</sup>, Alex Zelikovsky<sup>3</sup> and Yury Khudyakov<sup>1</sup>

<sup>1</sup>Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, USA {kki8, izd7, fyv6, jiv9, fld5,gzf7,yek0}@cdc.gov. <sup>2</sup>Universidade Federal de Uberlândia, Brazil, jyokosawa@icbim.ufu.br. <sup>3</sup>Georgia State University, Atlanta, GA {AlexZ@cs.gsu.edu}

**Abstract.** We present two new highly efficient error correction algorithms: (i) k-mer - based error correction (KEC); and (ii) empirical frequency threshold (ET). Both were compared to the recently published algorithm SHORAH to evaluate the relative performance using 24 experimental datasets obtained by 454-sequencing of amplicons with known sequences. We found that all three algorithms showed similar performance in terms of finding true sequences, but KEC and ET methods significantly outperformed SHORAH both in terms of their ability to remove false sequences and to estimate the frequency of true ones.

**Keywords:** HCV, quasispecies, pyrosequencing, error correction

### 1 Introduction

Hepatitis C virus (HCV) shows a very high level of sequence heterogeneity, which is responsible for its escape from neutralizing host immune responses and rapid development of drug resistance. Recent advances in high-throughput (HT) sequencing methods allow for analysis of the unprecedented number of HCV-genomic sequence variants from infected patients and present a novel opportunity for understanding HCV evolution, drug resistance and immune escape. However, owing to the massive scale of sequencing, sequence errors generated during HT sequencing require extensive computational processing with error correction algorithms in order to obtain high quality reads for genetic analysis. The key purpose of such algorithms is to discriminate between artifacts and actual sequences. This task becomes especially challenging for recognizing and preserving low-frequency natural variants in viral population.

SHORAH [5][6] is currently one of the best error correction algorithms available. It uses probabilistic clustering approach based on the Dirichlet process mixture. Another approach to error correction is based on the use of k-mers, or substrings of reads of a fixed length k [2][3][4]. These algorithms have good performance but high time- and memory-consumption needs, together with the possibility of errors introduced during the correction phase [5]. To overcome these disadvantages, the authors of EDAR algorithm [1] developed an approach for the detection and deletion



of sequence regions containing errors. This error deletion works well for shotgun experiments, but is unacceptable for the small amplicon reads commonly analyzed in viral samples.

In this paper, we present two new efficient error correction algorithms: (i) k-mer-based error correction (KEC); and (ii) empirical frequency threshold (ET). KEC uses the EDAR algorithm optimized for amplicon sequencing for the detection of error regions and a novel algorithm for correction of errors associated with homopolymers. KEC does not require a reference sequence and is, therefore, suitable for *de-novo* sequencing. The ET algorithm uses estimation of a frequency threshold for indels and haplotypes calculated from experimentally obtained clonal sequences, also correcting homopolymers. Both algorithms were compared to SHORAH to evaluate their relative performance using 25 experimental amplicon datasets with known sequences obtained using 454 sequencing.

## 2 Algorithms description

### 2.1. KEC algorithm

The scheme of KEC includes 4 steps: (1) Calculate k-mers and their frequencies (k-counts). We assume that k-mers with high k-counts (“solid” k-mers) are correct, while k-mers with low k-counts (“weak” k-mers) contain errors. (2) Determine the threshold k-count (error threshold), which distinguishes solid k-mers from weak k-mers. (3) Find error regions. The error region is the segment  $[i,j]$  of read such that for every  $p \in [i,j]$  the k-mer starting at the position  $p$  is considered weak. (4) Correct the errors in error regions.

Methods proposed in EDAR were used for steps 1 and 3. However, they were optimized using efficient data structures based on hash maps. The error threshold estimation from [1] is not applicable to the amplicon data. It was replaced by an algorithm based on the detection of local minima in smoothed distributions. We call error region  $x=[b,e]$  of a read  $r$  a tail, if either  $b = 1$  or  $e = n-k+1$  ( $n$  is the length of  $r$ ). Let  $l(x)$  be the length of  $x$ , and  $h_i(w)$  be a homopolymer of length  $i$  composed of nucleotide  $w \in \{A,T,G,C\}$ .

