

Dock no.: 1

Submission no.:71

**COMPREHENSIVE STRUCTURAL CLASSIFICATION OF LIGAND BINDING MOTIFS IN PROTEINS**

Akira R. Kinjo and Haruki Nakamura

Institute for Protein Research, Osaka University, Japan.

Abstract (long) Comprehensive knowledge of protein-ligand interactions should provide a useful basis for annotating protein functions, studying protein evolution, engineering enzymatic activity, and designing drugs. To investigate the diversity and universality of ligand-binding sites in protein structures, we conducted the all-against-all atomic-level structural comparison of over 180,000 ligand-binding sites found in all the known structures in the Protein Data Bank by using a recently developed database search and alignment algorithm. By applying a hybrid top-down-bottom-up clustering analysis to the comparison results, we determined approximately 3000 well-defined structural motifs of ligand-binding sites. Apart from a handful of exceptions, most structural motifs were found to be confined within single families or superfamilies, and to be associated with particular ligands. Furthermore, we analyzed the components of the similarity network and enumerated more than 4000 pairs of structural motifs that were shared across different protein folds.

Dock no.: 2

Submission no.: 5

**USING PROTEIN LOCAL STRUCTURE COMPARISON APPROACH TO DISCOVER LOCAL STRUCTURE CONSERVATION FOR ENZYME FAMILY**

Yu-Feng Huang<sup>1</sup>, Chi-Jun Sheu<sup>2</sup>, Tian-Wei Hsu<sup>2</sup>, Chia-Jui Yang<sup>2</sup>, Chien-Kang Huang<sup>3</sup>

<sup>1</sup> Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 106, R.O.C.

<sup>2</sup> Department of Engineering Science and Ocean Engineering, National Taiwan University, Taipei, Taiwan, 106, R.O.C.

<sup>3</sup> Corresponding author: [ckhuang@ntu.edu.tw](mailto:ckhuang@ntu.edu.tw)

Functional site stability and flexibility has been identified in enzyme family. Therefore, structural conservation can be observed for inferring the relationship between protein function and their local structural regions. Furthermore, functional hierarchical classification of enzyme family is a good starting point to discuss relationship between protein function and local region conservation because proteins in the same enzyme family share the same biochemical function. In this work, we apply the concept of mining frequent itemset to discover local structure conservation by utilizing alignment results generated by protein structure comparison tool. In this framework, to drive protein structure comparison for data mining is time-consuming task but the experimental results still provide a clue to the relationship between local conserved regions and protein functions in selected enzyme families. In our experiments, enzyme family can be utilized to identify proteins with the same functions, and it gives us stronger evidences that discovered local structures are correlated to protein functions. In order to distinguish between structural conservation and functional conservation, we use functional site or binding site information to connect the relationship between local structure conservation and protein function.

Dock no.: 3

Submission no.: 150

### **A sequence-based dual-model predictor of protein B-factors**

Yu-Cheng Liu<sup>1</sup>, Win-Li Lin<sup>1</sup>, Shien-Ching Hwang<sup>5</sup>, Yu-Feng Huang<sup>2</sup>, Chien-Kang Huang<sup>6</sup>,

and Yen-Jen Oyang<sup>2,3,4</sup>

<sup>1</sup> Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan, ROC

<sup>2</sup> Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, ROC

<sup>3</sup> Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan, ROC

<sup>4</sup> Center for Systems Biology and Bioinformatics, National Taiwan University, Taipei, Taiwan, ROC

<sup>5</sup> Department of Computer Science and Information Engineering, Ming Chuan University, Taipei, Taiwan, R.O.C.

<sup>6</sup> Department of Engineering Science and Ocean Engineering, National Taiwan University, Taipei 106, Taiwan, R.O.C.

### **Background**

The B-factor, which is also known as temperature factor or Debby-Waller factor, is an important structural flexibility index of the ground-state protein conformation. In particular, the B-factors associated with a segment of residues reflect the local flexibility of the corresponding protein tertiary substructure. Recent studies have shown that for certain families of proteins there exists a high degree of correlation between the B-factors and the protein functional sites, including antigenic regions, enzyme active sites, and nucleotide binding sites. This article presents a sequence-based predictor of B-factors with a dual-model approach.

### **Results**

The design of the dual-model approach has been aimed at exploiting the bi-modal distribution of B-factors in order to achieve higher prediction accuracy. In this article, the prediction accuracy is measured by Pearson correlation coefficient. Experimental results show that the dual-model predictor proposed in this article is capable of delivering superior correlation coefficient in comparison with two predictors reported in two latest articles.

### **Conclusion**

Though experimental results show that the dual-model proposed in this article really works more effectively than the conventional approach, it is of interest to continue investigating more advanced designs since there exists a strong correlation between B-factors and protein functional sites. In this respect, identifying additional physiochemical properties that are related to structural flexibility deserves a high degree of attention.

Dock no.: 4

Submission no.: 100

## Homology Modeling of Homo sapiens Shwachman-Bodian-Diamond Syndrome (SBDS)

<sup>1</sup>Sunil Kumar, <sup>2</sup>Niraj Kanti Tripathy and <sup>3</sup>Babu Manjasetty.

<sup>1</sup>Institute of Life Sciences

<sup>2</sup>Berhampur University, Berhampur

<sup>3</sup>Research & Industry Incubation Center, Dayananda Sagar Research Institute, Bangalore

Structural proteomics (SP) initiatives all around the world are implementing technologies that will permit to determine three dimensional protein structures at high throughput automated fashion. The studies are usually based on X-ray crystallography, an exceptionally powerful tool to study the structures of proteins, nucleic acids, and their complexes. The main focus of these initiatives are to solve the structures of a selected set of proteins derived from as many protein families as possible. At least one structure for each family is covered with reasonable sequence identity (~25%), which act as a template structure for homology modeling and provides the accurate models for the remaining family members. This strategy is very useful to generate the models of the proteins related to human diseases which are considerably very hard to determine the structures by experimentally. Improved methodologies and bioinformatics tools are readily available to generate the accurate models. In this report, we have generated a homology model of the Homo sapiens Shwachman-Bodian-Diamond Syndrome (SBDS). The SBDS gene is expressed in all tissues and encodes a protein of 250 amino acid residues (SwissProt Q9Y3A5). The function of this protein is not known and it has no primary sequence similarity to any other protein or structural domain that would indicate a possible function. However, biochemical studies suggest that SBDS protein may be involved in RNA metabolism or ribosome assembly. Many of the disease associated mutations and truncations are also identified. The model was constructed using the X-ray structure of archaeal ortholog (AF0491 from *Archaeoglobus fulgidus*- PDB1P9Q) with the MODELLER9v2 software. The final model obtained by molecular mechanics and dynamics method and was assessed by PROCHECK and VERIFY 3D graph, which showed that the final refined model is reliable. This model may serve as a valuable reference to improve understanding of the molecular basis of disease with a detailed view of the disease related mutations within the structural domain of the protein.

Dock no.: 5

Submission no.: 26

[Homology modeling of a sensor histidine kinase from \*Aeromonas hydrophila\*](#)

Fazil MHU Turabe, Sunil Kumar and Durg V Singh

Institute of Life Sciences

*Aeromonas hydrophila* has been implicated in extra-intestinal infection and diarrhea in humans. Targeting unique effectors of bacterial pathogens was considered an impressive strategy for drug design against bacterial variations to drug resistance. Two component systems of bacteria involving sensor histidine kinase (SHK) and its response regulators were considered one of the lucrative targets of drug design. This is the first report describing a 3D structure of a SHK of *A. hydrophila*. The model was constructed through homology modeling using the X-ray structure of PleD, a response regulator in conjunction with cdiGMP (PDB code 1w25) and HemAt sensor domain (PDB code 1OR4). The homology modeling was done by using the MODELLER9v2 software. The final model obtained by molecular mechanics and dynamics method was assessed by using PROCHECK and VERIFY 3D graph, assuring that the final refined model is reliable. Until the complete biochemical and structural data of SHK is determined by experimental means, this model can serve as a valuable reference for characterizing the protein and could be explored for drug targeting by design of suitable inhibitors.

Dock no.: 6

Submission no.: 6

[An In silico modeling of rice catalase-A and docking studies with sucrose](#)

<sup>1</sup>Sunil Kumar, <sup>2</sup>Naidu SubbaRao, <sup>1</sup>Sushmita Sahu, <sup>1</sup>Mamata Ray, <sup>1</sup>Priyanka Das, <sup>1</sup>Prosenjit Mondal and <sup>1</sup>Surendra Sabat.

<sup>1</sup>Institute of Life Sciences, Bhubaneswar, India

<sup>2</sup>Jawaharlal Nehru University, New Delhi, India

The interaction between osmolytes and the enzyme protein is known to bring stabilization of the protein for its efficient function. However, the mechanism of protein stabilization through interaction of osmolytes is essentially not yet fully understood. The present report describes an in silico analysis of interaction of sucrose with rice catalase-A protein by developing a 3-D modeling of the protein and its interaction with sucrose using GOLD docking program. The analysis indicated that sucrose can ligated forming hydrogen bonding with specific amino acid residues of the protein like, R43, N304, Y347, and Q308. The interaction also includes the participation of hydrophobic amino acid residues like E44, I46, F303, L350 and M368. Our analysis holds importance in further bio-chemical understanding of catalase activity in plants exposed to stress factors; salt-stress in particular leading for over synthesis sucrose that acts as an osmoticum maintaining the cell turgor pressure and also as an osmo-protectant against stress-induced denaturation of proteins.

Dock no.: 7

Submission no.: 41

**miRNARNA: A TOOL FOR miRNA REGULATION NETWORK IN ARABIDOPSIS**

Tze-Jung Yeh<sup>1</sup>, Ko-Chun Yang<sup>2</sup>, Tsung-Jui Chen<sup>3</sup>, Sheng-An Lee<sup>3</sup>, Tse-Yi Wang<sup>3</sup>,  
Kuang-Chi Chen<sup>4</sup>, and Cheng-Yan Kao<sup>2,3</sup>

<sup>1</sup> Institute of Biotechnology, National Taiwan University, Taipei 106, Taiwan.

<sup>2</sup> Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106, Taiwan.

<sup>3</sup> Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.

<sup>4</sup> Department of Medical Informatics, Tzu Chi University, Hualien, Taiwan.

**Background:** MicroRNA (miRNA) plays an important role in post-transcriptional regulation in various species such as human, sheep, nematode, Arabidopsis, tobacco, etc. In plant, pre-miRNAs are processed into mature miRNAs (18 to 24 nt) excised by Dicer-like proteins; furthermore, mature miRNAs bind to their target mRNAs by base-pairing in cytoplasm and lead to RNA cleavage or repression of translation.

**Method:** In this work, a system named miRNARNA was proposed, with the protein-protein interactions (PPIs) dataset and the miRNA-mRNA target information in Arabidopsis provided. Since miRNA can regulate mRNA expression and subsequently affect protein expression level, the miRNARNA system attempted to construct a global network including miRNA and their targets as well as PPIs. Furthermore, the Arabidopsis miRNA data were downloaded from miRbase (<http://microrna.sanger.ac.uk/>), experimental targets were collected from literatures, and via miRU (<http://bioinfo3.noble.org/miRNA/miRU.htm>) computational targets were predicted. The PPI network platform was created based on POINeT (<http://poinet.bioinformatics.tw/>) and additional Arabidopsis PPIs were collected from TAIR (<http://www.arabidopsis.org/>) and AtPID (<http://atpid.biosino.org/>) database.

**Result:** Since the function of miRNA regulation is still unclear, miRNARNA provides a visual network of miRNAs and their influenced proteins for biologists to investigate possible mechanisms and pathways. However, miRNAs and their targets may not express in the same period or not locate in the same organelles; it is in need to add miRNA array data and mRNA expression data to improve the accuracy in network prediction for the future study.

**Availability:** miRNARNA can be accessed from <http://mirnarna.bioinformatics.tw>

Dock no.: 8

Submission no.: 131

**A METHOD TO SPECIFICALLY PREDICT HISTONE DEMETHYLASE**

Xie Chao, Pitipol Meemak, Joyce Lin, Lim Shen Jean, Tanate Panrat, Tong Joo Chuan, Martti Tammi

May 25, 2009

Histone demethylases are important epigenetic modifier proteins. One class of histone demethylase is flavin containing amine oxidoreductase. However, existing methods can not effectively distinguish this class of histone demethylase from other types of amino oxidases. Here we developed a method that can solve this problem and specifically predict histone demethylase.

