PROGRAM & ABSTRACTS

The 5th International Conference on Computational Systems-Biology and Bioinformatics

NANYANG TECHNOLOGICAL UNIVERSITY

NUS National University of Singapore

Agency for Science, Technology and Research

King Mongkut's University of Technology Thonburi

The International Neural Network Society (INNS)

Welcome Message from the General Chairs

On behalf of the Organizing Committee, it is our greatest pleasure to welcome you to the 5th International Conference on Computational Systems-Biology and Bioinformatics (CSBio2014). The conference runs from 10th to 12th November 2014 at Nanyang Technological University, which is a global university on a rapid rise with sustainable research, and it is the worlds biggest engineering university. The conference venue, Nanyang Executive Centre, is located in the Yunnan Garden Campus of Nanyang Technological University, offering a premier venue for various events.

This years conference is jointly organised by Nanyang Technological University, National University of Singapore, Agency for Science, Technology and Research and King Mongkut's University of Technology Thonburi, and we continue the good practice of collocating with IES2014.

The CSBio brings together researchers and practitioners to meet at this event to exchange ideas and stimulate research collaborations as a result of rapid advances in the generation of high-throughput "omics" data (such as deep sequencing). To extract biological knowledge from the data, and translate it into benefits for society (e.g. better medicine and healthcare), novel computational tools and insights are needed for data analysis and building models.

We are keeping to the tradition in carefully selecting a handful of high-quality papers from many countries. These will be recommended for journal publications to the Journal of Bioinformatics and Computational Biology (JBCB) or the Journal of Medical Imaging and Health Informatics (JMIHI).

This year we are very privilege to have Prof Ng Huck Hui, executive director of Genome Institute of Singapore, to deliver his keynote titled Systems Biology of Stem Cells.

A series of social functions have been planned, which include a welcome reception, lunches, conference banquet at Faber Peak Singapore, and tours to the Singapore Cable Car as well as Garden by the Bay. These social activities serve as a good channel to establish new connections and foster everlasting friendship among fellow counterparts.

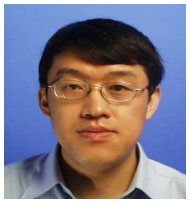
Last but not least, we would like to express our sincere gratitude to everyone involved in making the conference a success. Many thanks go to advisory board members, the organizing committees, the keynote and plenary speakers, the guest editors Prof (adj) Li Xiaoli, Prof Zheng Jie, and Prof Chou Siaw Meng for the two journals (JCBC and JMIHI), the program committee and reviewers, many of whom are active reviewers for the above journals, the conference participants, and of course, to all the contributing authors who will be sharing the interesting results of their research.

With our best wishes for a wonderful and stimulating stay in Singapore!

Conference Committee

General Chair
Kwoh Chee Keong, Singapore.
ASCKKWOH@ntu.edu.sg

General Co-Chair
Jonathan H. Chan, Thailand.
jonathan@sit.kmutt.ac.th



Program Chair
Li Xiaoli, Singapore.
xlli@i2r.a-star.edu.sg

Program Co-Chair
Zheng Jie, Singapore.
ZHENGJIE@ntu.edu.sg

Program Co-Chair
Chou Siaw Meng, Singapore.
MSMCHOU@ntu.edu.sg

Registration Chair
Huanjin Tang, Singapore.
htang@i2r.a-star.edu.sg

Publicity Chair
Wu Min, Singapore.
wumin@i2r.a-star.edu.sg

Local Arrangements Chair
Qi Cao, Singapore.

Secretary
Worrawat Engchuan, Thailand.

International Advisory Board

David W. Ussery
Oak Ridge National Lab, USA

Bairong Shen
Soochow University, China

Nikola Kasabov
Auckland University of Technology, New Zealand

Richard Wintle
The Centre for Applied Genomics, Canada

Stephen Wong
Houston Methodist Research Institute, USA

Supapon Cheevadhanarak
King Mongkut's University of Technology Thonburi, Thailand

Tom Lenaerts
Universit libre de Bruxelles, Belgium

Ong Yew Soon
Nanyang Technological University, Singapore

Jens B. Nielsen
Chalmers University of Technology, Sweden

Kwong-Sak Leung
Chinese University of Hong Kong, Hong Kong

Nikhil R. Pal
Indian Statistical Institute, India

Sissades Tongsima
National Center for Genetic Engineering and Biotechnology, Thailand

Sung-Bae Cho
Yonsei University, South Korea

Kay Chen Tan
National University of Singapore, Singapore

Yaochu Jin
University of Surrey, UK

PC members

Shandar Ahmad
National Institute of Biomedical Innovation, Japan

Tatsuya Akutsu
Kyoto University, Japan

Shunsuke Aoki
Kyushu Institute of Technology, Japan

Jonathan Chan
King Mongkut's University of Technology Thonburi, Thailand

Xin Chen
Nanyang Technological University, Singapore

Jin Chen
Michigan State University, USA

Jean-Paul Comet
University of Nice Sophia Antipolis, France

Frank Eisenhaber
Bioinformatics Institute, A*STAR, Singapore

Worrawat Engchuan
King Mongkut's University of Technology Thonburi, Thailand

Ge Gao
Peking University, China

Ivo Grosse
Martin Luther University of Halle-Wittenberg, Genmany

Chia-Lang Hsu
National Taiwan University, Taiwan

Hsuan-Cheng Huang
National Yang Ming University, Taiwan

Yaochu Jin
University of Surrey, UK

Tamer Kahveci
University of Florida, USA

Saowalak Kalapanulak
Bioinformatics and Systems biology Program, Thailand

Asif M. Khan
National University of Singapore, Singapore

Tsung Fei Khang
University of Malaya, Malaysia

Daisuke Kihara
Purdue University, USA

Jaebum Kim
Konkuk University, Korea

Sun Kim
National Institue of Health, USA

Yoo-Ah Kim
National Institue of Health, USA

Kengo Kinoshita
Tohoku University, Japan

Akihiko Konagaya
Tokyo Institute of Technology, Japan

Bartosz Krawczyk
Wroclaw University Of Technology, Poland

Igor Kurochkin
Bioinformatics Institute, A*STAR, Singapore

Chee-Keong Kwoh
Nanyang Technological University, Singapore

Teeraphan Laomettachit
King Mongkut's University of Technology Thonburi, Thailand