**Claim 1.** *Suppose, that the non-tail error region  $x$  was caused by a one-nucleotide error  $E$ . Let  $w$  be the last nucleotide of  $x$ . If  $E$  is a replacement, then  $l(x) = k$ . If  $E$  is an insertion in the homopolymer of length  $r$  ( $0 \leq r \leq k$ ), then  $l(x) = k-r+1$ ,  $x$  is followed by a homopolymer  $h_{r-1}(w)$ . If  $E$  is a deletion in the homopolymer of length  $m$ , then  $l(x) = k-m-1$  and if  $m \geq 1$ , then  $x$  is followed by a homopolymer  $h_m(c)$ , where  $c \neq w$ .*

Errors were identified and corrected in non-tail error regions using Claim 1, and then the corresponding prefixes or suffixes were deleted from reads for tails. The procedure was repeated until the dataset had no errors or the specified number of iterations was reached. Claim 1 considers only error regions with  $l(x) \leq k$ . The longer error regions correspond to the occurrence of  $>1$  errors separated by  $\leq k$  nucleotides. We found that this type of error is much less frequent and we correct it by a heuristics based on Claim 1.

**2.2. ET algorithm**

The key idea of the procedure is to calculate the frequency of erroneous sequences in amplicon samples where only a single sequence was expected. Each single-clone sample was processed in the following way: First, each sequence is aligned against a set of external references of all known genotypes. For each sequence the best match of the external set is chosen. The aligned sequence is clipped to the size of the chosen external reference. The 20 most frequent sequences that do not create insertions or deletions are selected, constituting the internal reference set. Each sequence is aligned against each member of the internal references set and its best match is chosen.

The frequency of erroneous indels and its standard deviation (s.d.) was calculated over all nucleotide positions for 15 single-clone samples. An indel threshold was defined as the average frequency of erroneous indels + 5 s.d. If a sequence contained an indel with a frequency lower than the threshold, the sequence was removed. Then all homopolymers of at least 4 nucleotides were identified, followed by removal of the insertions and replacement of the deletions by the repeated nucleotide. The frequency of erroneous sequences and its s.d. were calculated over the 15 single-clone samples. A sequence threshold was defined as the average frequency of erroneous sequences + 5 s.d. All sequences with a frequency lower than the threshold were removed. This procedure was applied to each mixture sample.

**3 Algorithms comparison**

Individual plasmid clones (n=10) containing different HCV hypervariable region 1 sequences were purified and sequenced using dye-terminator sequencing. A set of plasmid samples was generated. 14 samples contained a single clone. 10 samples contained 8 clones mixed together in different proportions (from 1% to 93%). The E1/E2 region (309 nt) was amplified from each sample and sequenced using GS FLX Titanium Series Amplicon kits. Low quality reads were removed using the GS Run Processor (Roche, 2010). Each sequence file was then analyzed using ET, KEC or SHORAH error correction algorithms. SHORAH was applied several times under different parameters and the best attained results are reported here. All results are summarized in Table.

**Table.** Test results of the single-clone (S) and mixture (M; n=8) samples. MT: Missing true sequences; FS: False sequences; MSE: root mean square error; HD: Average Hamming distance, averaged over all false sequences.

	ET				KEC				SHORAH			
	MT	FS	MSE	HD	MT	FS	MSE	HD	MT	FS	MSE	HD
S1	0	0	0.00	0	0	1	4.67	1	0	351	29.02	4.84
S2	0	0	0.00	0	0	0	0.00	0	0	269	30.12	4.44
S3	0	1	1.09	1	0	2	4.93	1.5	0	292	23.44	5.31
S4	0	1	0.98	2	0	1	2.84	1	0	271	44.68	5.39
S5	0	0	0.00	0	0	0	0.00	0	0	319	9.63	4.47