Dock no.: 9

Submission no.: 51

**EpiDB: A COMPREHENSIVE DATABASE AND ANALYSIS RESOURCE FOR EPIGENETIC FACTORS**

Pei Yu Lin<sup>1</sup>, Tin Wee Tan<sup>1</sup>, Xin-Yuan Fu<sup>1</sup>, Joo Chuan Tong<sup>1,2</sup>

<sup>1</sup>Department of Biochemistry, Yong Loo School of Medicine, National University of Singapore, Singapore 117597

<sup>2</sup>Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632

In the last decade, a variety of specialized databases have been developed to facilitate the study of epigenetic factors. Some of these databases contain basic sequences and related information, some store gene expression data, while others include disease-specific records. Each of these databases has a different focus with an emphasis on different groups of epigenetic factors. The majority of these databases either lack complete information on available epigenetic factors or are deficient in bioinformatic tools for analyzing the stored sequences.

To fill this important gap in existing resources, we report EpiDB, a comprehensive repository of epigenetic factors collected from in-house laboratory, literature reports and all existing general-purpose and specialist databases. Several features distinguish this database. First, the most important characteristics of epigenetic factors were extracted, manually verified, classified and stored in the database. Each entry is annotated with the following information, where available: i) sequence name, ii) nucleotide sequence, iii) protein sequences, iv) expression data, v) protein domains data, vi) protein-protein interaction data, vii) ChIP-chip data, viii) ChIP-seq data, ix) bibliographic references, and x) public database accessions.

Second, the database is integrated with a suite of bioinformatic tools to facilitate data analysis, visualization and retrieval, including keyword and sequence similarity searches. The relationship between epigenetic factors and transcription factors are also presented in cancer-related, stem-cell related and developmental systems.

Dock no.: 10

Submission no.: 173

**INFLUENZA DATABASE AND ANALYSIS TOOLS**

Wan Hui Chua, Bevin Ng and Adrian Danker.

Temasek Engineering School, 21, Tampines Avenue 1, Singapore 529757

The influenza virus is an emergent disease that poses a serious threat to the health and safety of the global community. The virus resides primarily in birds but also infects humans and other mammals. Recent reports of probable human-to-human transmission present a distinct threat of a pandemic. Detailed understanding of the interactions between virus and host would not only help us understand the emergence of disease outbreaks, but also facilitate the design of improved diagnostics, therapeutics and vaccines to prevent and control infection. In this work, we present IDAT, a freely accessible repository containing information related to influenza viral sequences. The database is integrated with a suite of analysis tools to facilitate data analysis, visualization and retrieval. A novel epitope mapping tool allows for detailed visualization of antigenic regions in the virus that are most suitable for the design of component-based vaccines.

Dock no.: 11

Submission no.: 172

**METHPRED: A SVM-BASED SYSTEM FOR PREDICTION OF PROTEIN METHYLATION SITE**

Ashwini A/P K.Gopinathan, Rathilalli D/O S.Elamparithi, Adrian Noel Danker

Temasek Engineering School, 21, Tampines Avenue 1, Singapore 529757

Protein methylation is an important post-translational modification (PTM) essential for many biological functions, including gene regulation and signal transduction. It is now known that methylation may occur at many residues, including arginine, lysine, histidine, alanine, proline, aspartic acid, glutamic acid and glutamine. Despite its discovery half a century ago, much remains unknown about protein methylation. Experimental identification of such sites is laborious and time consuming. In this work, we report MethPred, a SVM-based system for the study of methylation on arginine and lysine, two major protein methylation sites. The system is trained using data collected from 132 proteins and tested over multiple window sizes.

Dock no.: 12

Submission no.: 76

**DENVDB: A DENGUE VIRUS PROTEIN SEQUENCE DATABASE**

Benjamin Y.L. Tan, Shweta Ramdas, Muralidharan Anantharaman, Asif M. Khan, Tin Wee Tan<sup>1</sup>, J. Thomas August<sup>2</sup>

<sup>1</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

<sup>2</sup>Department of Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, Baltimore, Maryland,

United States of America

The amount of information relating to dengue viruses (DENVs) has been increasing significantly. The need to store this large body of data has led to a competitive interest in creating a comprehensive database. Specialized DENV databases including the virus variation resource (VVR) and dengueDB were recently established, however, they offer incomplete information and limited integrated functionality. For example, the sequence records in these database are not organized according to the individual proteins of DENV, but rather as polyproteins, and do not provide advance integrated tools to mine the database, such as sequence similarity search. These result in users resorting to manual mining of the database or using external tools. We present DENVDB, a database of DENV sequences classified according to the individual proteins of each serotype, annotated with strain name, geographical and temporal information, and provides several integrated functionalities to mine the database, such as BLAST search and alignment variability analysis. Downloads of multiple sequence alignments of the proteins of each serotype, corrected for misalignments, are also provided for other desired analysis by users. The availability of alignment data over various time points provides for historical analysis of the data, such as assessment of increase in sequence diversity over time or between time points. DENVDB represents an effort to consolidate disparate DENV sequence information into an integrated resource to facilitate dengue research. It is accessible at <http://proline.bic.nus.edu.sg/denvdb>.

Dock no.: 13

Submission no.: 70

**PHD: A DATABASE OF HISTONE DEMETHYLASES AND POLYAMINE OXIDASES ENRICHED WITH STRUCTURAL AND FUNCTIONAL DOMAIN ANNOTATIONS**

Zhiwei Ang<sup>1</sup>, Chu Qin<sup>1</sup>, Shweta Ramdas<sup>1</sup>, Xuanyao Liu<sup>1</sup>, Xiaoran Chai<sup>1</sup>, Shamiah Bafadhal<sup>1</sup>, Muralidharan Anantharaman<sup>1</sup>, Benjamin Yong Liang Tan<sup>1</sup>, Asif M. Khan<sup>1</sup>, Lim Shen Jean<sup>1</sup>, Tin Wee Tan<sup>1</sup>

<sup>1</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Histone demethylation is an epigenetic mechanism implicated in many essential biological processes. The histone lysine demethylases (KDM1) containing amine oxidase domain (AOD) and histone demethylases containing Jmjc domain (JHDM) are the two main families of histone demethylases. The number of characterized and predicted histone demethylase records in public databases is increasing rapidly; however, many remain to be reviewed, leading to the likelihood of errors in some records. A search for a given protein may thus yield unreliable results, and impede research progress. Moreover, KDM1 and polyamine oxidases (PAO) both have AOD domains, creating the possibility that many PAOs predicted from the AOD domain may actually be KDM1 and vice versa. Hence, a database of histone demethylases and PAO (PHD) was created to facilitate our efforts in characterizing and understanding histone demethylases. PHD contains non-redundant curated and reviewed polypeptide sequences enriched with structural and functional domain annotations and relevant literature references, providing users with a single comprehensive and reliable resource. Further, entries are organized into clusters based on domain architecture to improve the functional prediction of putative protein entries. Users can query the PHD database by keyword, or domain architecture clusters, and view the results as elaborate summaries with graphical representations of domain architecture. Sequence searches against the database can also be performed using BLAST. To identify potential histone demethylases, PHD is equipped with a feature to scan a sequence for matches to a set of sequence motifs identified from the analysis of characterized proteins found in PHD records.

Dock no.: 14

Submission no.: 128

## **A modified greedy search method for efficient Bayesian Network Inference**

Qing Zhang<sup>1</sup> and Dianjing Guo<sup>1</sup>

<sup>1</sup>Department of Biology and the State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong SAR, China

### **Background**

Bayesian network (BN) has been successfully used to infer the gene regulatory relationships from microarray dataset because their ability to handle noisy or missing data, to handle hidden variables (such as proteins that may affect mRNA steady state level), and to describe locally interacting processes. However, one major limitation of BN approach is the computational cost because the calculation time grow more than exponentially with the dimension of the dataset.

### **Results**

In this paper, we compare the gene regulation networks inferred from Graphical Gaussian Model (GGM) and BN model and propose a modified searching method for selecting the highest Bayesian score. Particularly, our method uses GGM to limit the searching space based on the partial correlation value.

### **Conclusion**

We demonstrate that using our method can achieve ~90% accuracy and save ~50% of computational time compared to the classical greedy search.

Dock no.: 15

Submission no.: 60

**FASTR3D: A FAST AND ACCURATE SEARCH TOOL FOR SIMILAR RNA 3D STRUCTURES**

Chin-En Lai<sup>1</sup>, Ming-Yuan Tsai<sup>1</sup>, Yun-Chen Liu<sup>1</sup>, Chih-Wei Wang<sup>1</sup>, Kun-Tze Chen<sup>1</sup> and Chin Lung Lu<sup>1,2</sup>

<sup>1</sup>Institute of Bioinformatics and Systems Biology

<sup>2</sup>Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan

In recent years, there is a fast growing interest in non-coding RNAs (ncRNAs) because they play essential roles in many cellular processes. However, the function of most ncRNAs has yet to be determined. Likewise to proteins, a common and useful approach for annotating the function of an ncRNA is by searching databases for similar RNA molecules whose functions are already known. Actually, a more reliable way for determining the functions of ncRNAs is from the analysis on the structure level, since structures of molecules are typically more evolutionarily conserved than their sequences. In this study, we have developed a web server, called FASTR3D ("Fast and Accurate Search Tool for RNA 3D structures"), based on a hashing algorithm that is able to fast and accurately find structural similarities for a query of RNA molecule in the PDB database. Currently, it allows the user to input three types of queries: (1) a PDB code of an RNA tertiary structure (default), optionally with specified residue range, (2) an RNA secondary structure, optionally with primary sequence, in the dot-bracket notation and (3) an RNA primary sequence in the FASTA format. In the output page, FASTR3D will show the user-queried RNA molecule followed by a detailed list of identified structurally similar RNAs. Particularly, when queried with RNA tertiary structures, FASTR3D provides a graphical display to show the structural superposition of the query structure and each of identified structures. FASTR3D, which is now available online at <http://bioalgorithm.life.nctu.edu.tw/FASTR3D/>, can serve as a useful tool in the study of structural biology.

Dock no.: 16

Submission no.: 11

**ARCD: Appetite Regulating Compound Database**

Biplab Bhattacharjee<sup>1</sup>, Anantharamanan .R<sup>1</sup>, Sushil Kumar Middha<sup>4</sup>, Jayadeepa .R.M<sup>2</sup>, Indhuja. R<sup>2</sup>, Seema Vaidya<sup>3</sup>, Shoba.G<sup>3</sup>, Prema .G<sup>3</sup>, Sreedevi .S<sup>3</sup>, Usha T<sup>4</sup>, Anirudh Chowdhary<sup>1</sup>, Animesh Acharjee<sup>5</sup>

<sup>1</sup>Institute Of Computational Biology, Bangalore, India

<sup>2</sup>Dr. G.R .Damodaran College of Science, Coimbatore, India

<sup>3</sup>Post Graduate Department of Studies and Research in Bio-Informatics, Karnataka State Women's University, Bijapur, India

<sup>4</sup>Maharani Lakshmi Ammani College for Women, Bangalore, India

<sup>5</sup>Laboratory of Plant Breeding, Wageningen University, Netherlands

Appetite is one of the root causes which play a crucial role in obesity and its health allied diseases. Dieting alone is ineffective in treating Obesity. Obesity is one of the hazards to the human mankind which affects the health of the developed world. Even though there are various control measures for treating obesity, there is a wide expectation for using the natural molecules as the Anti-Obesity Drugs. The person termed as obese when his/her BMI (Body Mass Index) is more than 30 kgm<sup>-2</sup>. Appetite Regulating Compound Database is composed of more than 200 natural and synthetic compounds annotated from wide range of journals like Medline, PubChem, Mary Ann Liebert, Blackwell synergy, IngentaConnect, Scirus, Chemfinder, Wiley and others. Appetite Regulating Compound Database Users can perform simple and advanced searches based on Appetite suppressing or Appetite inducing molecules related to physical, chemical and Biological properties. Appetite Regulating Compound Database provides the details of hormones and the receptor involved in the Appetite Regulation and their biochemical pathways. These compounds can be visualized, downloaded, and analyzed by users who range from academicians to pharmaceutical researchers.