Hyunju Lee
Gwangju Institute of Science and Technology, Korea

Jeongjin Lee
Soongsil University, Korea

Tom Lenaerts
Universit libre de Bruxelles, Belgium

Jinyan Li
University of Technology, Sydney, Australia

Xiaoli Li
Institute for Infocomm Research, A*STAR, Singapore

C P Lim
Deakin University, Australia

Fenglou Mao
University of Georgia, USA

Osamu Maruyama
Kyushu University, Japan

Sebastian Maurer-Stroh
A*STAR, Singapore

Asawin Meechai
King Mongkut's University of Technology Thonburi, Thailand

Niranjan Nagarajan
Bioinformatics Institute, A*STAR, Singapore

Jin-Wu Nam
Hanyang University, Korea

Somnuk Phon-Amnuaisuk
Institut Teknologi Brunei, Brunei

Treenut Saithong
King Mongkut's University of Technology Thonburi, Thailand

Yasubumi Sakakibara
Keio University, Japan

Christian Schnbach
Nazarbayev University, Republic of Kazakhstan

Kumar Selvarajoo
Keio University, Japan

Bairong Shen
Soochow University, China

Soo-Yong Shin
Asan Medical Center, Korea

Kyung-Ah Sohn
Ajou University, Korea

Chuan-Kang Ting
National Chung Cheng University, Taiwan

Wanwipa Vongsangnak
Soochow University, Thailand

Lusheng Wang
City University of Hong Kong, Hong Kong

Min Wu
A*STAR, Singapore

Xuegong Zhang
Tsinghua University, China

Shihua Zhang
University of Southern Canlifornia, China

Yun Zheng
Fudan University, China

Jie Zheng
Nanyang Technological University, Singapore

Shuigeng Zhou
Fudan University, China

Shanfeng Zhu
Fudan University, China

Zexuan Zhu
Shenzhen University, China

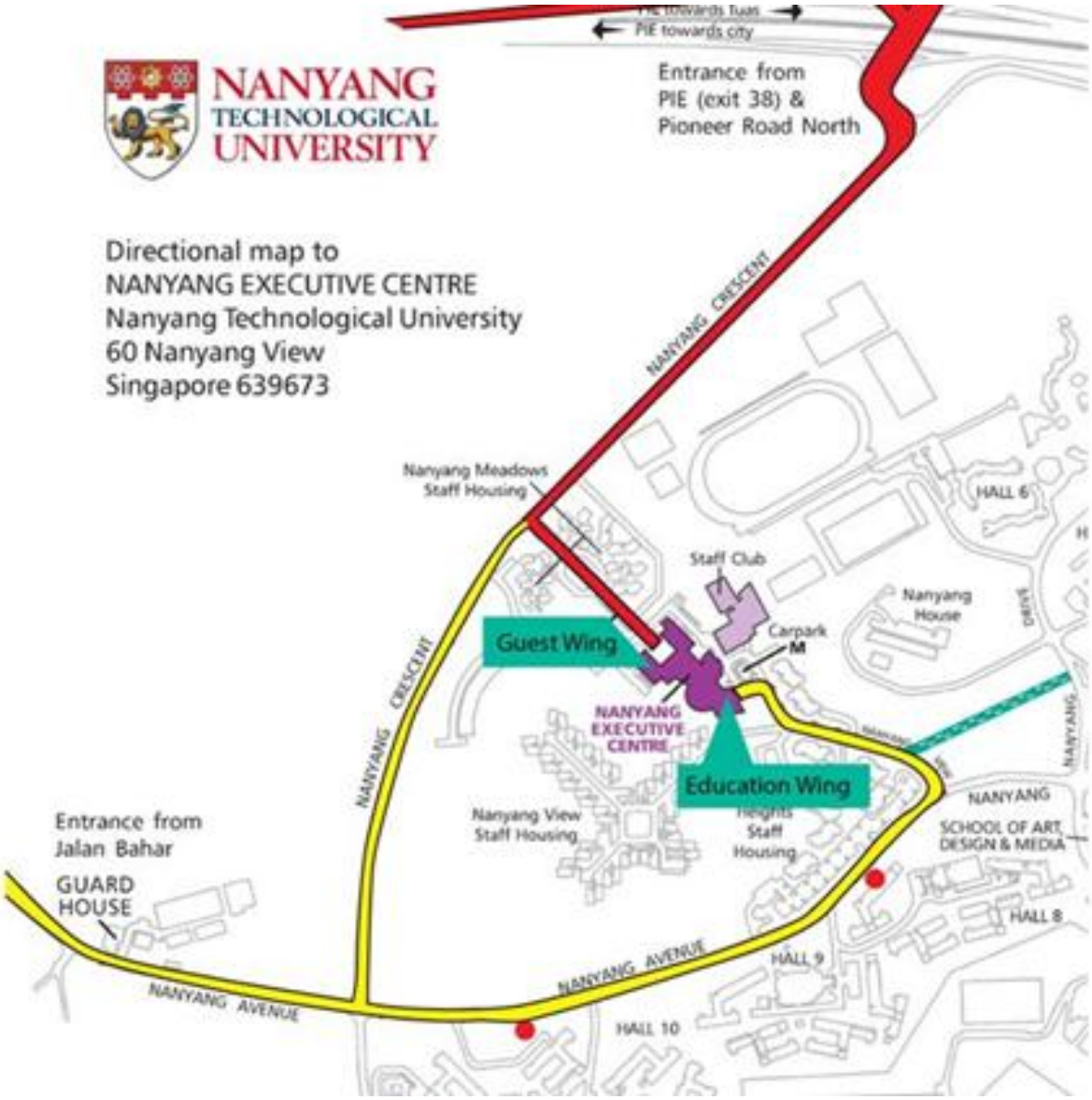
Conference Venue and Transportation

The Nanyang Executive Centre is the home of NTUs Executive Programmes and continuing education. It provides an excellent environment away from the city and distractions of the office for executive learning. The 170 guestrooms and suites have been designed to meet the needs of both training executives and business travellers.All the guestrooms are equippedwith awork area and othermodern amenities to ensure they provide guests with a pleasant stay. Complimentary Wi-Fi is available to all guests in their rooms, lobby and function spaces. Further, facilities such as swimming pool, gym, football field, basketball court, etc., are available around NEC.