## ISBRA 2011 Short Abstracts

S6	0	1	5.26	2	0	1	6.10	1	0	194	18.70	3.94
S7	0	0	0.00	0	0	1	5.80	1	0	496	21.52	6.70
S8	0	0	0.00	0	0	0	0.00	0	0	262	14.37	4.58
S9	0	0	0.00	0	0	0	0.00	0	0	183	6.23	6.97
S10	0	0	0.00	0	0	0	0.00	0	0	288	7.77	5.11
S11	0	1	0.53	2	0	0	0.00	0	0	717	24.71	5.03
S12	0	0	0.00	0	0	0	0.00	0	0	611	25.94	5.52
S13	0	0	0.00	0	0	0	0.00	0	0	156	5.53	4.93
S14	0	0	0.00	0	0	0	0.00	0	0	161	6.83	6.60
Mean	0.00	0.29	0.56	0.50	0.00	0.43	1.74	0.39	0.00	326.43	19.18	5.27
M1	0	0	1.17	0	0	1	0.87	1	0	320	1.23	4.51
M2	0	0	1.50	0	0	0	1.75	0	0	738	3.70	4.44
M3	0	0	2.92	0	0	0	3.55	0	0	638	3.65	4.25
M4	0	0	2.18	0	0	0	2.30	0	0	577	2.88	5.20
M5	0	0	0.34	0	7	0	7.00	0	0	214	0.91	7.37
M6	1	0	2.20	0	1	0	1.97	0	1	394	2.48	4.54
M7	0	0	1.20	0	0	0	1.97	0	0	499	2.04	5.00
M8	1	0	0.89	0	1	0	2.31	0	1	336	3.09	5.54
M9	0	0	2.23	0	6	0	9.25	0	0	643	6.56	4.49
M10	1	0	3.53	0	1	0	4.21	0	2	637	5.88	5.32
Mean	0.30	0.00	1.82	0.00	1.60	0.10	3.52	0.10	0.40	499.60	3.24	5.07

All methods found the correct sequence in each single-clone sample. However, ET and KEC retained the lower number of false sequences. Similarly, ET and KEC showed lower number of false sequences than SHORAH in each mixed samples. All three algorithms were successful in identifying most of true sequences, with ET being the most accurate. KEC did not detect true sequences representing ~1% in mixtures M5 and M9. The low root mean square error between observed and expected frequencies of true sequences indicates a high accuracy of ET and KEC, whereas SHORAH has much higher MSE, owing to the detection of a greater number of false sequences. Analysis of the Hamming distance between false sequences and their closest match shows that false sequences retained by KEC and ET are genetically closer to true sequences than sequences retained by SHORAH.

### 4 Conclusions

SHORAH, ET and KEC perform equally efficient in finding true sequences. However, KEC and ET outperform SHORAH in removing false sequences and estimating the sequence frequency. At the same time, in contrast to SHORAH and ET, KEC does not require a reference sequence. Both algorithms, KEC and ET, are highly

suitable for rapid recovery of high quality sequences from reads obtained by deep sequencing of genomic regions from heterogeneous viruses such as HCV and HIV.

## References

1. Zhao, X., Palmer, L., Bolanos, R., Mircean, C., Fasulo, D., Wittenberg, D.: EDAR: An efficient error detection and removal algorithm for next generation sequencing data. *Journal of computational biology*, 17(11), 1549 — 1560 (2010)
2. Chaisson, M.J., Brinza, D., Pevzner, P.A.: De novo fragment assembly with short mate-paired reads: does the read length matter? *Genome Res.* 19, 336–346 (2009).
3. Chaisson, M.J., Pevzner, P.A.: Short read fragment assembly of bacterial genomes. *Genome Res.* 18, 324–330 (2008).
4. Pevzner, P., Tang, H., Waterman M.: An Eulerian path approach to DNA fragment assembly: *Proc. Natl. Acad. Sci. USA*, 9748—9753 (2001).
5. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N.: Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *Journal of Computational Biology*, 17, 417—428 (2009).
6. Zagordi O, Klein R, Däumer M, Beerenwinkel N.: Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Research*, 38 (21), 7400—7409 (2010).

## An Efficient Constraint Planning Algorithm for Distributed Bio-Ontologies\*

Ming Fang, Weiling Li, Rajshekhar Sunderraman

Department of Computer Science,  
Georgia State University,  
Atlanta, GA 30303

{[mfang1@student.gsu.edu](mailto:mfang1@student.gsu.edu), [wli16@student.gsu.edu](mailto:wli16@student.gsu.edu), [raj@cs.gsu.edu](mailto:raj@cs.gsu.edu)}

**Abstract.** The data of Semantic Web exist in machine readable format called RDF, in order to promote data exchange on the web based on their semantics. Due to the nature of biological data, bio-ontologies tend to be very large, distributed, and interconnected. Thus, maintaining constraints and enforcing data consistency become very challenging. In previous study, we conducted a pioneer study and presented a framework for checking global constraints and ensuring integrity on data that span multiple ontologies. As an update is issued to a single site, global constraints that can be potentially violated are broken down into sub constraints that only involve a very small subset of ontologies. The checking of sub constraints runs effectively in parallel and returns results about each subset. The collection of these results determines the violation of global constraints.