Dock no.: 17

Submission no.: 133

**CONSTRUCTION OF SHRIMP KNOWLEDGE BASE AND ADVANCED ANALYSIS WORKFLOW**

<sup>1</sup>Pitipol Meemak, <sup>1</sup>Amornrat Phongdara, <sup>2</sup>Martti Tammi

<sup>1</sup>Prince of Songkla University

<sup>2</sup>National University of Singapore

Several large genomes have been completely sequenced, but despite of shrimp's economic and scientific importance no large-scale sequencing project has been initiated yet. The availability of a shrimp genome will provide a basis for increased understanding of the fundamental biology of shrimp and immediate applications for selective breeding and increase the knowledge of genomes themselves. In comparison to agriculture species, such as pig, cow and chicken whose genome has been completely sequenced, relatively little has been done for aquaculture species. The knowledge of these species is also fragmented, in particular of shrimp, despite of its greatest production value in aquaculture. Moreover, almost no work has been done on the miRNA regulation on the shrimp species. In this study, we constructed the EST – miRNA analysis workflow specific for shrimps with the emphasis on *P. monodon*. The workflow is including EST based prediction of miRNA pre-cursors, prediction of miRNA targets, SNP analysis, conserved domain analysis using RPS-BLAST, protein structure homology modeling and other standard EST analysis tools.

This work will contribute to a greatly increased number of molecular markers, which are important in the genetic improvement of the shrimp species and improved lineages for increased disease resistance and growth. Importantly, a bioinformatics resource is essential to provide the necessary analysis tools, data collection and maintenance for continued successful research.

Dock no.: 18

Submission no.: 123

**EPIANN: A ARTIFICIAL NEURAL NETWORK (ANN)-BASED SYSTEM FOR PREDICTING PROMISCUOUS MHC BINDING PEPTIDES**

Khyrul Noor Redhza B Roslani<sup>1</sup>, Sze Min Koo<sup>1</sup>, Kwong Yuew Chung<sup>1</sup>, Joo Chuan Tong<sup>2,3</sup>

<sup>1</sup>Temasek Polytechnic, 21 Tampines Avenue 1, Singapore 529757;

<sup>2</sup>Department of Biochemistry, Yong Loo School of Medicine, National University of Singapore, Singapore 117597;

<sup>3</sup>Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632

The major histocompatibility complex (MHC) molecules play an important role in cell-mediated immune response. They bind short endogenous peptides derived from cytosolic proteins for presentation to T-cell receptors. T-cell recognition of MHC class I complex will trigger a cascade of immunological events that lead to the clearance of the pathogen. Peptides that can bind to multiple MHC molecules are termed promiscuous peptides. They are important targets for subunit vaccine design because they are relevant to a larger percentage of human population. Experimental identification of such peptides is time-consuming, costly and not applicable for large-scale screening. Here, we describe an artificial neural network (ANN)-based system for predicting promiscuous MHC class I binding peptides. The algorithm is generalized and applicable to cross-MHC class I ligand prediction.

Dock no.: 19

Submission no.:112

**DEMox: A DATABASE OF DEMETHYLASES AND POLYAMINE OXIDASES**

Wei Hsien Lee<sup>1</sup>, Sumitro Harjanto<sup>1</sup>, Yiping Ruan<sup>1</sup>, Junjia Timothy Huang<sup>1</sup>, Ana Lisa Gomes<sup>1</sup>,  
Xue Wei Amelia Leong<sup>1</sup>, Tingfeng Elvin Li<sup>1</sup>, Asif M. Khan<sup>1</sup>, Lim Shen Jean<sup>1</sup>, Tin Wee Tan<sup>1\*</sup>

<sup>1</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Epigenetics is an important study of heritable changes in the genome function, without a change in the DNA sequence, which controls numerous essential biological processes, such as eukaryotic gene regulation and cellular differentiation. Demethylation of DNA is one of the most significant epigenetic modifications, and histone lysine demethylase (LSD1) and Jumonji C domain containing histone demethylases (JHDM) are the two key classes of demethylases. The currently available data and information on these enzymes in public databases are poorly characterized, making it difficult to differentiate them from other proteins that share common domains, such as the polyamine oxidases (PAO) that contain C-terminal amine oxidase domain (AOD) common to LSD1. In this study, we created DeMox, a specialized eukaryotic protein sequence database of LSD1 and JHDM demethylases, as well as polyamine oxidases. A combination of BLAST and Conserved Domain Database searches, complemented by domain-based manual curations, were applied to organize the data collected from the NCBI Entrez Protein Database. A total of 237 sequences comprising of LSD1 (85), JHDM (129), and PAO (23), were finally obtained and further classified as hypothetical or experimentally validated. DeMox serves as a platform for detailed analysis of these sequences, facilitated by the various tools provided on the database, including BLAST search against the database and advanced keyword search with filtering capabilities. Download of alignment data, manually corrected for misalignments, are also provided. The database is currently at the prototype stage as we are exploring application of sequence profiles and machine learning techniques for classification of domains of demethylases, to facilitate characterization of putative or unknown demethylases. The methodology employed in this study for the development of DeMox is generic and applicable to other epigenetic modifications, for example methylation, ubiquitination, and acetylation.

Dock no.: 20

Submission no.: 136

**TITLE: CELL SEGMENTATION ALGORITHMS USING MATLAB®**

Suwei Kum

Nanyang Polytechnic

The purpose of this study was to test out the various algorithms that I had learned during my time in school and applied it towards this study. The images are H&E images on breast cancer. The aim of this study is to separate the cell from the nucleus and to let doctors have a closer look at cells that are malignant and benign. The study will show images that are segmented with the algorithms that I have learned like thresholding, Sobel Segmentation with Gradient Magnitude, Color Based Segmentation, and Active Contours without edges and a Cell Segmentation Algorithm.

Dock no.: 21

Submission no.: 91

**IMPROVED TOOLS FOR ANALYZING ORTHOLOGY AND CHROMOSOMAL LOCALISATION**

Natascha May Thevasagayam<sup>1</sup>, Rajini Sreenivasan<sup>2</sup>, Woei Chang Liew<sup>2</sup>, Sebastian Maurer-Stroh<sup>3\*</sup>, Laszlo Orban<sup>2\*</sup>

**1** Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore,

**2** Reproductive Genomics Group, Temasek Life Sciences Laboratory, Singapore,

**3** Bioinformatics Institute, (A\*STAR) Agency for Science, Technology and Research, Singapore

Although several tools to study chromosomal synteny and annotate sequences are available, these are either sophisticated for simple use by biologists or do not contain enough useful annotations to provide a biologically significant result. In this study, two new tools were developed over the Orthology Matrix (OMA) to provide these functionalities and enhance the value of its data. OMA is a comprehensive database of ortholog relationships between proteins of different species, with each protein entry annotated with valuable information. The first tool that was developed for OMA facilitated inter-species chromosome mappings to investigate synteny between chromosomes of various species. The second tool allowed chromosomal mapping of user provided sequences against protein sequences in OMA via BLAST to obtain useful annotations and chromosome locations. These tools were applied to study the potential existence of sex chromosomes in zebrafish. Chromosome mappings revealed high synteny between the sex chromosomes of five mammals with zebrafish chromosome 14 (Chr14), while the sex chromosome of medaka (LG 1) was highly syntenic with Chr1 of zebrafish. Mapping of experimentally obtained sequences of genes that were potentially sex-related against the zebrafish protein-coding loci also showed that these genes were distributed across all the zebrafish chromosomes, although chromosomes 21 and 5 had consistently higher degree of representation of these genes. Therefore, it was deduced that no single zebrafish chromosome exhibited predominant localisation of sex-related genes. This supported the experimental observations in our laboratory that indicate the absence of sex chromosomes in zebrafish. Through this application, it was demonstrated that the newly developed tools provided a simplified method of studying relationships between gene content in different genomes as well as deriving useful information from experimentally obtained data.

Dock no.: 22

Submission no.: 53

**TAP HUNTER: A SVM-BASED SYSTEM FOR PREDICTING VARIABLE LENGTH TAP-BINDING PEPTIDES**

Tze Hau Lam<sup>1</sup>, Ee Chee Ren<sup>2,3</sup> and Joo Chuan Tong<sup>1,4,\*</sup>

<sup>1</sup>Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632.

<sup>2</sup>Laboratory of Immunogenetics and Viral Host-Pathogen Genomics, Singapore Immunology Network, 8A Biomedical Grove, #03-06, Immunos, Singapore 138648.

<sup>3</sup>Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

<sup>4</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

<sup>1</sup>Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632. <sup>2</sup>Laboratory of Immunogenetics and Viral Host-Pathogen Genomics, Singapore Immunology Network, 8A Biomedical Grove, #03-06, Immunos, Singapore 138648. <sup>3</sup>Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597. <sup>4</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

The transporter associated with antigen processing (TAP) plays an important role in the human leukocyte antigen (HLA) class I antigen processing and presentation pathway. They translocate antigenic peptides generated by the proteasome from the cytosol into the lumen of the endoplasmic reticulum (ER) for loading on HLA class I molecules. The ligated HLA class I complexes then leave the ER and are transported to the cell surface for presentation to T cell receptors.

Because TAP-binding preferences may significant impact T cell epitope selection, there is great interest in applying computational techniques to systematically discover these elements. Such efforts, however, have been significantly hindered by the ability of TAP translocators to bind variable length peptides, making naïve modeling methods difficult to apply.

We describe TAP Hunter, a web-based computational system for predicting TAP-binding peptides. A novel encoding scheme, based on representations of TAP peptide fragments, allows the identification of variable-length TAP ligands using support vector machine (SVM) as the prediction engine. The system was rigorously trained and tested using 1,137 experimentally verified peptide sequences. The results showed that the system has good predictive ability with area under the receiver operating characteristics curve ( $A_{ROC}$ )  $\geq 0.91$ .

Dock no.: 23

Submission no.: 143

**DETECTION OF EPILEPSY EVENTS IN THE EEG BY MIMETIC SIGNATURES AND NONLINEAR DYNAMIC**

**INDICATIONS**

Ta-Cheng Chen<sup>1</sup>, Jing-Doo Wang<sup>2</sup>, Jiunn-I Shieh<sup>3</sup>, Kuei-Jen Lee<sup>1</sup>, Wenpin Hu<sup>1</sup>, Pei-Chun Chang<sup>1</sup> and Hsiang-Chuan Liu<sup>1</sup>

<sup>1</sup>Department of Bioinformatics,

<sup>2</sup>Asia University, Department of Computer Science and Information Engineering, Asia University

<sup>3</sup>Department of Information Science and Applications, Asia University

Epilepsy is a chronic neurological disorder that is characterized by recurrent unprovoked seizures. These seizures are due to abnormal, excessive or synchronous neuronal activity in the brain. Mimetic methods are widespread use in diagnosis to detect the sharp changes of EEG data. For a neural network such as brain, nonlinearity is necessary to describe the complexity of dynamic system. In this study, we used fuzzy measure and fuzzy integration to extract the features that combining mimetic signatures and some nonlinear dynamic indications such as Hurst exponent, sample entropy, and detrended fluctuation in time series of epilepsy electroencephalogram regarding different physiological and pathological brain states. After that, support vector machine was adopted as the classifier to discriminate the electroencephalogram regarding different brain states of epilepsy patients.

We concluded that using our algorithm we could discriminate the electroencephalogram regarding different physiological and pathological brain states of epilepsy patients.