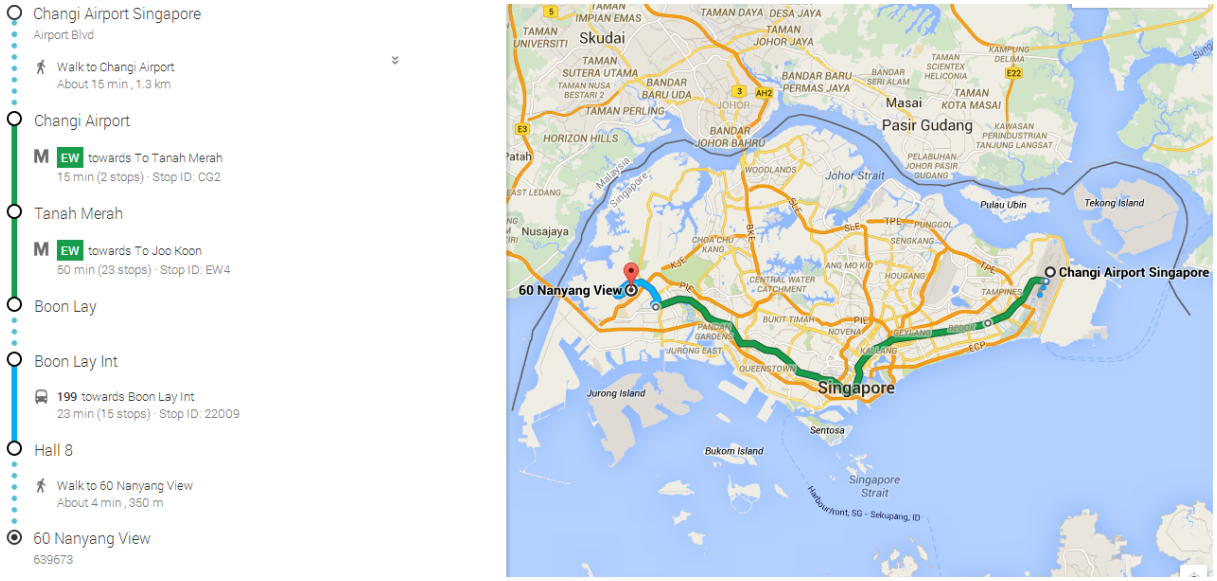
Nanyang Executive Centre (General enquiries)

60 Nanyang View, Singapore, 639673.
Tel: +65 6790 6699/ 6790 6697
Email: ntu-nec@ntu.edu.sg
http://www.ntu.edu.sg/nec/virtualTour/Pages/VirtualTour.aspx.

Nanyang Executive Centre



Map of Nanyang Executive Centre



Transportation form Singapore Changi Airport to Nanyang Executive Centre

Walk from Hall 8 to Nanyang Executive Centre

Keynote Speech: Systems Biology of Stem Cells

Professor Huck-Hui NG

Genome Institute of Singapore
nghh@gis.a-star.edu.sg

Date/Time: 08:30 – 09:30, Tuesday, 11 Nov. 2014
Venue: Auditorium, NTU

Biosketch of the Speaker

Professor Huck-Hui NG is the Executive Director of the Genome Institute of Singapore. The Genome Institute of Singapore was started as a national flagship programme to tap into the innovation and impact of deciphering the genetic blueprint of mankind. Today, it houses more than 200 researchers working on different aspects of Human Genomics (Human Genetics, Infectious Diseases, Cancer Therapeutics and Stratified Oncology, Stem Cell Genomics, Cancer Stem Cell Genomics, Translational Genomics, Computational and Systems Biology).

Huck-Hui NG graduated from the National University of Singapore with a first class Honor degree in Molecular and Cell Biology and obtained his PhD from the University of Edinburgh. He spent the next few years working at the Harvard Medical School as a Damon Runyon-Walter Winchell research fellow. Upon return to Singapore, he started his research programme on Stem Cell Genomics at the Genome Institute of Singapore.

His lab works on different aspects of Systems Biology of Stem Cells. Specifically, his group uses genome wide approaches to dissect the transcriptional regulatory networks in embryonic stem cells with the aim to identify key nodes in this network. This had led to the first paper on the whole genome and unbiased mapping of key transcription factors in mouse embryonic stem cells. His group also conducted the whole genome genetic screen for human embryonic stem cells. More recently, his lab has begun to investigate the reprogramming code behind the induction of pluripotency in somatic cells. Huck-Hui Ngs works have been published in journals such as Cell, Science and Nature. He also sits on the Editorial Board of International Journals such as Genes & Development. His papers have received over 12,000 citations and his H-index is 44.

His research work has earned him several prestigious national and international accolades including the Singapore Youth Award (2005 and 2010), the National Science Award 2007, the HUGO Chens New Investigator Award 2010 and the Presidents Science Award 2011.

Huck-Hui NG holds Adjunct Professor appointment at the National University of Singapore (Departments of Biochemistry and Biological Sciences) and the Nanyang Technological University (School of Biological Sciences). He is also the President of the Stem Cell Society, Singapore.

Abstract

Embryonic stem (ES) cells are characterized by their ability to self-renew and remain pluripotent. Transcription factors have critical roles in the maintenance of ES cells through specifying an ES-cell-specific gene expression program. Deciphering the transcriptional regulatory network that describes the specific interactions of these transcription factors with the genomic template is crucial for understanding the design and key components of this network. To gain insights into the transcriptional regulatory networks in ES cells, we use chromatin immunoprecipitation coupled to ultra-high-throughput DNA sequencing (ChIP-seq) to map the locations of sequence specific transcription factors. These factors are known to play different roles in ES cell biology. Our study provides new insights into the integration of these regulators to the ES cell-specific transcription circuitries. Collectively, the mapping of transcription factor binding sites identifies new features of the transcriptional regulatory networks that define ES cell identity. Using this knowledge, we investigate nodes in the network which when activated, will jump-start the ES cell-specific expression program in somatic cells.

Technical Programme

Technical Programme – Day 1: Tuesday, 11 Nov. 2014

08:00 onwards	Registration
08:15 – 08:30	Opening of CSBio

08:30 – 09:30	CSBio Keynote: Systems Biology of Stem Cells by Professor Huck-Hui NG, Genome Institute of Singapore
Venue	Auditorium, NTU

10:30 – 11:00	Coffee Break

12:00 – 13:00	Lunch Break

Session	Interaction, Biological Network and Pathway
Time	13:00 – 15:00
Venue	Auditorium, NTU
Chair	Kumar Selvarajoo

13:00 – 13:20	Invited Speech: Emergent Properties in Cell Signaling and Transcriptome-wide Response Kumar Selvarajoo
13:20 – 13:40	Investigating noise tolerance in an efficient engine for inferring biological regulatory networks. Asako Komori, Yukihiro Maki, Isao Ono and Masahiro Okamoto.
13:40 – 14:00	Predicting Essential Genes and Synthetic Lethality via Influence Propagation in Signaling Pathways of Cancer Cell Fates. Fan Zhang, Min Wu, Xuejuan Li, Xiaoli Li, Chee Keong Kwoh and Jie Zheng.
14:00 – 14:20	Computationally Predicting Protein-RNA Interactions Using Only Positive and Unlabeled Examples. Zhanzhan Cheng, Shuigeng Zhou and Jihong Guan.

14:20 – 14:40	Network Reconstruction vs. Edge Enrichment: Which is Better in Protein Function Prediction Based on Protein-Protein Interaction Networks? Wei Xiong, Luyu Xie, Jihong Guan and Shuigeng Zhou.
14:40 – 15:00	Detection of Highly Overlapping Communities in Complex Networks. Madhusudan Paul, Rishav Anand and Ashish Anand.