In this work, we present an efficient constraint planning algorithm for distributed bio-ontologies. This algorithm serves as the key part of the global constraint checking framework. This algorithm takes a number of distributed but interconnected bio-ontologies and a set of global constraints expressed in logic programming as inputs, and produces a set of sub-constraints in Semantic Web query language SPARQL for constraint checking. An working example is presented at the end.

**Keywords:** Semantic Web, OWL, Distributed Bio-Ontology, Integrity Constraints.

### 1 Introduction

OWL ([1]) has been widely adopted in areas like science and commerce. One reason for its popularity is its ability to formally describe complex concepts and relationships among concepts. More importantly, OWL provides a way to facilitate automated reasoning at both the conceptual and the instance level ([2]). Although numerous bio-ontologies, such as Cancer Ontology, Gene Ontology, Human Disease Ontology, are available in RDF ([3]) form, there is still a huge demand for developing more OWL ontologies for various purposes.

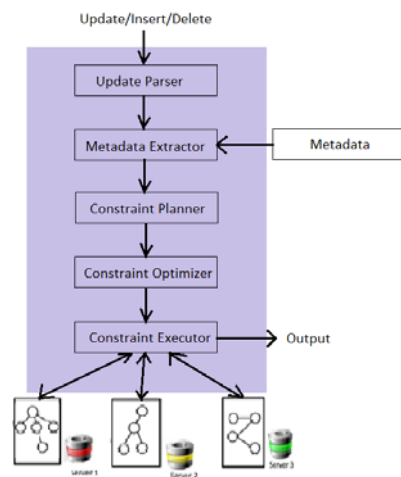
\* This work is supported by Molecular Basis of Disease, GSU

2 **Ming Fang, Weiling Li, Rajshekhar Sunderraman**

In previous studies, we uncovered the global constraint violation issue. As an initial investigation as well as an evidence to show ontologies are interdependent and dynamic, we explored a group of well-established and well-known ontologies from the biomedical field. We have investigated 81 bio-ontologies from The Open Biological and Biomedical Ontologies [4] and the Ontology Lookup Service [5]. In average, one biomedical ontology references to/depends on three other bio-ontologies. Although some ontologies in the set do not change very often, the majority of these ontologies updates 21.31 times/year, with 7010.57 lines/update in average.

Since there is no formal mechanism available to enforce data consistency among distributed ontologies while they are interconnected and constantly changing, we designed a frame that is faster and of less network traffic than the naive method. The infrastructure in figure 1 contains:

- **Update Parser:** Parses a user specified update, and return involved ontology objects.
- **Metadata Extractor:** Extract the set of global constraints that could be potentially violated by the update statement.
- **Constraint Planner:** runs an effective algorithm to generate sub constraints that will be dispatched to remote ontology sites.
- **Constraint Optimizer:** reorganizes the order of sub constraints in order to achieve higher efficiency.
- **Constraint Executor:** Execute sub constraints in parallel, and made decisions about the global constraint upon receiving the results of sub constraints.



**Fig. 1.** Internal Architecture of Constraint Checker

In section 2, we will introduce the constraint planning algorithm that is the central piece of the Constraint Planner. It is efficient because it breaks a global constraint that requires bringing in multiple larger ontologies into several sub-constraint

queries that are sent to remote ontologies in parallel. In section 3, we will provide a simple working example of this algorithm to conclude this poster.

## 2 The Constraint Planning Algorithm

The inputs of this algorithm contain an update statement **U** to ontology site **S**, and a list of global constraints **C**. As we step through the constraint checker to the constraint planning stage, we will also have the following two pieces of information available: (1) Ontology Object List (OOL) that identifies what predicates are to change and their new values; (2) Constraint-Source Table (CST) that specifies the sites involved in this update **U** for each global constraint in **C**. The output will be a list of sub-constraints for each global constraint in **C** affected by the update **U**.