Dock no.: 24

Submission no.: 75

**BIOINFORMATIC ANALYSIS OF THE SWINE-ORIGIN INFLUENZA A VIRUS**

Melissa Wei Shan Chan<sup>1</sup>, Shi Ya Au Yong<sup>1</sup>, Christine Lee<sup>1</sup>, Joo Chuan Tong<sup>2,3</sup>

<sup>1</sup>Nanyang Girls High School, 2 Linden Road, Singapore 288683

<sup>2</sup>Department of Biochemistry, Yong Loo School of Medicine, National University of Singapore, Singapore 117597

<sup>3</sup>Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632

An outbreak of swine-origin influenza in humans, caused by the influenza type A H1N1 virus and first detected in Mexico on March 2009, has spread to more than 33 countries and caused disease in more than 6497 patients across five continents. First recognized clinically in pigs during the Spanish flu pandemic in 1918, the influenza A lineage of swine influenza virus (SIV) are zoonotic pathogens capable of infecting a wide variety of animals. Clinical presentations of swine-origin influenza A virus (S-OIV) include fever, cough, sore throat, diarrhea and vomiting. As of 14 May 2009, the severity of this disease is such that the mortality rate in Mexico is ~2.5% (60/2446), but appears self-limiting in other countries. The origin of this new virus and its links to human, avian and swine influenza remains unknown. We have studied the evolutionary patterns of S-OIV in human, swine and avian over the past ninety years across five continents using 41,012 publicly available sequences from the NCBI Influenza Database. The effects of the hosts and geographic distance on the genetic diversity of S-OIV lineages are discussed.

Dock no.: 25

Submission no.: 122

**SPATIO-TEMPORAL ANALYSIS OF SWINE-ORIGIN INFLUENZA A VIRUS**

Yiting Zhao<sup>1</sup>, Ming Li Gan<sup>1</sup>, Kwong Yuew Chung<sup>1</sup>, Joo Chuan Tong<sup>2,3</sup>

<sup>1</sup>Temasek Polytechnic, 21 Tampines Avenue 1, Singapore 529757

<sup>2</sup>Department of Biochemistry, Yong Loo School of Medicine, National University of Singapore, Singapore 117597

<sup>3</sup>Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632

A novel swine-origin influenza A (H1N1) virus (S-OIV), first detected in Mexico on March 2009, has spread to more than 33 countries and caused disease in more than 6497 patients across five continents. To date, its links to human, avian and swine influenza remains unknown. In type A influenza virus, it has been found that genetic and geographical lineages are positively correlated, suggesting an isolation-by-distance effect. Swine influenza is common in pigs in the mid-western United States, Mexico, Canada, South America, Europe, Kenya, Mainland China, Taiwan, Japan and other parts of eastern Asia. Here, we present a novel spatio-temporal algorithm that incorporates sequence data, year of isolation, host and source country to trace the evolutionary history of this new virus. The approach is scalable and effective in discovering mutations that occur geographically over time.

Dock no.: 26

Submission no.: 113

## **CONSERVATION AND VARIABILITY OF H5N1 NEURAMINIDASE PROTEIN: IMPLICATIONS FOR DRUG**

### **DESIGN**

Chew Shi Ling<sup>1</sup>, Lee Ping Ting<sup>1</sup>, Lee Shi Qi Rachel<sup>1</sup>, Liau Yin Ting<sup>1</sup>, Yang Yating Adonsia<sup>1</sup>, Sarah Alice Davies<sup>2</sup>, Asif M. Khan<sup>1</sup>, Lim Shen Jean<sup>1</sup>, Tin Wee Tan<sup>1\*</sup>

<sup>1</sup> Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, 117597 <sup>2</sup> Department of Chemistry, University of Bath, 1 South, Bath, BA2 7AY

Avian influenza H5N1 has been a pandemic threat to the world because of its capability of avian – human transmission. The continuous mutation of viral glycoproteins impairs human response to vaccines or drugs, such as neuraminidase inhibitors (NAIs) whose effectiveness relies on the binding efficiency to the catalytic site. In this study, we focused on identifying amino acids conserved and variable across all reported avian H5N1 NA protein sequences (1072), and mapping the conserved amino acids structurally to identify potential targets for prophylactic and therapeutic purposes. Two clades were observed in the phylogenetic tree generated, indicating a high conservation of the protein. A total of 126 residues were identified as completely conserved in the NA protein (460 residues), representing ~27% of the protein length. The amino acids critical for the catalytic site pocket formation of the NA protein was found to exhibit variability: 2 out of 7 sites were not completely conserved. These diverse amino acid positions in the catalytic site could probably explain in part why existing drugs targeting the site are losing their effectiveness. This was supported by a structural and chemical bond analysis, which highlighted the importance of the diverse positions in binding to currently marketed drugs, thus, implying that variation at these sites can affect the drug efficacy. In addition, we observed that a number of non catalytic, completely conserved sites were fully or partially exposed on the virus surface, forming conspicuous structures. These regions represent attractive sites for future research work to identify alternative candidate targets sites for design of inhibitory drugs or neutralizing vaccines.

Dock no.: 27

Submission no.: 69

### **EVOLUTIONARY DIVERSITY OF HEPATITIS A VIRUS PROTEINS**

Natascha May Thevasagayam<sup>1</sup>, Maheshwar Ramakrishnan<sup>1</sup>, Gayathri Rathamani<sup>1</sup>, Vinupriya Ganapathy<sup>1</sup>, Ishak Darryl Irwan<sup>1</sup>, Rajoshi Ghosh<sup>1</sup>, Hanaa Goolbar<sup>1</sup>, Yun Le Go<sup>1</sup>, Asif M. Khan<sup>1</sup>, Tin Wee Tan<sup>1</sup>, J. Thomas August<sup>2\*</sup>

<sup>1</sup> Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore,

<sup>2</sup> Department of Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America.

The Hepatitis A virus is unique among picornaviruses as it has a relatively higher degree of conservation, and human strains seem to have been geographically associated. However, only few studies have been performed to investigate the extent of evolutionary diversity of HAV in recent years. In this study, we describe a systematic analysis of the conservation and variability of the entire HAV proteome. The study focused on the analysis of peptides of length 9 amino acids or more, for immunological applications, across 3,053 sequences of the 11 HAV proteins reported in the NCBI GenPept database (as of March, 2008). Numerous evolutionarily stable nonamer sites with entropy value of  $\leq 1$  were identified across the proteome via an entropy-based diversity analysis of the HAV sequences. Each protein had a different pattern of diversity, and interestingly, structural proteins were found to be generally less diverse than the non-structural. This is in contrast to our observation of entropy of common human pathogens, such as HIV, influenza A, dengue and West Nile virus. Maximum entropy was observed in the structural protein 1B, and non-structural proteins 3A and 3D, while the non-structural protein 3C was the most conserved. The representation (frequency) of nonamers that were variant to the predominant peptide at the stable positions across the proteome was low ( $\leq 10\%$  of the HAV sequences analyzed). Forty fragments of length 9-41 amino acids, representing  $\sim 26\%$  of the HAV polyprotein length, were identified to be evolutionarily completely conserved in all HAV sequences analyzed. The complete conservation of these immunologically relevant peptides over the history of HAV implies their possible role in diagnostics, therapeutics and peptide-specific vaccine development.

Dock no.: 28

Submission no.: 72

**MODELING HOST RESPONSE TO CHIKUNGUNYA INFECTION USING SUPPORT VECTOR MACHINES**

Jia Lin Tan<sup>1</sup>, Lisa F. P. Ng<sup>2,3</sup> and Joo Chuan Tong<sup>3,4,\*</sup>

<sup>1</sup>Department of Biological Sciences, National University of Singapore, 8 Medical Drive, Singapore 117543

<sup>2</sup>Singapore Immunology Network, 8A Biomedical Grove, #04-06, Immunos, Singapore 138648

<sup>3</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597

<sup>4</sup>Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632

Chikungunya virus (CHIKV) has re-emerged as an important infection in South-East Asia and the Pacific region, causing considerable morbidity with even some cases of fatality. Epidemic resurgence of disease was reported in the Democratic Republic of Congo in 2000, in Indonesia during 2001-03, in India during 2005-06 in Malaysia in 2006 and in Singapore in 2008. Most notably, the outbreaks in several Indian Ocean islands in 2005-2006 (Réunion Island, Maldives, Mayotte, Mauritius and Seychelles) infected more than 1.5 million people. In this era of globalization, the threat of such disease epidemics should not be underestimated as such public health events could cripple public health systems and economies. At present, there is no specific or effective treatment for CHIKV, and patient management is largely symptomatic relief and primarily anti-inflammatory drugs. Given the expanding geographic range of CHIKV and its potential to rapidly cause large scale epidemics, It has become important to understand the immune and pathogenic mechanisms active during CHIKV infections in order to guide the development of targeted and effective control and treatment strategies. More recently, global analyses on the specific involvement of inflammatory cytokines and chemokines have showed that IL-1 $\alpha$ , IL-6, and RANTES were associated with disease severity. Further identification of such factors and their related partners will be crucial the development of modulators to reduce disease severity and halt disease progression. Using STAT3 and NF- $\kappa$ B pathways as model systems, we present a novel support vector machine (SVM)-based system that allows the identification of potential biomarkers that are involved in CHIKV infections.

Dock no.: 29

Submission no.: 165

## **Introducing the *Fungome***

Ranganath, G and Sivaramaiah Nallapeta

Newgen Biotech, Hyderabad, India.

### **Background**

Fungal oncology is a recent addition to “-oncology” that deals with infectious and pathogenic fungi and their role in causing cancer in nascent stage. With genome sequencing burgeoning, it would be interesting to know if there are any fungal genes and the complementary set of those genes in other fungal species responsible for the disease. Given the number of fungal genomes being sequenced, bio information leveraged in these genomes, we extend a strategy through *fungome* to comprehend the omics-es of fungal oncogenes, specific to human cancer.

### **Methodology**

With over 25 fungal genomes sequenced and over 3 genome sequence releases, our goals through fungome is beyond clinical outreach and diagnosis. While we are inclined to understand and decipher the challenges the umpteen fungal genomes have in store for us, we use omics based approach, narrow the down the scale of searching genes to understand the role of protein-coding genes, besides exploring horizontally gene transferred (HGT) genes that remain a daunting task.

### **Conclusion**

Our work delves on the human mycoses which belong to families viz Aspergillosis, Blastomycosis, Candidiasis, Coccidioidomycosis, Cryptococcosis, Histoplasmosis, Paracoccidiomycosis, Sporotrichosis, and Zygomycosis as these could be deleterious strains/families for fungal cancer. The complete genome sequence of the human pathogenic fungi and the level of gene acquisition in the form of pathogenicity and genomic islands within the species/families are not yet studied in great detail and several questions remain unanswered. In addition, our studies on the significant role of horizontal gene transfer (HGT) and Vertical Gene Transfers (VGT) in the evolution, ecology, and virulence of cancer causing fungus has indeed proved to understand and probe the “functionally-related” genes which may be mutated during evolution or transferred between strains. Identifying the most pathogenic organisms from the fungal families and collecting the list of selected annotated proteins would be our key focus. As this work progresses, utilizing protein clusters could mean identifying similar functioning proteins that are more involved in pathogenesis.

Dock no.: 30

Submission no.: 81

**Investigation of the molecular evolution of nitrogen fixation using nucleotide triplet based condensed matrix method and probing for the unit of selection**

Saubashya Sur<sup>1</sup>, Arnab Sen<sup>1</sup>, Asim K Bothra<sup>2</sup>, Uttam K Mondal<sup>2</sup>

<sup>1</sup>NBU Bioinformatics Facility, Department of Botany, University of North Bengal, Siliguri-734013, India.

<sup>2</sup>Cheminformatics Bioinformatics Laboratory, Department of Chemistry, Raiganj University College, Raiganj-733134, India.

2

**Background:** Evolutionary history of nitrogen fixation has been retraced by analyzing *nif* H, D & K genes, and whole genomes from cyanobacteria, proteobacteria, clostridia, actinobacteria, green sulfur bacteria and archaea using an alignment free nucleotide triplet based condensed matrix method. This novel technique takes into account a set of invariants in a DNA sequence. It determines the resemblance among DNA sequences using the invariants, to infer upon characteristics of the DNA primary sequence. Qualitative and quantitative differences between DNA sequences at intraspecific and interspecific level are recognized.

**Results:** Our findings indicate that structural properties of DNA sequences are guided by different descriptors and invariants. Our results support evidence for polyphyletic origin, occurrence of horizontal gene transfer and gene duplication events. Duplications in portions of genes, operons or stretches of nucleotide sequences occurred during transformation of primitive nitrogenase to the present form, playing a crucial role in producing genes with similar properties. Our results for whole genomes differ from the whole genome and whole proteome phylogeny derived with composition vector method using CVTree. High degree of heterogeneity present among *nif* genes implied action of mutation and selection pressures with unlike intensities. Findings obtained for GC content, GC3, CAI and Nc values support the aforesaid fact. These genes being subjected to different recombination events and dissimilar selection pressures never evolved as a unit.