15:00 – 15:30	Coffee Break

Session	Medical Data Diagnosis, Classification and Simulation
Time	15:30 – 17:30
Venue	Auditorium, NTU
Chair	Shirley Sue

15:30 – 15:50	Microarray-based cancer diagnosis using an integrative gene-set analysis approach. Worrawat Engchuan, Asawin Meechai, Sissades Tongsima and Jonathan Chan.
15:50 – 16:10	Comparison of Different ROI Sizes for Breast Density Classification: An Experimental Study. Vipul Sharma and Sukhwinder Singh.
16:10 – 16:30	Real-time Analysis of Vital Signs Using Incremental Data Stream Mining Techniques with a Case Study of ARDS under ICU Treatment. Simon Fong, Shirley Siu, Suzy Zhou, Jonathan Chan, Sabah Mohammed and Jinan Fiaidhi.
16:30 – 16:50	Numerical Simulation of Contaminant Control in Multi-patient Intensive Care Unit of Hospital Using Computational Fluid Dynamics. Tikendra Nath Verma and Shobha Lata Sinha.
16:50 – 17:10	Efficient Variation-based Feature Selection for Medical Data Classification. Simon Fong, Justin Liang, Shirley Siu and Jonathan Chan.
17:10 – 17:30	Docking Score Calculation Using Machine Learning with an Enhanced Inhibitor Database. Masato Okada, Hayato Ohwada and Shin Aoki.

17:40 – 21:00	Award Banquet

Technical Programme – Day 2: Wednesday, 12 Nov. 2014

Session	Sequence, Structure and Parallel Computing
Time	08:20 – 10:40
Venue	Auditorium, NTU
Chair	TBA

08:20 – 08:40	FQC: a novel approach for efficient compression, archival and dissemination of fastq datasets. Anirban Dutta, Mohammed Monzoorul Haque, Tungadri Bose, Cvsk Reddy and Sharmila Mande.
08:40 – 09:00	Grid-assembly: an oligo-nucleotide composition based partitioning strategy to aid metagenomic sequence assembly. Tarini Shankar Ghosh, Varun Mehra and Sharmila S Mande.
09:00 – 09:20	PSOVINA: the hybrid particle swarm optimization algorithm for protein-ligand docking. Marcus C. K. Ng, Simon Fong and Shirley W. I. Siu.
09:20 – 09:40	Identifying SNP-SNP interactions with CNAs in lymphoma susceptibility. Tse-Yi Wang, Yen-Ho Chen and Kuang-Chi Chen.
09:40 – 10:00	PARALLELFRAPPE: parallel version of admixture calculation. Alongkot Burutarchanai and Prabhas Chongstitvatana.
10:00 – 10:20	Massively Parallel Tool for Protein Structure Comparison. Ahmad Salah, Kenli Li and Tarek Gharib.
10:20 – 10:40	Embedding Assisted Prediction Architecture for Event Trigger Identification. Yifan Nie, Wenge Rong, Yiyuan Zhang, Yuanxin Ouyang and Zhang Xiong.

10:40 – 11:00	Coffee Break

13:00 – 14:00	Lunch Break

14:00 – 18:00	Free Tour

Abstracts

Day 1: Tuesday, 11 Nov. 2014

Session	Interaction, Biological Network and Pathway
Time	13:00 – 15:00
Venue	Auditorium, NTU
Chair	Kumar Selvarajoo

13:00 – 13:20
Invited Speech: Emergent Properties in Cell Signaling and Transcriptome-wide Response

Kumar Selvarajoo[1,a]

[1] *Institute for Advanced Biosciences Keio University, Japan*
E-mail: [a] *kumar@ttck.keio.ac.jp*

The recent surge in systems biology efforts is revealing many complexities that cannot be understood using traditional causal relationship approaches. For example, gene expressions between identical single cells are shown to be stochastic in time, and clonal population of cells display heterogeneity in the abundance of a given protein per cell at any measured time. Despite this, at population level, cells are able to execute well-defined biological processes such as growth, division, differentiation and immune response. How do living systems overcome single cell fluctuations to produce deterministic population response? Our research focuses on fundamental cell behaviors. We investigate the dynamic instructive cell signaling and highthroughput transcriptome-wide behaviors of immune, cancer and embryonic development cells, from single cells to population scale. As a general feature, we have found that single cell fluctuations reduce when cells form ensembles, due to the reduction of dominant stochastic and transcriptome-wide noise as the number of cells or molecules within a cell is increased. This resulted in deterministic cellular response that can be modeled using linear response rules. Adopting the rules, we have successfully predicted and experimentally verified novel signaling features, such as missing intermediates and crosstalk mechanisms, and have identified novel targets for controlling proinflammatory response and cancer apoptosis. Specifically, I will provide an overview on toll-like receptors, tumor necrosis factor (TNF) receptor, and TNF-related apoptosis-inducing ligand receptor signaling. I will also show RNA-Seq based single cell transcriptome-wide data analysis for embryonic developmental process (from oocyte to ES cell stage).

13:20 – 13:40
Investigating noise tolerance in an efficient engine for inferring biological regulatory networks

Asako Komori[1,a], Yukihiro Maki[2,c], Isao Ono[3,d] and Masahiro Okamoto[1,2,b]

[1] *Department of Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University Fukuoka, 8128582, Japan*
[2] *Advanced Research Center for Synthetic Systems Biology, Kyushu University Fukuoka, 8128582, Japan*
[3] *Graduate School of Computational Intelligence and Systems Science, Tokyo Institute of Technology Tokyo, 2268502, Japan*
E-mail: [a] *a-komori,* [b] *okahon@brs.kyushu-u.ac.jp,* [c] *ykhrmaki@tos.bbiq.jp,* [d] *isao@dis.titech.ac.jp*

Biological systems are composed of biomolecules such as genes, proteins, metabolites, and signaling components, which interact in complex networks. To understand complex biological systems, it is important to be capable of inferring regulatory networks from experimental time series data. In previous studies, we

developed efficient numerical optimization methods for inferring these networks, but we have yet to test the performance of our methods when considering the error (noise) that is inherent in experimental data. In this study, we investigated the noise tolerance of our proposed inferring engine. We prepared the noise data using the Langevin equation, and compared the performance of our method with that of alternative optimization methods.

Keywords: S-system formalism; real-coded genetic algorithm; gene regulatory networks; systems biology; system identification.