### Constraint Planning Algorithm:

```

For each constraint c in the list of global constraints C
  For each site s from CST that is affected by constraint c in the update
    If site s is not where the update happens
      Then generate sub-constraints in the form of SPARQL
        queries from all the predicates (available in OOL) that
        reference to site s using appropriate conditions. Include
        arithmetic queries when necessary.
    Elseif site s is where the update happens
      Then If there are variables whose values are from s
        Then generate sub-constraints similar to the above case
        If there are variables whose values are from remote sites
          Then generate queries to retrieve values from those
          remote sites first, then use those values to generate sub-
          constraints using similar method above
    End for
  End for
End for
    
```

## 3 An Example

In this example, we used the Anatomical Entity Ontology and Cancer Ontology in Protégé environment. The global constraint we used was a useful OWL-style constraint called Specific Individual Type. It requires that the explicitly declared individual of a concept or relationship (property) in the instance data must be the most specific one. This constraint can be expressed in logic programming as:

4 **Ming Fang, Weiling Li, Rajshekhar Sunderraman**

$\leftarrow inAnatomicalPart(x, y) \wedge inAnatomicalPart(x, z) \wedge cancer(x) \wedge anatomicalPart(y) \wedge anatomicalPart(z) \wedge isSubClassOf(y, z).$

Upon receiving the result of the first two queries, by running the constraint planning algorithm, we will have sub-constraint query 3 targeted at Cancer Ontology and queries 4,5,6 targeted at the Anatomical Entity Ontology.

$C_1 = \text{SELECT } ?x, ?y$ $\text{WHERE}$ $\{ ?x \text{ rdf: } inAnatomicalPart$ $?y \}$	$C_2 = \text{SELECT } ?x, ?z$ $\text{WHERE}$ $\{ ?x \text{ rdf: } inAnatomicalPart$ $?z \}$	$C_3 = \text{SELECT } ?x$ $\text{WHERE}$ $\{ ?x \text{ rdf: } hasClass$ $Cancer \}$
$C_4 = \text{SELECT } ?y$ $\text{WHERE}$ $\{ ?y \text{ rdf: } hasClass$ $AnatomicalPart \}$	$C_5 = \text{SELECT } ?z$ $\text{WHERE}$ $\{ ?z \text{ rdf: } hasClass$ $AnatomicalPart \}$	$C_6 = \text{SELECT } ?y, ?z$ $\text{WHERE}$ $\{ ?y \text{ rdf: } isSubClassOf$ $?z \}$

In this way, we avoid bringing both ontologies onsite for constraint checking.

## 4 References

- [1] M. K. Smith, C. Welty, and D. L. McGuinness. OWL Web Ontology Language Guide. World Wide Web Consortium, Recommendation REC-owl-guide-20040210, 2004. <http://www.w3.org/TR/owl-guide/>.
- [2] G. Antoniou and F. Van Harmelen. Web Ontology Language: OWL. In Handbook on Ontologies in Information Systems, pages 67–92. Springer, 2003.
- [3] G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. World Wide Web Consortium, Recommendation REC-rdf-concepts-20040210, 2004. [<http://www.w3.org/TR/rdf-concepts/>].
- [4] “The Open Biological and Biomedical Ontologies,” [<http://www.obofoundry.org>]
- [5] “Ontology Lookup Service,” [<http://www.ebi.ac.uk/ontology-lookup/ontologyList.do>]



## **Screening of natural compounds for the treatment of Diabetes Mellitus Type 2 -An *Insilico* approach**

Ajit Kumar Pandey<sup>1</sup>, Sonia Avasthi<sup>1</sup>, Pranjali Pandey<sup>1</sup>