**Conclusions:** Completed genomes of nitrogen fixing microorganisms put forward newer outlook in studying microbial evolution, horizontal gene transfer, gene duplications etc. Our methodology using the condensed matrix method is novel and offer significant insights into the evolution of nitrogen fixation. The results indicate agreement of *nif* H, *nif* K and *nif* D phylogenies, with respect to horizontal gene transfer and gene duplication events as foremost evolutionary force in the studied gene types and whole genomes. Our work throws light on better understanding of the diversity of nitrogen fixation and the technique employs the power of categorization of DNA sequences by invariants to recognize the qualitative and quantitative properties

Genome analysis

Dock no.: 31

Submission no.: 19

**APPLICATION OF DNA BARCODE FOR THE IDENTIFICATION OF FRESHWATER FISH IN KOREA**

Sungmin Kim<sup>1</sup>, Hae-Seok Eo<sup>2</sup>, Hyeyoung Koo<sup>3</sup>, Jun-Kil Choi<sup>3</sup> and Won Kim<sup>1</sup>

<sup>1</sup>School of Biological Sciences, College of National Sciences, Seoul National University, Seoul 151-747, Korea

<sup>2</sup>School of Computational Sciences, Korea Institute for Advanced Study, Seoul, Korea

<sup>3</sup>Department of Biological Science, Sangji University, Wonju 220-702, Gangwon, Korea

DNA barcoding is a species identification technique using a short DNA sequence from a standard position in the genome. Our group has examined the utility of cytochrome oxidase I (COI) to provide taxonomic resolution in Korean freshwater fish. Here, we introduce a highly efficient and accurate molecular diagnostic web tool for the identification of freshwater fish in Korea. The web tool combines a similarity search method (BLAST) and a statistical model (profile hidden Markov model, profile HMM). We also present a freshwater fish-specific DNA chip based on the optimized set of oligonucleotide probes of the web tool. The applications of DNA barcode, Molecular Identification System of Freshwater Fish (MISF) and DNA chip, allow the species identification of freshwater fish with high efficiency and accuracy.

Dock no.: 32

Submission no.: 164

**DNA COPY-NUMBER ESTIMATION AND INTERSPECIES COMPARISON OF CHROMOSOMAL ALTERATION PATTERNS BASED ON ARRAYCGH DATA**

Shinya Akatsuka and Shinya Toyokuni

Nagoya University Graduate School of Medicine

Genomic instability such as chromosomal alterations and aneuploidy is frequently observed in cancer cells. Many studies have shown that DNA copy-number changes are deeply involved in carcinogenesis and cancer progression. Array-based comparative genomic hybridization (arrayCGH) has recently become a popular tool to identify DNA copy-number changes at high resolution throughout the genome. Similar to gene expression profiles, DNA copy-number profiles can be used as markers for diagnosis or prognosis of cancer. However, there is a problem in interpreting the data because of the principle of arrayCGH technique. Every single experiment of arrayCGH yields ratios of DNA copy-number at each chromosomal location to the mean over the whole genome in a test sample. Because the mean copy-number of test genomes is an unknown parameter, true copy-number for every location cannot be determined directly from measured fluorescent signals. Furthermore, samples of cancer tissue generally include normal cell contamination, leading to compression of the ratio values according to the contamination proportion. Here we propose a statistical method to estimate DNA copy-number per tumor cell at each genomic location by considering ploidy change in tumor genomes and proportion of tumor cells in samples. We evaluated this method by applying it to data obtained from rat carcinogenesis experiments conducted in our laboratory. In addition, we developed a procedure of data processing for comparing arrayCGH profiles among different species utilizing synteny maps. Then, we applied this procedure to comparison between data of the rat carcinomas and several types of human cancer.

Dock no.: 33

Submission no.: 99

**GENCODE ANNOTATION OF THE HUMAN GENOME: IDENTIFYING CODING AND NONCODING GENES**

Gloria Despacio-Reyes, Mark Thomas, Jeff Almeida-King, Clara Amid, If Barnes, Alex Bignell, Denise Carvalho-Silva, Adam Frankish, Toby Hunt, Mike Kay, Marie-Marthe Suner, Jonathan Mudge, Gavin Laird, Rhoda Kinsella, Laurens Wilming, Jeena Rajan, Elizabeth Hart, Ed Griffiths, James Gilbert, Stephen Trevanion, Charles Steward, Jennifer Harrow and Tim Hubbard.

The Wellcome Trust Sanger Institute

An accurate and detailed annotation system is crucial for improving our understanding of genome information, and can enrich its biological importance and meaning. Hence, as part of the scale up of the Encyclopedia of DNA Elements (ENCODE) project, the Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute is providing high quality manual annotation of the human genome. All annotation is based on transcriptional evidence (mRNA, EST or protein) according to an established set of guidelines. There is an emphasis on the annotation of alternative splice variants and pseudogenes, together with poly-adenylation features. In addition to protein coding transcripts, noncoding transcripts are also annotated, either as part of coding loci or as independent structures. The annotation of both coding and non-coding transcripts is important, given the increasing significance of functional noncoding RNAs (ncRNAs). An example of this is the HOX gene clusters, which are important protein coding genes that also have a high number of ncRNA transcripts. Annotation of noncoding transcripts for the HOXA cluster on human chromosome 7 was further supported by the findings of recent experimental studies of ncRNAs that are transcribed in coordination with the HOX genes. All of our annotation is publically available and can be viewed using the Vertebrate Genome Annotation (VEGA; <http://vega.sanger.ac.uk>) browser or other genome browsers such as Ensembl or UCSC.

Dock no.: 34

Submission no.: 88

**GENOMIC VARIATION AMONG BURKHOLDERIA PSEUDOMALLEI ISOLATES IN MALAYSIA**

Lye Siew Fen<sup>1</sup>, Abdul Munir Abdul Murad<sup>1,2</sup>, Wan Kiew Lian<sup>1,2</sup> & Sheila Nathan<sup>1,2\*</sup>

1 School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

2 Malaysia Genome Institute, UKM-MTDC Smart Technology Centre, Bangi, Selangor, Malaysia

*Burkholderia pseudomallei* is the causative agent of melioidosis, a fulminating disease of humans and animals. As an endemic disease in Southeast Asia and northern Australia, melioidosis presents diverse symptoms in infected patients. Genomic differences among the clinical strains might explain the diverse clinical manifestations which range from asymptomatic to septicemia. The potential of obtaining and eliminating important genes among the strains may lead to differences in virulence levels. Five strains of *Burkholderia pseudomallei* isolated from Malaysia has been shown to kill the nematode, *Caenorhabditis elegans* but with different lethality rates. To identify the gene variation among these five isolates, genome re-sequencing using the New Generation Sequencing technology (NGS), Illumina Solexa platform was undertaken. SSAHA2 and AMOS were utilized to map and assemble the short reads according to the reference sequence, *Burkholderia pseudomallei* K96243. Gene identities were predicted using GLIMMER3 and GENEMARK based on a reciprocal best hit against genes in the reference genome. An assembly coverage of approximately 90% compared to the reference genome was obtained. The reciprocal blast returns of 70%-75% of the genes in each of the five isolates were similar to the reference genome. Interestingly, out of the 16 genomic islands in *Burkholderia pseudomallei* K96243, all the genes in genomic islands 4, 5, 6, 11 and 12 were absent from all five local isolates. Within the rest of the genomic islands, the number of missing genes varied between isolates when compared to *Burkholderia pseudomallei* K96243. The identified gene variation among the isolates will be analyzed against the corresponding lethality rates and lead to a correlation between genomic and virulence levels.

Dock no.: 35

Submission no.: 43

***In silico* study of molecular interaction of snake venom Phospholipase A2 with inhibitors**

R. K. Sanjukta<sup>1</sup>, A. Nargotra<sup>2</sup>, B. K. Konwar<sup>1</sup> and Rajiv Das K<sup>3</sup>

<sup>1</sup>Bioinformatics Infrastructure Facility, Tezpur University, Assam

<sup>2</sup>Indian Institute of Integrative Medicine, Jammu

<sup>3</sup>Bioinformatics Infrastructure Facility, Rajiv Gandhi University, Itanagar

Phospholipase A2s (PLA2; EC 3.1.1.4) are widely distributed in snakes, lizards, bees and mammals. Snake venom PLA2s are small in size and possess diverse pharmacological and toxicological functions including neurotoxicity, myotoxicity, cardiotoxicity, anticoagulant and hemolytic activities. Therefore, they represent good model for study of protein structure and function relationship. In the present work, protein-ligand interaction of snake venom PLA2 was analyzed using computational models. Out of 116 3-D structures of snake venom PLA2 retrieved from protein data bank; 46 belonging to families elapidae and viperidae were obtained. Active sites for these PLA2 structures were analyzed using MOE. The analysis depicted the same active site pocket of 29 such structures having residues namely Ala\_102, Ala\_18, Cys\_29, Asp\_49, Phe\_5, Gly\_30, His\_48, Ile\_19, Leu\_2, Pro\_68, Trp\_31, Val\_47, Thr\_23, Tyr\_52 and Tyr\_22. Furthermore, a detailed docking simulation study of ligands into the active site of 29 PLA2s was carried out using Cerius2 in order to identify the position of the inhibitor (ligand) binding and affinity towards receptor (proteins). From the study, it was concluded that the ligands 9AR (9-hydroxy aristolochic acid), ELD [(9E)-octadec-9 enamide] and IDA [(2-carbamoylmethyl-5-propyl-octahydro-indol-7-yl) acetic acid] of snake venom PLA2 having PDB\_Id 1FV0, 1RGB and 1OXL, respectively are specific to their receptors, whereas the inhibitor of 1OXL receptor protein belonging to viperidae *Daboia russellii russellii* could be a better one for the rest of the receptors. Results of docking could be exploited in finding inhibitors for specific target proteins and thus to design potent toxin-specific antivenoms.

Dock no.: 36

Submission no.: 92

In-silico identification of giant intestinal fluke, *Fasciolopsis buski* (Lankester, 1857)

Looss, 1899 using sequence motifs as barcodes and RNA secondary structures of ITS2

rDNA

V. Tandon<sup>1</sup>, P. K. Prasad<sup>1</sup>, D. K. Biswal<sup>2</sup>, L. M. Goswami<sup>1</sup> and A. Chatterjee<sup>3</sup>

<sup>1</sup>Department of Zoology,

<sup>2</sup>Bioinformatics Centre

<sup>3</sup>Department of Biotechnology & Bioinformatics, North-Eastern Hill University, Shillong 793022, India.

**Background:** Infection of *Fasciolopsis buski*, an zoonotic intestinal fluke of pigs, is a zoonoses in South and Southeast Asia and is commonly prevalent in regions across Northeast India. Populations of the fluke collected from different parts of the region exhibit variations in gross morphology. Most phylogenetic studies using current molecular methods have focused on primary DNA sequence information. DNA sequence motifs from the internal transcribed spacer (ITS) of the nuclear rRNA repeat and also, the RNA secondary structures are particularly useful in systematics because they include characteristics that give “morphological” information, which is not found in the primary sequence. The present study was undertaken to demonstrate the RNA secondary structure from the sequence analysis of the ribosomal DNA of *Fasciolopsis buski* obtained from the intestine of freshly slaughtered pig, *Sus scrofa domestica* at local abattoirs, so as to supplement the morphological criteria and primary sequence data of the parasite.

**Results:** We describe herein the ITS2 sequence of the parasite collected from swine hosts. Phylogenetically, primary sequence analysis of *F. buski* based upon Maximum Parsimony tree resembles closely the other members of the Family Fasciolidae and Paragonimidae based upon Neighbor-Joining tree, showing significant expectation values in the alignment. However, secondary structure analysis and Bayesian analysis phylogeny showed closest resemblance with the members of both Paragonimidae and Fasciolidae, proving that they are closely related phylogenetic groups. ITS2 sequence motifs allowed an accurate in-silico distinction of the giant intestinal fluke. The data indicate that ITS2 motifs ( $\leq 50$  bp in size) can be considered promising tool for trematode species identification. Using molecular morphometrics that is based on ITS2 secondary structure homologies, phylogenetic relationships with various isolates of several trematode species have been discussed.