13:40 – 14:00
Predicting Essential Genes and Synthetic Lethality via Influence Propagation in Signaling Pathways of Cancer Cell Fates

Fan Zhang[1,a], Min Wu[2,e], Xuejuan Li[1,b], Xiaoli Li[2,f], Chee Keong Kwoh[1,c], Jie Zheng[1,3,d]

[1] *School of Computer Engineering, Nanyang Technological University, Singapore 639798*
[2] *Data Analytic Department, Institute for Infocomm Research, A*STAR, Singapore 138632*
[3] *Genome Institute of Singapore, A*STAR, Singapore 138672*
E-mail: [a] *fzhang005@e.ntu.edu.sg,* [b] *lixj,* [c] *asckkowh,* [d] *zhengjie@ntu.edu.sg,* [e] *wumin,* [f] *xlli@i2r.a-star.edu.sg*

A major goal of personalized anti-cancer therapy is to increase the drug efects while reducing the side-efects as much as possible. A novel therapeutic strategy called synthetic lethality (SL) provides a great opportunity to achieve this goal. SL arises if mutations of both genes lead to cell death while mutation of either single gene does not. Hence, the SL partner of a gene mutated only in cancer cells could be a promising drug target, and the identification of SL pairs of genes is of great significance in pharmaceutical industry. In this paper, we propose a hybridized method to predict SL pairs of genes. We combine a data-driven model with knowledge of signalling pathways to simulate the influence of single gene knock-down and double genes knock-down to cell death. A pair of genes is considered as an SL candidate when double knock-down increases the probability of cell death significantly, but single knock-down does not. The single gene knock-down is confirmed according to the human essential genes database. Our validation against literatures shows that the predicted SL candidates agree well with wet-lab experiments. A few novel reliable SL candidates are also predicted by our model.

Keywords: Synthetic Lethality; Signaling Pathways; Data-driven.

14:00 – 14:20
Computationally Predicting Protein-RNA Interactions Using Only Positive and Unlabeled Examples

Zhanzhan Cheng[1,a], Shuigeng Zhou[1,b] and Jihong Guan[2,c]

[1] *Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science Fudan University, 220 Handan Road, Shanghai 200433, China*
[2] *Department of Computer Science and Technology Tongji University, 4800 Cao'an Road, Shanghai 201804, China*
E-mail: [a] *zhanzhancheng,* [b] *sgzhou@fudan.edu.cn,* [c] *jhguan@tongji.edu.cn*

Protein-RNA interactions (PRIs) are considerably important in a wide variety of cellular processes, ranging from transcriptional and post-transcriptional regulations of gene expression to the active defense of host against virus. With the development of high throughput technology, large amounts of protein-RNA interaction information is available for computationally predicting unknown PRIs. In recent years, a number of computational methods for predicting PRIs have been developed in the literature, which usually artificially construct negative samples based on verified non-redundant datasets of PRIs to train classifiers. However, such negative samples are not real negative samples, some even may be un-

known positive samples. Consequently, the classifiers trained with such training datasets can not achieve satisfactory prediction performance.

In this paper, we propose a novel method PRIPU that employs biased-SVM for predicting Protein-RNA Interactions using only Positive and Unlabeled examples. To the best of our knowledge, this is the first work that predicts PRIs using only positive and unlabeled samples. We first collect known PRIs as our benchmark datasets and extract sequence-based features to represent each PRI. To reduce the dimension of feature vectors for lowering computational cost, we select a subset of features by a filter-based feature selection method. Then, biased-SVM is employed to train prediction models with different PRI datasets. To evaluate the new method, we propose a new performance measure called explicit positive recall (EPR), which is specifically suitable for the task of learning positive and unlabeled data.

Experimental results over three datasets show that our method not only outperforms four existing methods, but also is able to predict unknown PRIs.

Keywords: Protein-RNA Interactions; Biased-SVM; Prediction.

14:20 – 14:40
Network Reconstruction vs. Edge Enrichment: Which is Better in Protein Function Prediction Based on Protein-Protein Interaction Networks?

Wei Xiong[1], Luyu Xie[1], Jihong Guan[2] and Shuigeng Zhou[1,a]

[1] *School of Computer Science, and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China*
[1] *Department of Computer Science & Technology, Tongji University, Shanghai, China*
E-mail: [a] *sgzhou@fudan.edu.cn*

Background: The high-throughput technologies have led to vast amounts of protein-protein interaction (PPI) data, and a number of approaches based on PPI networks have been proposed for protein function prediction. Unfortunately, these PPI networks face serious data quality challenges of false positives (noise) and false negatives (incompleteness), which adversely affects the performance of protein function prediction.

Methodology: To boost the performance of protein function prediction based on PPI networks, two major types of approaches were proposed to construct more robust and reliable PPI networks, including network reconstruction and edge enrichment. Although various implementations of these two types of approaches were reported, and definite performance improvements were achieved, there lacks a systematic performance comparison between these two types of approaches. To answer this question, in this paper we carry out a comprehensive performance comparison of these two types of approaches. Concretely, we first reconstruct and enrich PPI networks by using protein sequence similarity, local similarity indices and global similarity indices, and then compare the prediction performance of these reconstructed and enriched networks with that of the original networks of two real PPI datasets.

Conclusions: Experimental results demonstrate that the enriched networks achieves more accurate predictions than the original networks and the reconstructed networks. Meanwhile, the reconstructed network performs better than the original network of the BioGRID dataset that contains only physical interactions. However, the original network of the STRING dataset that contains known and predicted interactions outperforms the reconstructed network. This means that network reconstruction is more effective in relatively small and incomplete PPI networks. Our empirical study also validates that sequence similarity is more effective than global similarity and local similarity in PPI network enrichment.

14:40 – 15:00
Detection of Highly Overlapping Communities in Complex Networks

Madhusudan Paul[1,a], Rishav Anand[1,b] and Ashish Anand[1,c]

[1] *Department of Computer Science and Engineering Indian Institute of Technology Guwahati, India*
E-mail: [a] *madhusudan,* [b] *anand.ashish@iitg.ernet.in,* [c] *rishav@alumni.iitg.ernet.in*

Detecting communities in complex networks is one of the most important aspects to understand complex systems. In reality, many of these communities are highly overlapping in nature, i.e., several nodes belong to more than three communities. Identification of highly overlapping communities are strongly demanded in many applications such as systems biology and social networks. Although there are algorithms for detecting overlapping communities, majority of these are unable to detect highly overlapping communities properly. The performance of these algorithms falls sharply when overlapping nodes belong to more than three communities. In this paper, we propose an extension of existing overlapping community detection algorithm, namely Greedy Clique Expansion (GCE). Due to lack of unavailability of real networks with complete information of ground-truth communities, firstly, we experiment on state-of-the-art synthetic benchmark datasets and observe that our proposed extension exhibits better performance when overlapping nodes belong to more than three communities. We also experiment on real datasets and observe competitive performance. The proposed extension can be applied on networks with highly overlapping community structure such as protein-protein interaction networks.

Keywords: Overlapping community detection; Greedy Clique Expansion; complex net- works.