<sup>1</sup>Amity Institute of Biotechnology, Amity University

Lucknow, Uttar Pradesh, India

### **Abstract**

In the past two decades there has been an explosive increase in the number of cases of Diabetes Mellitus worldwide, particularly type 2 diabetes (T2D). The modern lifestyles, abundant nutrient supply, reduced physical activity, and obesity increases the risk of T2D. About 60% to 90% cases of T2D are supposed to be related with obesity. Therefore the aim of present work was to examine the potential anti-diabetic agents from natural compounds using *Insilico* techniques. PPAR gamma was chosen as the target and its structure was retrieved from PDB (ID: 1I7I). From the literature survey total 120 small natural molecules having anti-diabetic potential were chosen as lead molecules. These molecules were subjected to receptor-ligand interaction using various softwares. ADME and Toxicity analysis was done using ADME TOX web. Three natural compounds namely Chlorogenic acid, Hesperidin and Lochnerine showed better ligand binding score in comparison with the reference drugs. Hesperidin which is the principal compound present in citrus fruits showed better ligand binding affinity towards the PPAR Gamma (Peroxisome Proliferator-Activated Receptor Gamma). Thus it can be concluded that these compounds can have therapeutic importance for the treatment of T2D.

### **Introduction**

Diabetes is an important cause of amputations of lower body members resulting from a non-traumatic origin, as well as blindness and kidney failure. Diabetes mellitus may present with characteristic symptoms such as thirst, polyuria, blurring of vision, and weight loss. The long-term effects of diabetes mellitus include progressive development of the specific complications of retinopathy with potential blindness, nephropathy that may lead to renal failure. There are certain natural and synthetic PPAR inhibitors called anti-diabetic agents, which will prevent or slow down the pathway and hence regulate diabetes.

### **Methodology**

The methodology included - Collection of target molecule by literature study. Retrieval of information of drugs from drug bank. Retrieval of structures for the collected molecules performing energy minimization of the retrieved small molecules using Marvin Sketch

4.1.13.Retrieval of 3D structure of the drug target using PDB. Docking of different molecules obtained in Quantum 3.3.0 .IC50 calculations of screened molecules and drugs obtained in Quantum 3.3.0. Docking of the screened molecules in Argus lab 4.0.1. Docking of screened molecules and drugs in Hex 6.1 .Viewing the Hydrogen bond length of the Receptor - Ligand complex using Swiss PDB Viewer 4.0.1.Calculating the ADMET of the screened values using ADME TOX. The best three results obtained were analysed under Quantum 3.3.0, Argus lab 4.0.1 and Hex 6.1. Their IC50 value was also analysed using Quantum 3.3.0.1. Graphs were then plotted by analysing the values.

## Results

Fig.1: Table showing Natural compounds and Quantum results

S.No	Compound Name	G-Bind energy[kJ/mol]	RMS [A]	Lipinski's rule	Selected
1	Bakuchiol	-16.71	30.94	No	No
2	Corosolic Acid	-21.82	22.74	No	No
3	Fagomine	-17.66	29.65	Yes	No
4	Marsupsin	-19.52	26.96	Yes	No
5	Pinitol	-13.74	33.16	No	No
6	Catechin	-17.19	34.85	No	No
7	Catharantine	-21.17	34.38	Yes	No
8	Chlorogenic Acid	-23.72	32.08	No	Yes
9	Hesperidin	-36.41	30.92	No	Yes
10	Lochnerine	-22.61	23.55	Yes	Yes

Fig.2: Table showing Drugs and Quantum results

Drug	G-bind[kJ/mol]	RMS
Gemfibrozil	-19.89	33.15
Pioglitazone	-23.33	25.23
Troglitazone	-22.87	27.75