**Conclusion:** As has already been demonstrated for many parasitic helminthes, ITS sequences serve as effective genetic markers for molecular discrimination of species. Further, ITS2 RNA secondary structures provide a valuable tool for identifying closely related species because they contain more unique information than the usual primary sequences.

Dock no.: 37

Submission no.: 46

**COMPARATIVE IN SILICO ANALYSES OF MAP KINASE MOTIFS AND FUNCTION PREDICTION OF PLASMODIUM FALCIPARUM MAP KINASE-2 (PfMAP2)**

Farah Aida Dahalan & Hasidah Mohd Sidek

School of Biosciences & Biotechnology, Faculty of Science & Technology, Universiti Kebangsaan Malaysia

Mitogen-activated protein (MAP) kinases, a family of enzymes central to signal transduction processes, are highly conserved in eukaryotes. PfMAP2, one of two MAP kinases identified in *Plasmodium falciparum* is believed to be involved in cell cycle and growth of the parasite.

The aim of the present *in silico* investigation is to determine similarities of *Plasmodium* MAP2 sequence including that of the kinase activation domain within the *Plasmodium* species; and with the human and mouse genomes. In addition, the PfMAP2 sequence analysis is carried out to predict subcellular localization and potential substrates for the kinase in *Plasmodium falciparum*.

Homology search using BLAST showed 87-90% similarity between the PfMAP2 sequence and its orthologues in *P.berghei*, *P.chabaudi*, *P.vivax* and *P.yoelli*. Sequence comparison of *Plasmodium* MAP2 activation domain with the human and mouse genomes using ClustalW indicated <36% homology. The *Plasmodium* MAP2 activation motif is TSH. Based on Signal Peptide and SubLoc analysis, PfMAP2 is predicted to be a non-secretory protein localised in the cytoplasm and the mitochondria. Using Predikin, substrates identified with substrate-determining residue (SDR) matrix scores of >85% for PfMAP2 consists of SMAD4 (94.01%), Elk-1 (89.91%), RSK3 (88.91%), RSK1 (88.91%), NFAC2 (88.43%) and ATF2 (86.09%).

The data obtained from this study validates what is already known about PfMAP2; regarding its potential as an attractive antimalarial drug target. The TSH activation motif in *Plasmodium* is different from that in its eukaryotic human and rodent hosts (TEY). In addition, low similarity of MAP2 sequence exists between the parasite and its mammalian hosts. Subcellular localisation of the kinase in the cytosol and the mitochondria as well as the identification of several potential substrates for the PfMAP2 will allow further investigation on the functional aspects of this kinase in malarial infection.

Dock no.: 38

Submission no.: 57

**HIGH-THROUGHPUT GENE EXPRESSION ANALYSIS REVEALS BURKHOLDERIA PSEUDOMALLEI SURVIVAL STRATEGIES IN MACROPHAGE CELLS**

SYLVIA CHIENG<sup>1</sup>, LAURA CARRETO<sup>2</sup> & SHEILA NATHAN<sup>1</sup>

1 School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

2 CESAM & Departamento de Biologia, Universidade de Aveiro, 3810-193 Aveiro, Portugal

**BACKGROUND:** *Burkholderia pseudomallei*, the causative agent of melioidosis, an endemic disease to Southeast Asia and Northern Australia, is able to survive and multiply within phagocytic and non-phagocytic cells. The interaction between the bacterium and host macrophage cells is crucial in understanding the strategies used in bacterial survival and progression of the disease. **RESULTS:** Here, we report the application of high density microarrays to unravel the pathogen's expression profile over a 6 hour time course of infection of human monocyte-like U937 cells, as compared to in vitro grown bacteria. High-throughput single channel microarray data was filtered and median normalized before determination of differentially expressed genes based on Significance Analysis of Microarrays (SAM). Using multiclass SAM and hierarchical clustering, 4086 out of the 5728 predicted genes of *B. pseudomallei* were found to be differentially regulated in at least 1 time point. About 14.8% (n=605) of the total differentially expressed genes were regulated throughout the monitored infection period. Most of these genes were downregulated and were involved in metabolism, cell envelope and motility, replication, transport and regulatory functions. Similar regulatory patterns have been observed in *Salmonella* sp., *Shigella* sp. and *Escherichia coli*, indicating a common theme in bacterial infection. Survival of *B. pseudomallei* within macrophage cells also involves up-regulation of tss-5 type VI secretion system proposes a novel virulence mechanism. Some of the consistently up-regulated genes were involved in anaerobic metabolism and ABC transport systems.

**CONCLUSION:** Strategies of adaptation to hostile environments, avoidance of immune responses and regulation of novel virulent determinants are important for *B. pseudomallei* survival in macrophage cells.

Dock no.: 39

Submission no.: 4

## **Insights into rubber biosynthesis in *Parthenium argentatum* and**

### ***Hevea brasiliensis***

Jayaraman Muthukumar<sup>1</sup>, Anmol J. Hemrom<sup>1</sup>, Nagarajan Arumugam<sup>1</sup>, Mannu Jayakanthan<sup>1</sup> and Durai Sundar<sup>2\*</sup>

<sup>1</sup>Centre for Bioinformatics, School of Life sciences, Pondicherry University, Pondicherry 605014, India

<sup>2</sup> Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology (IIT), New Delhi 110016, India

#### Background:

Natural rubber is an important polymer produced by plants and made up of isoprene units derived from isopentenyl diphosphate (IPP). Although more than 2000 plant species are known to produce natural rubber, currently there are two important commercial sources, *Hevea brasiliensis* (the Brazilian rubber tree) and *Parthenium argentatum* Gray (guayule). Natural rubber biosynthesis requires three distinct biochemical processes such as (i) initiation, (ii) elongation and (iii) termination. A comparative analysis of the enzymes farnesyl diphosphate (FPP) synthase in *Parthenium argentatum* and cis-prenyl transferase (CPT) in *Hevea brasiliensis* that play a vital role in initiation and elongation stages for biosynthesis of cis-1, 4-polyisoprene has been undertaken in this study.

#### Results:

The comparative sequence analysis of FPP synthase and CPT to their identified similar sequences, was carried out to understand the evolutionary relationship among different species. Homology modeling and binding pocket analysis was performed to understand the structure-function relationship of FPP synthase and CPT. The structural templates farnesyl diphosphate synthase (Source: *Gallus gallus*) [PDB ID: 1UBX] for FPP synthase and undecaprenyl diphosphate synthase (Source: *Micrococcus Luteus* B-P 26) [PDB ID: 1F75] for CPT were selected for homology modeling. The Ramachandran plots were developed for modeled structures of FPP synthase and CPT, which showed 95.9% and 92.6% of amino acid residues occurring in favored regions. These models were deposited into Protein Model Database [PMDB ID: PM0075218 & PM0075509]. The substrate and cofactor binding site residues R103, L149, A184, Y197, L211, H214, E223, T226, D332, K246, Y306, K313 of FPP synthase and Y4, E7, R20, K21, G22, K154, K178, D193, E231, T232, R233 of CPT were identified by using binding pocket analysis.

#### Conclusion:

The computational analysis of the enzymes involved in initiation and elongation of cis-1, 4-polyisoprene in rubber biosynthesis provided invaluable insights into the putative initiation and elongation factors for FPP synthase and CPT.

Dock no.: 40

Submission no.: 56

**BIOINFORMATICS ANALYSIS OF HLA CLASS II ASSOCIATIONS WITH METHICILLIN-RESISTANT STAPHYLOCOCCUS AUREUS (MRSA)**

Anni Cai<sup>1</sup>, Aofei Lu<sup>1</sup>, Carmen Hoo<sup>1</sup>, Joo Chuan Tong<sup>2,3</sup>

<sup>1</sup>Raffles Girls School, 20 Anderson Road, Singapore 259978

<sup>2</sup>Department of Biochemistry, Yong Loo School of Medicine, National University of Singapore, Singapore 117597

<sup>3</sup>Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632

Methicillin-resistant *Staphylococcus aureus* (MRSA) is a bacterium responsible for difficult-to-treat infections in humans, and is particularly hazardous to hospital patients since having evolved a resistance to common antibiotics. In 2005, 19,000 people died from MRSA infections in the United States, and an average of 6.3 out of every 100,000 infections resulted in death.

The human leukocyte antigen (HLA) class II molecules play a central role in cell-mediated immune response. They present exogenous antigens, such as those from bacteria to CD4<sup>+</sup> helper T-cells. Upon recognition, the binding triggers an immune cascade which leads to the clearance of the foreign substance. To date, much remains unknown with regards to the HLA class II restriction patterns in MRSA.

To assess the impact of HLA class II variation and MRSA infection, computational predictions of T-cell epitopes that bind to 11 common HLA class II alleles (DRB1\*0101, \*0301, \*0401, \*0404, \*0405, \*0701, \*0802, \*1101, \*1302, \*1501, and DRB5\*0101) were performed using three computational systems derived from the Immune Epitope Database and Analysis Resource (IEDB): Average Relative Binding (ARB) matrix, Stabilized Matrix Method (SMM) and profile-based method. These data are also compared to mutation patterns in MRSA family of bacteria to provide a description of the residues crucial for binding and specificity.

Dock no.: 41

Submission no.: 21

## **Development of T-cell Epitope Predictors**

**Wu LanLan<sup>1</sup>, Tan Eng Chian<sup>1</sup>, Aaron Lim Shi Fan<sup>1</sup>, Chung Kwong Yuew<sup>1</sup>, Tong Joo Chuan Victor<sup>2</sup>**

<sup>1</sup> Temasek Polytechnic, 21 Tampines Avenue 1, Singapore 529757

<sup>2</sup> Institute for Infocomm Research, 1 Fusionopolis Way #21-01 Connexis, Singapore 138632

This poster describes the development of web-based T-cell epitope predictions through the use of Support Vector Machine (SVM) learning algorithms and through the recognition of binding sites of an unknown peptide to a specific HLA (Human Leukocyte Antigen). These recognised T-cell epitopes are targets for vaccine and immunotherapy development. The prediction system is able to be deployed for large-scale virtual screening initiatives of unknown peptides.

Dock no.: 42

Submission no.: 62

**COMBINING CLUSTER ANALYSIS OF DYNAMICS OF GENE EXPRESSION PROFILES**

Ying-Kai Jhong<sup>1</sup>, Tse-Yi Wang<sup>2</sup> and Kuang-Chi Chen<sup>1\*</sup>

1 Department of Medical Informatics, Tzu Chi University, Hualien, Taiwan

2 Bioinformatics Lab, Dept. of CSIE, NTU, Taipei, Taiwan

**Background:**

Time-series microarray data are produced under different biological conditions, which allow investigators to study the behaviors of genes over time and across conditions. Although various clustering methods have been proposed to group longitudinal array data, the best method often varies from one target data to another. Combination of multiple methods has been shown to be effective.

**Results and Discussion:**

This article proposes a two-stage procedure to analyze temporal microarray profiles. At the first stage, several model-based clustering methods are applied to group the gene-expression profiles. Next, support vector machine (SVM) method is carried out to combine the results of previous step. The combination of heterogeneous and well-performing methods may achieve the best performance. We apply this procedure to simulated and real biological data. Our algorithm could identify the correct number of clusters efficiently, and detect the relevant functional categories of temporal profiles.

**Conclusion:**

This procedure combines model-based statistical approaches with machine learning technique. It is flexible to implement different clustering methods for different purposes and application domains. The use of fusion results of multiple methods provides a better performance than a single method.

**Method:**

In order to depict temporal gene expression profiles, the Bayesian model-based clustering and hidden Markov model (HMM) are applied. The Bayesian method uses polynomial functions to describe the changes of expression profiles. The HMM characterizes the cyclic profile and the horizontal dependencies of time-course data. In the fusion step, SVM is implemented to combine all different clustering results by adopting clustering results as its features.