15:00 – 15:30: Coffee Break

Session	Medical Data Diagnosis, Classi– –fication and Simulation
Date/Time	15:30 – 17:30
Venue	Auditorium, NTU
Chair	Shirley Sue

15:30 – 15:50
Microarray-based cancer diagnosis using an integrative gene-set analysis approach

Worrawat Engchuan[1,a], Asawin Meechai[2,b], Sissades Tongsima[3,c] and Jonathan Chan[1,d]

[1] *Data and Knowledge Engineering Laboratory, School of Information Technology King Mongkuts University of Technology Thonburi, Bangkok, Thailand*
[2] *Department of Chemical Engineering, Faculty of Engineering King Mongkuts University of Technology Thonburi, Bangkok, Thailand*
[3] *Biostatistics and Informatics Laboratory, Genome Institute National Center for Genetic Engineering and Biotechnology*
E-mail: [a] *worrawat.eng@st.sit.kmutt.ac.th,* [b] *asawin.mee@kmutt.ac.th,* [c] *sissades@biotec.or.th,* [d] *jonathan@sit.kmutt.ac.th*

Cancer is a complex disease that cannot be diagnosed reliably using only single gene expression analysis. Hence it is fast becoming the norm to do gene-set analysis instead. This work provides a tool (Pathway Activity Toolbox: PAT) that incorporates three gene-set analysis methods, namely ANOVA-based feature set (AFS), Negatively Correlated Feature Set (NCFS-i) and Conditional-Responsive gene (CORG)-based. These methods identify subsets of phenotype-relevant genes, which will be used to transform the gene expression levels to gene-set activities. As a case study, the classifications using the expression levels of these gene subsets without transformation and the gene-set activities are compared in terms of the performance for disease diagnosis of three common cancer

types. Additionally, several gene-set collections and classifiers can be optimally chosen for each disease. We demonstrate the performance of PAT by running it using three gene-set collections from MSigDB and three traditional classifiers on eight actual microarray datasets of lung, colorectal and breast cancers. Five-by-Two-fold cross-validation is used for parameters optimization and the overall classification performance is evaluated by cross dataset validation. The comparison results show that the selections of gene-set analysis methods, gene-set collections, and classifiers can be made differently depending on the disease. The results show that the use of Gene-Set-Analysis (GSA)-based gene markers is preferred to gene-set activity markers for disease diagnosis. PAT can be accessed via `http://pat.sit.kmutt.ac.th` from which its java library for gene-set analysis, simple classification and a database with eight benchmark datasets of three cancers can be downloaded.

Keywords: Microarray; Gene expression analysis; Gene set; Classification; Feature selection; Breast cancer; Lung cancer; Colorectal cancer.

15:50 – 16:10
Comparison of Different ROI Sizes for Breast Density Classification: An Experimental Study

Vipul Sharma[1,a] and Sukhwinder Singh[1,b]

[1] *UIET, Panjab University, Sector-25 Chandigarh, 160014, INDIA*
E-mail: [a] *vipuls85@gmail.com,* [b] *sukhdalip@pu.ac.in*

In this paper, an experimental study has been done to find the optimal size of Region of Interest (ROI) for the classification of breast density. ROI is a subpart of the image that contains very important information related to the diagnosis. Since an ROI is used as representative of the image, and all further computations and diagnosis depends upon the ROI, therefore, it is very crucial to select an appropriate image area as ROI. For this purpose, different sized ROIs are selected from mammogram in such a way that their central pixel is same. The effect of ROI size is evaluated in terms of the performance of differentiating between the fatty and dense breast tissue. From the experimental results it has been found that ROI of size 200200 pixels results in maximum overall accuracy of 97.2% with 100% sensitivity. The experimental results encourage the use of 200200 pixels ROI for classification of breast density.

Keywords: Region of Interest; Breast Density; Classification; Texture Features; Feature Selection.

16:10 – 16:30
Real-time Analysis of Vital Signs Using Incremental Data Stream Mining Techniques with a Case Study of ARDS under ICU Treatment

Simon Fong[1,a], Shirley Siu[1,b], Suzy Zhou[2,c], Jonathan Chan[3,d], Sabah Mohammed[4,e] and Jinan Fiaidhi[4,f]

[1] *Department of Computer and Information Science University of Macau Macau SAR*
[2] *Department of Products Management Mozat Pte Ltd Singapore*
[3] *School of Information Technology King Mongkut's University of Technology Thonburi Bangkok, Thailand*
[4] *Department of Computer Science Lakehead University Thunder Bay, Canada*
E-mail: [a] *ccfong,* [b] *shirley.siu@umac.mo,* [c] *suzyzhou@mozat.com,* [d] *jonathan@sit.kmutt.ac.th,* [e] *sabah.mohammed,* [f] *jinan.fiaidhi@lakeheadu.ca*

Analysing data streams of vital signs has been a popular topic in research communities with techniques mainly of detection, classification and prediction. One drawback for data classification/prediction is that the data mining model is built based on a full set of stationary data. Updating the model for sustaining the classification accuracy often needs the whole dataset including the evolving data to be accessed. This nature of model rebuilding dampers the possibility of mining vital signs in real-time and at high speed. Unfortunately all the past papers in the literature were based on traditional data mining

models. In this paper, a data stream mining model which is flexible in configuring with different incremental data stream learning methods is tested as the real-time classification engine for mining vital data streams. A computer simulation experiment is conducted that is based on a case study of adult respiratory distress syndrome under twelve-hours of ICU treatment. The results indicate promising possibilities of performing real-time prediction by the proposed model.

Keywords: Vital Signs Analysis, Data Stream Mining, Naviie Bayes, Support Vector Machine, Optimized Very Fast Decision Tree.

16:30 – 16:50
Numerical Simulation of Contaminant Control in Multi-patient Intensive Care Unit of Hospital Using Computational Fluid Dynamics

Tikendra Nath Verma[1,a] and Shobha Lata Sinha[1,b]

[1] *National Institute of Technology Raipur, Department of Mechanical Engineering, Raipur, Chhattisgarh 492010, India*
E-mail: [a] *verma_tikendra@yahoo.co.in,* [b] *shobha_sinha1@rediffmail.com*

The complex hospital environment requires special attention to ensure healthy indoor air quality (IAQ). It is necessary to protect to patient and healthcare workers against hospital acquired pathogens/infections. A CFD analysis supported by measurement data has been carried out to simulate the temperature distribution, air flow pattern in the ICU and the contaminant dispersion from the patient. Numerical model solves conservation of mass, momentum and energy equations. The contaminated particle (infection) which is simulated with a Lagrangian particle tracking model and the same has been carried out by simulation. CFD analysis is used to simulate and compare the removal of contaminated particle using a number of different ventilation systems. The study concerns about the health risk of the airborne diseases (bacteria, fungus, viruses etc.) from the patient to the other people in the ICU of hospital. A healthy environment can be achieved by minimizing the risk of contamination through appropriate filtration and air distribution scheme. It is observed that stagnant zone near the doctor and patient is not healthy. Therefore suitable ventilation arrangement and position of patient (bed) must be provided for healthy environment in the ICU.