Fig 3: Docking Results of natural molecules and drugs from Quantum 3.3.0

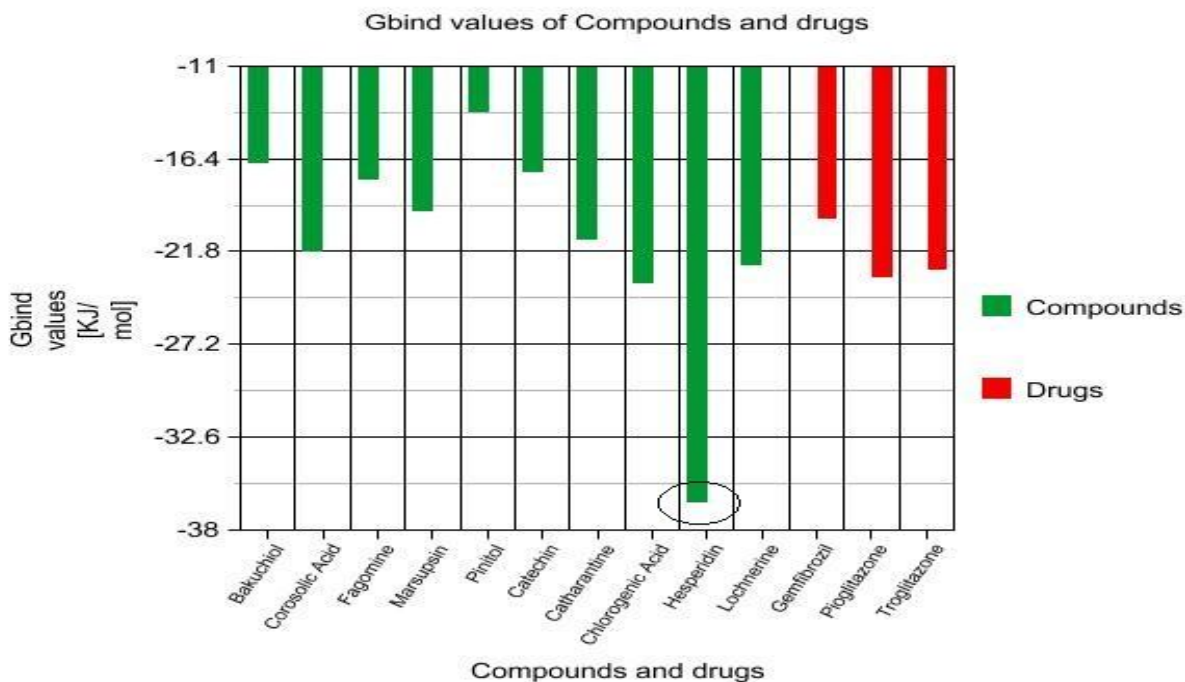
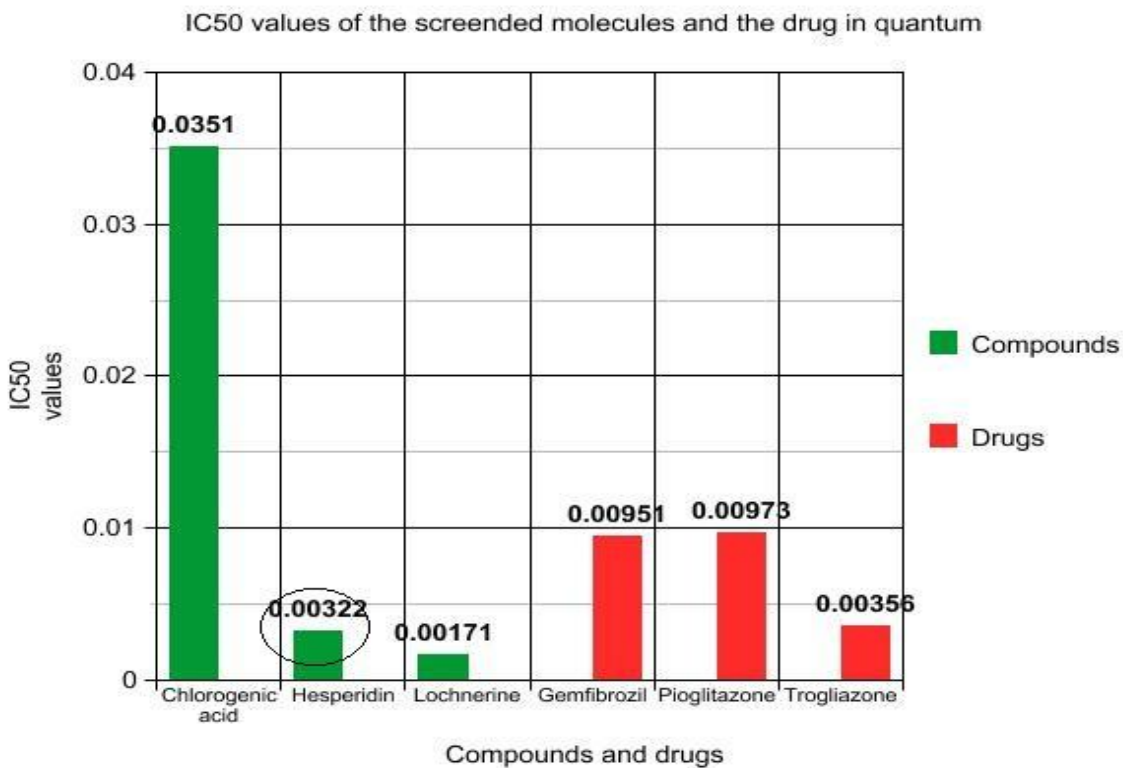