Dock no.: 43

Submission no.: 28

## **Computation of Hydrophobic Nature of Proteins Based on Carbon Content**

M.Vijayasathy, V.Jayaraj# and E.Rajasekaran

Departments of Biotechnology and Computer Applications Periyar Maniammai University Vallam, Thanjavur – 613403  
Tamil Nadu, India.

here are more than 55 computational methods available on the Internet (<http://www.expasy.ch/cgi-bin/protscale.pl>) for computing protein properties from sequence information. About 22 such programs are focused on hydrophobicity of proteins. The factor which influences hydrophobicity is that the amount of carbon. This carbon content in proteins has been studied extensively by our group in recent years. It is reported that proteins prefer to have 31.44% carbon for its survival. The distribution of this carbon along the protein sequences are also studied in detail. The studies again conclude that proteins prefer to have 31.44% carbon all along the sequence both locally and globally. But the fact is that proteins accumulate more carbon in the active site for its activity. Due to this reason the carbon distribution is altered accordingly. We have developed a software tool to study these properties and brought to Internet for a wider usage. The program can be accessed at [www.rajasekaran.net.in/tools/carbana.html](http://www.rajasekaran.net.in/tools/carbana.html). The program accepts protein sequence in fasta format or flat file and outputs a carbon content with atom number. The output can be plotted to look for active site, hydrophobic, hydrophilic region along the sequence. The outputs are similar to hydropathy plots but superior over those methods.

Dock no.: 44

Submission no.: 68

**IN SILICO CHARACTERIZATION OF ARGININE AND LYSINE METHYLATION SITES**

Wei Hsien Lee<sup>1</sup>, Tin Wee Tan<sup>1</sup>, Joo Chuan Tong<sup>1,2</sup>

<sup>1</sup>Department of Biochemistry, Yong Loo School of Medicine, National University of Singapore, Singapore 117597

<sup>2</sup>Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632

Protein methylation is an important and reversible post-translational modification of proteins that governs cellular dynamics and plasticity through many protein pathways and interactions. Such methylation events occur ubiquitously in nature, from eukaryotes to prokaryotes and are commonly mediated by a protein family of methyltransferases that require S-adenosylmethionine as a donor substrate for methyl group transfer. However, despite having been first discovered half a century ago, much remains unknown with regards to the regulatory roles of protein methylation.

To identify amino acid residues and essential physicochemical properties that are characteristics of protein methylation, an analysis was performed on a dataset of 330 methylated sites, comprising of 94 lysine and 236 arginine methylated sites from 145 eukaryotic and bacterial protein sequences. Several patterns of expression occur with a higher frequency in lysine and arginine methylated sites. In addition, the difference in the physicochemical properties between lysine methylated sites, arginine methylated sites and non-methylated sites were also examined. The results of these analyses and possible functional implications will be discussed.

Dock no.: 45

Submission no.: 58

**The microRNAs complementary to mRNAs, target binding sites in the STS of some hypothetical proteins**

Chirag Matkar<sup>1\*</sup>, Saraswathy Nagendran,<sup>2</sup> and Arun Gupta<sup>3#</sup>

1. D Y Patil Biotechnology & Bioinformatics Institute, D Y Patil University, Pune, India.

2. School of Pharmacy & Technology Management, Narsee Monjee Institute of Management Studies. Mumbai, India.

3. School of Computer Sciences, Devi Ahilya University, Indore (MP), INDIA #Present address: EMBL-EBI, Hinxton, Cambridge CB10 1SD, United Kingdom.

The microRNAs or miRNA genes encode short (18-24 nucleotides) non-protein coding RNAs, that regulate gene expression in eukaryotes through post-transcriptional suppression of mRNA translation. We considered Sequence Tagged Sites (STS) of some of the potential hypothetical proteins which are known to be involved in diseases and hope our annotation will facilitate the study of their contribution to biological processes and disease.

Dock no.: 46

Submission no.: 3

**INSILICO COMPARISON OF PROTEOMES FROM NITROGEN FIXING ORGANISMS: INSIGHTS INTO VARIATIONS AND AMINO ACID COMPOSITIONS**

Arnab Sen<sup>1</sup>, Saubashya Sur<sup>1</sup>, Subarna Thakur<sup>1</sup>, Asim Bothra<sup>2</sup>, Louis Tisa<sup>3</sup>

<sup>1</sup> Bioinformatics Facility, University of North Bengal, Siliguri 734013, India

<sup>2</sup> Raiganj College Raiganj, India;

<sup>3</sup> Department of Microbiology, University of New Hampshire, Durham, NH, USA

The physical properties of the proteins are quite significant in portraying the functions of microbes. Analysis of theoretical proteomes from micro organisms offer an opportunity to understand the molecular nature of functionality of the proteins associated with metabolic functions. Over the last few decades a number of works related to genetics, physiology, molecular biology and biochemistry of nitrogen fixing microorganisms have been performed, however the availability of proteomes sequences for these bacteria have opened up the possibility to undertake comparative proteome analysis to garner novel insights. Characterization of the proteomes at the amino acid composition level is expected to solve many unanswered questions relevant to the way of existence of the nitrogen fixers. In this work, some proteomes of nitrogen fixing microorganisms were amino-acid adaptation index, amino acid frequency, isoelectric point, protein energetic cost, aromaticity and hydrophobicity with special reference to the proteins associated with the nitrogen fixation mechanism. Isoelectric point, amino acid frequencies and energetic costs of the proteins were correlated with GC3 and GC content of the organisms. Our results provide a basis for resolution of theoretical proteome comparison in nitrogen fixers. Our findings report significant insights into the variations of the proteomes as well differences in amino acid compositions. The patterns reflected in the results clearly depict the interrelationships between, amino acid frequency, isoelectric point, protein energetic cost, aromaticity and hydrophobicity. Proteomes are influenced by compositional bias. Our analysis is expected to provide newer dimensions in the biology of nitrogen fixers.

Sequence analysis

Dock no.: 47

Submission no.: 96

**Identification of susceptible SNPs for the development of head and neck cancer in Indian population**

Pinaki Mondal<sup>1</sup>, Indian Genome Variation Consortium<sup>2</sup> and Susanta Roychoudhury<sup>1</sup>

<sup>1</sup>Molecular and Human Genetics Division, Indian Institute of Chemical Biology, Kolkata, India,

<sup>2</sup>Nodal Laboratory, Institute of Genomics and Integrative Biology, New Delhi, India

Genetic polymorphisms in genes controlling cellular processes such as cell cycle, DNA repair and apoptosis may modulate the risk for the development of cancer. Availability of large number of SNPs in the human genome allows us to find putative risk alleles. Head and Neck Squamous Cell Carcinoma (HNSCC), estimated worldwide to be the sixth most common cancer, is one of the most prevalent cancers in India. We intend to identify putative SNPs in the DNA repair genes that may confer risk to the development of HNSCC in Indian population. With this goal, we used HAPMART to download necessary of all SNPs covering eleven important DNA repair genes. An in house program MARKER was used to screen all the SNPs in these genomic regions (how many Mb) based on several parameters such as physical distance between each SNP, average Hardy-Weinberg P values, average heterozygosity, average  $r^2$  value, average minor allele frequency, etc in four ancestral HAPMAP populations to identify SNPs that we can use in the case-control study in Indian population. Common probable LD blocks structure of HAPMAP populations was designed using a program Haploview. Different bioinformatics prediction servers used to elucidate possible functionally important SNPs. Genotype data of different selected SNPs in seven of eleven above-mentioned genes and some other cancer related genes in 552 normal Indian samples from different ethnic groups were analyzed to find out effectiveness of our SNPs selection method. Several tools (Genecount, Dispan, etc.) were used to analyze frequency, distribution of SNPs and possible population clusters in Indian population.

Dock no.: 48

Submission no.: 40

## **Gold Standard for Single Nucleotide Polymorphism (SNP) Nomenclature**

Danny Chiang Choon Poo<sup>1</sup>, Yi Situ<sup>2</sup> and James Tzia Liang Mah<sup>3</sup>

1 Department of Information Systems, National University of Singapore

2 Department of Computer Science, National University of Singapore

3 Data Mining Department, Institute for Infocomm Research (I2R), A\*STAR

### **Motivation**

A lack of standardization in single nucleotide polymorphism (SNP) nomenclature causes much regrettable confusion in the community. The number of SNPs in public databases has been growing exponentially over the years and it has become essential for an SNP gold standard for nomenclature. While there have been several attempts to uniquely identify SNPs in the past none has been outstandingly successful. This paper discusses the conflicting nomenclatures that are commonly used and introduces a gold standard for SNP nomenclature that greatly minimizes ambiguities and miscommunications among researchers.

### **Results**

The proposed gold standard for SNP nomenclature has the following characteristics: precise, unambiguous, and unique. The location of the SNP can be precisely located within the chromosome number and position. The definition of the SNP nomenclature is unambiguous; each SNP has a definitive meaning, misinterpretation is therefore not possible. The identified SNP is unique since there is only one identifier for each SNP. The structure of an SNP identifier consists of a tuple of three values: chromosome number, chromosome position, and observed polymorphism. For example, "21:31760592:C>T" denotes an SNP at nucleotide 31760592 of chromosome 21 of the reference sequence, with C changed to a T. Chromosome number and position gives the precise location of the SNP, and the polymorphism described adopts the substitution, deletion and insertion nomenclature for single nucleotide recommended by den Dunnen and Antonarakis (2000).

Dock no.: 49

Submission no.: 95

**Molecular characterization of the intestinal fluke, *Artyfechinostomum sufrartyfex***

**(Trematoda: Echinostomatidae) using a combinatorial approach of molecular**

**morphometrics and CBC analysis in the internal transcribed spacer regions of rDNA**

Lalit Mohan Goswami<sup>1</sup>, Veena Tandon<sup>1\*</sup>, Devendra Kumar Biswal<sup>2</sup>, Pramod Kumar Prasad<sup>1</sup> and Anupam Chatterjee<sup>3</sup>

<sup>1</sup> Department of Zoology,

<sup>2</sup> Bioinformatics Centre,

<sup>3</sup> Department of Biotechnology and Bioinformatics, North Eastern Hill University, Shillong, Meghalaya, 793022, India

*Artyfechinostomum sufrartyfex* Lane, 1915 is an echinostome intestinal fluke of pigs causing echinostomiasis, a zoonosis. *A. sufrartyfex* infection in humans has been sporadically reported to occur in few Southeast Asian countries including India. However, another echinostome species viz., *A. oraoni* has been found to be of endemic occurrence among the Oraon tribe of West Bengal in India. Discrete identification of the species implicated in infection becomes difficult if only based on morphological criteria of adult or operculate egg stages of the fluke. With the use of molecular tools assisting the conventional diagnostic techniques and various tree construction methods, molecular morphometrics approach with compensatory base change analysis that serve as reservoirs for evolutionary changes, phylogeny of *A. sufrartyfex* is discussed.

**Results**

The ITS regions of *A. sufrartyfex* DNA were successfully amplified, by using the universal primers. The obtained sequences (Accession = EF027100, EF27101) were compared with other trematodes obtained from the Genbank database. We determined the taxonomic position of *A. sufrartyfex* using a combinatorial approach of Molecular morphometrics and compensatory base change count analysis. The secondary structures generated were aligned and CBC distance count matrix generated showed few CBCs as expected. Moreover, ITS motifs were searched for further validation and the sequence motif patterns had exact or perfect matches with species from Echinostome family with significant E-value and high percent identity scores. The Neighbour-Joining (NJ) and Maximum Parsimony (MP) trees also complied with the above mentioned in-silico analysis.

**Conclusion**

This study has provided the first molecular characterization of this species using internal transcribed spacer regions as carriers of evolutionary information. The various in silico study of ITS sequence accessions of *A. sufrartyfex* and its homologues showed close similarity with members of Echinostomatidae and *Fasciolopsis buski* as well.