Keywords: Numerical simulation; ICU; Contaminant particle; Lagrangian model.

16:50 – 17:10
Efficient Variation-based Feature Selection for Medical Data Classification

Simon Fong[1,a], Justin Liang[1], Shirley Siu[1] and Jonathan Chan[2,b]

[1] *Department of Computer and Information Science, University of Macau Taipa, Macau SAR*
[2] *School of Information Technology, King Mongkut's University of Technology Thonburi Bangkok, Thailand*
E-mail: [b] *jonathan@sit.kmutt.ac.th,* [a] *ccfong@umac.mo*

Medical data which collected from sophisticated and sometimes different instruments are described by a large number of feature variables and a historical archive of patients records, known as multivariate medical dataset. In biomedical data mining, classification model is often built upon such dataset for predicting which particular type of disease that a new instance of record belongs to. One of the challenges in inferring a classification model with good prediction accuracy is to select the relevant features that contribute to maximum predictive power. Many feature selection techniques have been proposed and studied in the past, but none so far claimed to be the best. In this paper, a novel and efficient feature selection method called Clustering Coefficients of Variation (CCV) is proposed. CCV is based on a very simple principle of variance-basis which finds an optimal balance between generalization and overfitting. Through a computer simulation experiment,

eight medical datasets with substantially large number of features are tested by CCV in comparison to three popular feature selection techniques. Results show that CCV outperformed them in all aspects of averaged performances and speed. By the simplicity of design it is anticipated that CCV will be a useful alternative of feature selection method for tasks of medical data classification especially with those datasets that are characterized by many features.

Keywords: Feature selection; Classification; Medical datasets.

17:10 – 17:30
Docking Score Calculation Using Machine Learning with an Enhanced Inhibitor Database

Masato Okada[1,a], Hayato Ohwada[1,b] and Shin Aoki[1,c]

[1] *Faculty of Science and Technology, Tokyo University of Science Noda, Chiba, 278-8510 Japan*
E-mail: [a] *okada@ohwada-lab.net,* [b] *ohwada@rs.tus.ac.jp,* [c] *shinaoki@rs.noda.tus.ac.jp*

This paper describes a machine-learning method for docking score calculation with the Database of Useful Decoys: Enhanced (DUD-E). This database includes both good inhibitors (ligands) and poor inhibitors (decoys), allowing machine learning to predict appropriate docking scores of a ligand and the associated decoys. This property enables us to find new inhibitor candidates with high accuracy and to screen many compounds with excellent performance. The proposed method can also be applied to any enzymes without the use of the molecular structure of an enzyme, outperforming a number of traditional docking software tools in both predictive accuracy and generality. We selected 10 enzymes from DUD-E and conducted a comparative study using two docking software tools. The classification performance was obtained from an experiment where 2869 ligands were predicted from 2985 actual ligands, and 8923 decoys were predicted from 8955 registered decoys. Such excellent performance is visualized by Receiver Operating Characteristic (ROC) curves at various threshold settings.

Keywords: Machine learning; In silico screening Docking score.

Day 2: Wednesday, 12 Nov. 2014

Session	Sequence, Structure and Parallel Computing
Time	08:20 – 10:40
Venue	Auditorium, NTU
Chair	TBA

08:20 – 08:40
FQC: a novel approach for efficient compression, archival and dissemination of fastq datasets

Anirban Dutta[1], Mohammed Monzoorul Haque[1], Tungadri Bose[1], Cvsk Reddy[1] and Sharmila Mande[1,a]

[1] *Bio-Sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services Limited, 54-B, Hadapsar Industrial Estate, Pune 411013, Maharashtra, India*
E-mail: [a] *sharmila.mande@tcs.com*

Sequence data repositories archive and disseminate fastq data in compressed format. In spite of having relatively lower compression efficiency, data repositories continue to prefer GZIP over available specialized fastq compression algorithms. Ease of deployment, high processing speed and portability are the reasons for this preference. This study presents FQC – a fastq compression method that, in addition to providing significantly higher compression gains over GZIP, incorporates features necessary for universal adoption by data repositories/end-users. This study also proposes a novel archival strategy which allows sequence repositories to simultaneously store and disseminate lossless as well as (multiple) lossy variants of fastq files, without necessitating any additional storage requirements. For academic users, Linux, Windows and Mac implementations (both 32 and 64-bit) of FQC are freely available for download at: `https://metagenomics.atc.tcs.com/compression/FQC`.

Keywords: Data compaction and compression; Algorithms for biological data management; NGS data; Sequencing data archival.

08:40 – 09:00
Grid-assembly: an oligo-nucleotide composition based partitioning strategy to aid metagenomic sequence assembly

Tarini Shankar Ghosh[1], Varun Mehra[1] and Sharmila S Mande[1,a]

[1] *Biosciences R&D Division, TCS Innovation Labs, 54-B Hadapsar Industrial Estate Pune, Maharashtra 411013, India*
E-mail: [a] *sharmila@atc.tcs.com*

Metagenomics approach involves extraction, sequencing and characterization of the genomic content of entire community of microbes present in a given environment. In contrast to genomic data, accurate assembly of metagenomic sequences is a challenging task. Given the huge volume and the diverse taxonomic origin of metagenomic sequences, direct application of single genome assembly methods on metagenomes are likely to not only lead to an immense increase in requirements of computational infrastructure, but also result in the formation of chimeric contigs. A strategy to address the above challenge would be to partition metagenomic sequence datasets into clusters and assemble separately the sequences in individual clusters using any single-genome assembly method. The current study presents such an approach that uses tetra-nucleotide usage patterns to first represent sequences as points in a 3D-space. The 3D-space is subsequently partitioned into 'Grids'. Sequences within overlapping grids are then progressively assembled using any

available assembler. We demonstrate the applicability of the current Grid-Assembly method using various categories of assemblers as well as different simulated metagenomic datasets. Validation results indicate that the Grid-Assembly approach helps in improving the overall quality of assembly, in terms of the purity and volume of the assembled contigs.

Keywords: Metagenomics; Sequence Assembly; Tetra-nucleotide frequency.