Fig 4: Docking results of screened molecules and drugs from Hex 6.1 Fig 5: IC50 calculations of the screened molecules and the drugs in Quantum 3.3.0



## SIGNIFICANCE AND CONCLUSION

The characteristic symptoms of diabetes mellitus are polydipsia, polyuria, polyphagia, retinopathy-blurring of vision, nephropathy and weight loss. The long-term effects of diabetes mellitus include progressive development of the specific complications like Retinopathy with potential blindness; nephropathy that may lead to renal failure, and neuropathy with risks of foot ulcers, amputation together with features of autonomic dysfunction, including sexual dysfunction. On top of this People with diabetes are at an increased risk of cardiovascular, peripheral vascular and cerebro-vascular diseases. **There are certain natural and synthetic PPAR inhibitors called anti-diabetic agents, which will prevent or slow down the pathway and hence regulate diabetes. It was found that molecules like Chlorogenic acid, Hesperidin and Lochnerine were showing reliable pharmacokinetics and pharamacodynamics features more than the reference drug. Further it was found that among the three molecules screened, Hesperidin was found to have the best ligand binding energy score. Thus, it can be taken as a diabetes mellitus type 2 curing drug candidate and it can be taken as an effective inhibitor of the PPAR pathway. Further *invivo* and *invitro* studies have to be conducted for concrete evidence against anti diabetic activity of these compounds and their target specificity for PPAR gamma.**

### References

- [1] Krenisky JM, Luo J Reed, MJ Carney JR Isolation and antihyperglycemic activity of bakuchiol from *Otholobium pubescens* ( Fabaceae), a Peruvian medicinal plant used for the treatment of diabetes. *Biol Pharm Bull.* 1999 Oct; 22(10):1137-40.
- [2] Sivakumar G, Vail DR, Nair V, Medina-Bolivar F, Lay JO Jr .Plant-based corosolic acid: future of anti-diabetic drug? *Biotechnol J.* 2009 Dec; 4(12):1704-11.
- [3] Yasunori Banba, Chiemi Abe, Hideo Nemoto, Atsushi Kato, Isao Adachi and Hiroki Takahata Asymmetric synthesis of fagomine and its congeners *Tetrahedron: Asymmetry*, 12 (6), 17 April 2001, 817-819
- [4] Manickam M, Ramanathan M, Jahromi MA, Chansouria JP, Ray AB Antihyperglycemic activity of phenolics from *Pterocarpus marsupium*. *J Nat Prod.* 1997 Jun; 60(6):609-10.
- [5] Davis A, Christiansen M, Horowitz JF, Klein S, Hellerstein MK, Ostlund RE Jr Effect of pinitol treatment on insulin action in subjects with insulin resistance. *Diabetes Care.* 2000 Jul; 23 (7):1000-5.
- [6] Choi JH, Chai YM, Joo GJ, Rhee IK, Lee IS, Kim KR, Choi MS, Rhee SJ. Effects of green tea catechin on polymorphonuclear leukocyte 5'-lipoxigenase activity, leukotriene B4 synthesis, and renal damage in diabetic rats. *Ann Nutr Metab.* 2004; 48(3):151-5. Epub 2004 May 6.
- [7] Un Ju Jung, Mi-Kyung Lee\*, Kyu-Shik Jeong and Myung-Sook Choi The Hypoglycemic Effects of Hesperidin and Naringin Are Partly Mediated by Hepatic Glucose-Regulating Enzymes in C57BL/KsJ-db/db Mice *The American Society for Nutritional Sciences J. Nutr.* 134:2499-2503, October 2004
- [8] Sartorelli DS et al. Different effects of coffee in the risk of type 2 diabetes according to meal consumption in a French cohort of women: The E3N/EPIC Cohort Study"