Dock no.: 50

Submission no.: 89

**ANALYSIS OF THE MOLECULAR MECHANISMS OF KNOWN AND PREDICTED DISEASE MUTATIONS IN LGI EPILEPSY GENES**

Vachiranee Limviphuvadh, Ling Ling Chua, Rabi'Atul Adawiyah Bte Rahim, Frank Eisenhaber, Sharmila Adhikari and Sebastian Maurer-Stroh

Bioinformatics Institute, A\*STAR Singapore

Partial epilepsy with pericentral spikes (PEPS) is a subtype of familial temporal lobe epilepsy with the disease locus mapped to 4p15 but to date, a causative gene of the disease is yet to be identified. Through protein sequence analysis of all 53 genes that are mapped in proximity to the disease locus, we identified leucine-rich repeat LGI family, member 2 (LGI2) as the most likely candidate gene for the disease. LGI2 is a member of the LGI gene family and contains characteristic EPTP/EAR epilepsy-associated repeats. Mutations in the best studied member, LGI1/Epitempin, cause a different type of epilepsy called autosomal dominant lateral temporal epilepsy (ADLTE). Thorough literature analysis identifies tonic-clonic seizures as possibly shared phenotype between PEPS and ADLTE. Structural modeling of the EAR domain region and mapping of evolutionary conserved residues to the surface of the models, indicates a conserved preferential binding side and the possibility of shared interaction partners among the LGI family. Our localization experiments show that all LGI1 disease mutants fail to be secreted, accumulate in the ER and do not reach the Golgi. In contrast, the effects of the tested SNPs in LGI2 on secretion are minimal. However, when introducing the same mutation that caused secretion deficiency in LGI1 to its structurally homologous position in LGI2 we also observed lack of secretion of the LGI2 mutant. Inversely, a mutation that did not alter secretion in LGI2 also did not affect LGI1. This similarity of the effects of mutations suggests a common disease mechanism

Dock no.: 51

Submission no.: 94

## **SeqAnnotator a DNA sequence annotation system for genomic laboratories**

Pubudu S. Samarakoon, Rohan W. Jayasekara and Vajira H.W. Dissanayake

Human Genetics Unit, Faculty of Medicine, University of Colombo

In the last two decades the rapid growth of high throughput genotyping and sequencing generated large volumes of genomic data at an unprecedented speed. With the sheer volume of data and the increasing demand for the understanding of genes and proteins, bottleneck of genomic laboratories is shifting towards a new position, which is the issue of sequence annotation. SeqAnnotator is a software solution which was conceived with the aim of automating the annotation process of genomic and cDNA sequences in laboratories. Currently SeqAnnotator takes the DNA sequence of a gene as an input and produces an annotated sequence as a result. SNP ID's given by the National Center of Biotechnology Information (NCBI) dbSNP database, exons and introns, Expressed Sequence Tags (ESTs) and Sequence Tags Sites (STS) can be accurately assigned within the input sequence and presented in a graphical and textual format with the use of this software. Users are given the option to feed a file downloaded from the NCBI dbSNP database with the list of SNPs into the program. This will be used to assign SNPs in the input sequence. Information of each annotated feature of this program is extracted from the NCBI database collection. Perl programming language is used to write this program and Entrez EUtils and perl XML parsers are used to retrieve relevant information from the NCBI server. SeqAnnotator is been used in the SNP profiling for the Sri Lankan Genome Variation Database (SLGVD).

Dock no.: 52

Submission no.: 132

**ERROR CORRECTION FOR SECOND-GENERATION SEQUENCING TECHNOLOGIES**

Tanate Panrat<sup>1</sup>, Amornrat Phongdara<sup>1</sup>, and Marri T Tammi<sup>2,3</sup>

1 The Center for Genomics and Bioinformatics Research, Faculty of Science, Prince of Songkla University, Hatyai, Thailand.

2 Dept. of Biological Sciences and Dept. of Biochemistry, National University of Singapore, Singapore.

3 Karolinska Institutet, Dept. of Microbiology, Tumor and Cell Biology, Stockholm, Sweden.

Rapid advances in sequencing technologies have resulted in massive increase in speed and greatly reduced the sequencing cost. Unfortunately, the short read length makes *de novo* genome assembly task cumbersome even for small genome sizes and perhaps impossible for large mammalian genomes. This is due to the combination of repeated sequences and sequencing errors. Using a combination of Sanger sequencing, which produces long reads and a large number of short reads produced by second-generation technologies together with varied mate-pair lengths, greatly increases the quality of an assembly. However, the repeated sequences still cause assembly errors. Since most of the repeats are not identical, but contain a few differences, therefore correcting sequencing errors further improves the repeat resolution. Here we present a novel approach for repeat resolution based on machine-learning method. Our method does not attempt to correct sequencing errors, but instead directly classifies sequenced reads in to repeat classes. The repeat classes represent separate repeated genomic sequence locations.

Dock no.: 53

Submission no.: 171

**STATISTICAL SIGNIFICANCE OF PROTEIN INTERACTION AND METABOLIC DATA FOR SUBCELLULAR LOCALIZATION PREDICTION**

Gaurav Kumar, Helena Nevalainen and Shoba Ranganathan

Macquarie University

One of the important task of functional proteomics is to understand the subcellular localization (SCL) of eukaryotic proteins for the better understanding of their function. In the past, researchers have experimentally investigated and predicted the SCL of proteins in both eukaryotes and prokaryotes. Such experimental and predicted data has been archived in various protein databases. In spite of the increasing data, prediction of SCL remains a crucial problem. The indirect prediction by counting the neighbouring protein function is very useful for annotating function to the unknown protein with known functional neighbours. Similarly, it can be extended to understand the location of protein via its interacting neighbours. For example, a given protein is likely to be co-located with other proteins it interacts with or shares a common substrate with. Therefore, an integrated dataset in XML format is created by linking protein SCL information to interaction and metabolic pathway information for human.

The LOCATE database contains 64,637 human proteins with known or predicted SCL information [1]. The number of LOCATE proteins found in protein interaction and metabolic pathway databases vary considerably. This is partly due to the difference in their curation and retrieval system for archiving molecular information. Therefore, we incorporated five protein-protein interaction (HPRD, IntAct, MINT, DIP and BioGRID) and two metabolic pathway (KEGG and HUMANCYC) databases to increase the protein association coverage. A  $\chi^2$  statistic and loglinear analysis was performed to test the hypothesis that the interacting pairs of proteins reside in the same sub-cellular compartment.

The statistical tests on the integrated dataset support our hypothesis. It also provides an important biological insight regarding the role of membrane in various subcellular compartments for controlling the active and passive transportation.

Dock no.: 54

Submission no.: 45

## **COMBINING AGENT-BASED MODELS WITH STOCHASTIC SYNERGISTIC SYSTEMS FOR GENE REGULATORY NETWORKS**

Tse-Yi Wang<sup>1</sup>, Cheng-Yan Kao<sup>1</sup> and Kuang-Chi Chen<sup>2,\*</sup>

1 Bioinformatics Laboratory, Dept. of CSIE, NTU, Taipei, Taiwan

2 Department of MI, Tzu Chi University, Hualien, Taiwan

\* .Corresponding author

### Background

One of the most fundamental processes in a living cell is the regulation of gene expression. Its dynamic modeling or simulation in systems biology is the major theme increasingly attracting in post-genome era. We are motivated to combine agent-based models together with stochastic synergistic systems to characterize the gene regulatory networks both from the micro-level behaviors of individual molecules and the macro-level variations of integrated measures.

### Methods and Results

To model a gene regulatory network, an artificial agent-based model is well defined as general as possible with many molecules of various species interacting in the micro level, and a stochastic synergistic system is constructed to fit the realist experimental data, which are collected as integrated values in the macro level. Our combination method results that the hypothesized rules imposed on the behaviors of molecule agents can deductively infer the stochastic differential equations of the synergistic system. To implement our approach, we applied it to the eukaryote *Saccharomyces cerevisiae* data. For 2,819 target genes, their possible regulators were selected to quantify their dynamical transcription patterns. Several best fitting, worst fitting and vital genes were listed to show that our quantitative results agree well with the realist experimental data.

### Conclusion

Combining agent-based models together with stochastic synergistic systems affords a more complete perspective on gene regulatory networks, not only from the macro level but also from the micro level. It establishes a more solid linkage between *in-silico* simulation and real-world experiments.

Dock no.: 55

Submission no.: 175

**Towards understanding the role of natural genetic variation on lipid profiles.**

Husna Begum Begum and Markus Wenk

Department of Biochemistry, National University of Singapore

The completion of the initial phase of International Human Haplotype Map (the “HapMap”) project gave rise to many unanswered questions on how important natural variation in the genome plays a role in disease. Much work has been done in studying the genome, i.e. genomics as well as its active products such as RNA (transcriptomics) and proteins (proteomics). However, these fields are unable to explain the basis of genetic variation completely. The emerging field of lipidomics studies patterns of lipid profile across various pathologies. It is hoped that it would be able to complement the current research fields in providing further insights into the importance and role of natural genetic variations. This could be done by firstly determining with high accuracy, sensitivity and resolution the variation in around 400 lipid metabolites which cover different chemical classes and their various specific biological functions. Secondly, to understand using various bioinformatics tools the variations of metabolites and how genetic variations may be playing a role in these metabolite profile changes. Thus, it is aimed to link the fields of genomics and lipidomics to better understand the role of natural genetic variation on lipid profiles.

## Transcription analysis

Dock no.: 56

Submission no.: 83

Computational analysis of gene variants in SIM2 and ETS2 to identify their possible role in the functioning of the transcription factors.

Arpita Chatterjee and Kanchan Mukhopadhyay

Manovikas Biomedical Research and Diagnostic Centre, Manovikas Kendra, 482 Madudah, EM Bypass, Kolkata, 700107, India. Email: [arpita.chatterjee.2006@gmail.com](mailto:arpita.chatterjee.2006@gmail.com); [kanchanmvk@yahoo.com](mailto:kanchanmvk@yahoo.com)

Down's syndrome (DS) patients, commonly exhibiting trisomy 21, are prone to leukemia but resistant to solid tumors, especially breast cancer. Previous computational analysis by our group has revealed differential regulation of various tumor related genes by SIM2 and ETS2, two transcription factors (TFs) transcribed by genes located in the 21st chromosome and therefore, these TFs were speculated to have a role in the malignancy paradox of DS.

To evaluate the role of different gene variants on the expression and function of these two TFs, the present study was designed to identify various synonymous as well as non-synonymous variations.

From a large number of coding and noncoding genetic variations of SIM2 and ETS2, we have estimated the high risk variations by web-based computational methods. The risk of being deleterious for a non-synonymous SNP is calculated by SIFT and PolyPhen tool. The PupaSuite tool was used for predicting the effect of SNPs on the structure and function of the TFs. Risk levels were also verified by web-based programs like FastSNP, SNPeffect etc. Functional variations were tested computationally in respect to the post translational modification of the proteins.

High risk variations in SIM2 and ETS2, which showed deleterious /damaging effects, were sorted out. These variants could be useful as tools for studying functional alterations in the transcription factors and thus may throw some light in understanding the underlying disease pathology.

Dock no.: 61

Submission no.: 101

Towards Prediction of TCR Recognition of HLA/Peptide Complexes: A Structural Analysis

Stephanus Daniel Handoko, Chee Keong Kwoh and Yew Soon Ong

Nanyang Technological University

### **Background**

The key to adaptive immune response is TCR recognition of HLA/peptide complexes, which has to be preceded by HLA binding of some antigenic peptides. High polymorphism of the HLA genes has made the problem of exhaustively revealing the (TCR-)HLA/peptide interaction intractable. Both sequence- and structure-based predictors have been developed. However, despite the ability to produce better prediction quality, structure-based predictors are not generally suited for high-throughput screening. Therefore, it is desirable to harvest discriminative structure-based features to be used in conjunction with computational methods employed by sequence-based predictors. Internal coordinates, such as bond lengths, bond angles, and torsion angles, shall be computed and analyzed for each of the 9-mer antigenic peptides obtained from the Protein Data Bank.

### **Result**

All of the N-C $\alpha$ , C $\alpha$ -C, and C-N bond lengths as well as the N-C $\alpha$ -C, C $\alpha$ -C-N, and C-N-C $\alpha$  bond angles demonstrate only slight deviations. Most  $\omega$  torsion angles are fixed at about 180°, except some at 0°. In the absence of TCR, the  $\phi$  and  $\psi$  torsion angles are observed to be scattered in a range as large as 270°. In the presence of TCR, nonetheless, these torsion angles are observed to be clustered around some angular values.

### **Conclusion**

Consistent patterns observed from the  $\phi$  and  $\psi$  torsion angles may be an indication that an antigenic peptide, already bound by the HLA, may be further bound (recognized) by the TCR if it accurately has torsion angles that are of certain allowable values such that interaction with the TCR is geometrically possible. Based on the  $\phi$  and  $\psi$  torsion angles, a simple dichotomizer rule is established and applied successfully with 86.84% sensitivity and 88.89% specificity.