09:00 – 09:20
PSOVINA: the hybrid particle swarm optimization algorithm for protein-ligand docking

Marcus C. K. Ng[1,a], Simon Fong[1,b] and Shirley W. I. Siu[1,c]

[1]*Department of Computer and Information Science, University of Macau Avenida da Universidade, Taipa Macau S.A.R, China*
E-mail: [a]*marcus.ckng@gmail.com,* [b]*ccfong,* [c]*shirleysiu@umac.mo*

Protein-ligand docking is an essential step in modern drug discovery process. The challenge here is to accurately predict and efficiently optimize the position and orientation of ligands in the binding pocket of a target protein. In this paper, we present a new method called PSOVina which combined the Particle Swarm Optimization (PSO) algorithm with the efficient BFGS local search method adopted in AutoDock Vina to tackle the conformational search problem in docking. Using a large and diverse data set of 201 protein-ligand complexes from the PDBbind database, we assessed the predictive performance of PSOVina in comparison to the original Vina program. Our docking simulations showed that PSOVina has a remarkable speed-up of over 45% in terms of average docking time. Regarding docking accuracy, PSOVinas predicted ligand conformations are on average 7% lower in RMSD compared to Vina with an increase of prediction success rate by 4%. Our work proves that PSO is superior to Monte Carlo search method in molecular docking applications and lays the foundation for the future development of swarm-based algorithms in docking programs.

Keywords: particle swarm optimization; protein-ligand docking; flexible docking, conformational search; AutoDock; drug design.

09:20 – 09:40
Identifying SNP-SNP interactions with CNAs in lymphoma susceptibility

Tse-Yi Wang[1,a], Yen-Ho Chen[2,b] and Kuang-Chi Chen[2,c]

[1]*Laboratory of Molecular Anthropology and Transfusion Medicine, Department of Medical Research, Mackay Memorial Hospital, New Taipei City 25160, Taiwan*
[2]*Department of Medical Informatics, Tzu Chi University, Hualien 97004, Taiwan*
E-mail: [a]*tseyiwang,* [b]*yhchen1210@gmail.com,* [c]*chichen6@mail.tcu.edu.tw*

Genome-wide association studies (GWAS) identify many single nucleotide polymorphisms (SNPs) that are associated with diseases and are involved in their pathogenesis. The currently identified SNP variants only explain a portion of the heritability underlying the complex diseases. Some recent studies have focused on investigating SNP-SNP interactions to explore the missing aspects of the heritability. Several have looked for genetic variations other than SNPs such as copy number alterations (CNAs). However, few have considered both SNP-SNP interactions and CNAs comprehensively. In our study, we employed a fusion approach that incorporated the information of copy numbers to identify SNP-SNP interactions in analyzing a public dataset of 214 lymphoma cases. The copy numbers were first examined by clustering analysis, and then the SNP-SNP interactions were detected by the multifactor dimensionality reduction (MDR) method. The results showed that the identified SNP-SNP interactions with CNAs were more significantly associated with lymphoma susceptibility than the SNP-SNP interactions detected without regarding CNAs. Therefore, we conclude that com-

bining SNP-SNP interactions with CNAs provides a more comprehensive strategy for disease association studies.

09:40 – 10:00
PARALLELFRAPPE: parallel version of admixture calculation

Alongkot Burutarchanai[1,1] and Prabhas Chongstitvatana[1,b]

[1] *Department of Computer Engineering, Chulalongkorn University, Address Bangkok, 10330, THAILAND*
E-mail: [a] *Alongkot.bu@gmail.com,* [b] *prabhas.c@chula.ac.th*

ParallelFrappe is software for the analysis of individual admixture for Individual Identification by genetic data. The ParallelFrappe is a parallel version of Frappe program. This version can achieve speed up of the calculation with multicore processors. The program was implemented with a 64-bit platform to improve the performance on large dataset. The algorithm for the parallel machine was adapted to match with the multiprocessor architecture. The program also has a milestone feature that it can re-execute from the previous stop point which is very important for a long running job.

Keywords: Individual Identification; Admixture Analysis; Expectation Maximization.

10:00 – 10:20
Massively Parallel Tool for Protein Structure Comparison

Ahmad Salah[1,2,a], Kenli Li[3,b] and Tarek Gharib[4,5,c]

[1] *College of Information Science and Engineering, Hunan University Changsha, Hunan 410082, China*
[2] *College of Computers and Informatics, Zagazig University Zagazig, Egypt*
[3] *The National Supercomputing Center in Changsha and College of Information Science and Engineering, Hunan University Changsha, Hunan 410082, China*
[4] *Faculty of Computing and Information Technology, King Abdulaziz University Jeddah, Saudi Arabia*
[5] *Faculty of Computer and Information Sciences, Ain Shams University Cairo, Egypt*
E-mail: [a] *ahmad,* [b] *lkl@hnu.edu.cn,* [c] *tfgharib@kau.edu.sa*

At the center of computational structure biology, protein structure comparison, or alignment, is a key problem. The steady increase in the number of protein structures encourages the development of massively parallel methods and tools. While the focus of research is to propose data-analytical methods to tackle this problem, there are limited practical efforts to propose generic tools to run these methods in different, parallel environments other than the one these methods are targeting. In the present work, we propose a scalable tool to handle the rapid growth of the structural proteomic data using massively parallel environments. The proposed tool can run the sequential methods on parallel environments or extend the parallelism level from multicore to cluster. The tool requires no scripting or installation. The tool achieves a linear, close to optimal, speedup. The experimental results show no effect on the accuracy of the used methods. The tool is available at `http://biocloud.hnu.edu.cn/ppsc/`.

Keywords: Protein structure comparison; parallel tool; multicore; grid; speedup.

10:20 – 10:40
Embedding Assisted Prediction Architecture for Event Trigger Identification

Yifan Nie[1,a], Wenge Rong[2,3,b], Yiyuan Zhang[2,e], Yuanxin Ouyang[2,3,c] and Zhang Xiong[2,3,d]

[1] *Sino-French Engineering School, Beihang University, Beijing 100191, China*
[2] *School of Computer Science and Engineering, Beihang University, Beijing 100191, China*
[3] *Research Institute of Beihang University in Shenzhen, Shenzhen 518057, China*
E-mail: [a] *yifan.nie@ecpk.buaa.edu.cn,*

[b] *w.rong,* [c] *oyyx,* [d] *xiongz@buaa.edu.cn,*
[e] *yyzhang@cse.buaa.edu.cn*

Molecular events normally have significant meanings since they describe important biological interactions or alternations such as binding of a protein. As a crucial step of biological event extraction, event trigger identification has attracted much attention and many methods have been proposed. Traditionally those methods can be categorised into rule-based approach and machine learning approach and machine learning based approaches have demonstrated its potential and outperformed rule-based approaches in many situations. However, machine learning based approaches still face several challenges among which one notable one is how to model semantic and syntactic information of different words and incorporate it into the prediction model. There exist many ways to model semantic and syntactic information, among which word embedding is an effective one. Therefore, in order to address this challenge, in this study, a word embedding assisted neural network prediction model is proposed to conduct event trigger identification. The experimental study on commonly used dataset has shown its potential. It is believed that this study could offer researchers insights into semantic-aware solutions for event trigger identification.

Keywords: Neural networks; Word embedding; Event trigger identification.