The 7th International Conference on Systems Biology (ISB 2013)

Local Organizer

Organizers



Sponsors



August 23-25, 2013
Huangshan, China

Schedule & Locations

Date	Time	Conference Room Taoyuan 桃源厅	Conference Room Jianjiang 渐江厅
August 22 Thursday	15:00-23:30	Registration (hotel lobby)	
	18:00-19:00	Supper	
	19:30-21:30		Board member meeting of ORSC-CSB
August 23 Friday	08:30-08:50	Opening Session	
	08:50-09:50	Plenary Session P1	
	09:50-10:20	Coffee break	
	10:20-12:20	Plenary Session P2	
	12:30-13:30	Lunch	
	14:00-16:00	Session A1: Highlight I	Session B1: Genomics
	16:00-16:20	Coffee break	
	16:20-18:00	Session A2: Highlight II	Session B2: Bioinformatics
	18:00-20:00	Welcome Reception	
August 24 Saturday	08:30-10:10	Session A3: Complex diseases I	Session B3: Proteomics I
	10:10-10:30	Coffee break	
	10:30-12:10	Session A4: Complex diseases II	Session B4: Proteomics II
	12:30-13:30	Lunch	
	14:00-15:40	Session A5: Network Biology I	Session B5: Systems Biology I
	15:40-16:00	Coffee break	
	16:00-17:40	Session A6: Network Biology II	Session B6: Systems Biology II
	18:00-20:00	Banquet	
August 25 Sunday	08:00-18:00	One day tour to Huangshan (Yellow Mountain). Departure at 8:00 from lobby.	

ISB2013 Program

August 23-25, Huangshan, China

*The program subjects to revision based on further information and Ad Hoc presentation requests.

August 22 (Thursday) Registration

15:00-22:00	Registration, check in and pick up registration package at hotel lobby.
18:00-19:00	Supper
19:30-21:30	Board member Meeting for ORSC-CSB (Conference Room Jianjiang)

August 23 (Friday) Technical Sessions

08:00-08:30	Registration for late arrivals (hotel lobby)

Location	Conference Room Taoyuan 桃源厅
08:30-08:50	Opening Session Chair: Luonan Chen
08:50-09:50	Plenary Session P1 Chair: Luonan Chen
08:50-09:50	*Mining important agronomic trait genes by evolutionary genomics* **Wen Wang**, Kunming Institute of Zoology, Chinese Academy of Sciences
09:50-10:20	Coffee break
10:20-12:20	Plenary Session P2 Chair: Wen Wang
10:20-11:20	*Genome-Phenome Association Analysis under Complex Structures* **Eric Xing**, Carnegie Mellon University
11:20-12:20	*Hierarchical Multilabel Classification in Disease Diagnosis Using Public Gene Expression Data* **Haiyan Huang**, University of California, Berkeley

12:30-13:30	Lunch

Location	Conference Room Taoyuan 桃源厅	Conference Room Jianjiang 渐江厅
14:00-16:00	Session A1: Highlight I Chair: Junwen Wang	Session B1: Genomics Chair: Jiayan Wu
14:00-14:20	*#88: HaploShare: Identification of extended haplotypes shared by cases and evaluation against controls* Dingge Ying, Pak Chung Sham, David Keith Smith, Lu Zhang, Yu Lung Lau and **Wanling Yang**	*#18: Tissue Significances Tests on DNA Binding Sequence Motifs for Human Genes* Hua Yu and **Xiujun Gong**
14:20-14:40	*#81: The lysine modifications: data resources and computational analysis* Zexian Liu, Tianshun Gao and **Yu Xue**	*#31: Codon-Based Encoding for DNA Sequence Analysis* **Byeong-Soo Jeong** and Ho-Jin Choi

14:40-15:00	*#54: How do the evolutionary events of gene and protein domain contribute to the increasing biological complexity?* **Dong Yang**, Dingchen Li, Chong Ma, Ying Jiang and Fuchu He	*#33: GPU-Meta-Storms: Computing the similarities among massive microbial communities using GPU* Xiaoquan Su, **Xuetao Wang**, Jian Xu and Kang Ning
15:00-15:20	*#89: QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data* Qian Zhou, **Xiaoquan Su**, Anhui Wang, Jian Xu and Kang Ning	*#47: Reinitiation enhances reliable transcriptional responses in eukaryotes* **Bo Liu**, Zhangjiang Yuan, Kazuyuki Aihara and Luonan Chen
15:20-15:40	*#84: Stochastic modelling of biochemical systems with multi-step reactions and inference of model parameters* **Qianqian Wu**	*#8: Imputing missing values for Genetic Interaction Data* Yishu Wang, Lin Wang, Dejie Yang and **Minghua Deng**
15:40-16:00	*#86: Finite element simulation of ion channel systems: continuum model, numerical method, and software platform* Benzhuo Lu, **Bin Tu** and Shiyang Bai	*#66: A novel discretization method for processing digital gene expression profiles* **Jibin Qu**, Jinxia Zhang, Baogui Xie, Yong Wang, Xiang-Sun Zhang and Chenyang Huang
16:00-16:20	**Coffee break**	
16:20-18:00	**Session A2: Highlight II** **Chair: Shihua Zhang**	**Session B2: Bioinformatics** **Chair: Minghua Deng**
16:20-16:40	*#87: GWAS3D: detecting human regulatory variants by integrative analysis of genome wide associations, chromosome interactions, and histone modifications* Mulin Jun Li, Lily Yan Wang, Zhengyuan Xia, Pak Chung Sham and **Junwen Wang**	*#55: Gene Ontology Based Housekeeping Gene Selection for RNA-seq Normalization* Yu-Lun Lu, Chi-Pong Sio, Chien-Ming Chen, Chih-Kai Hsu, Guan-Chung Wu and **Tun-Wen Pai**
16:40-17:00	*#82: Two straws make a jewel: screening the big data for disease diagnosis signature* Guoqin Mai, Miaomiao Zhao, Youxi Luo and **Fengfeng Zhou**	*#6: DMS model Calibration Using Genetic Algorithm* **Bo Qu**, Albert Gabric and Jiaojiao Xi
17:00-17:20	*#83: Identifying the critical transition for complex diseases based on a small number of samples* **Rui Liu**, K. Aihara and L. Chen	*#3: A systematic study on GPCR prototypes: did they really evolve from prokaryotic genes?* Zaichao Zhang, **Jiayan Wu**, Yongbing Zhao, Zhewen Zhang and Jingfa Xiao
17:20-17:40	*#53: Identifying disease genes and module biomarkers by differential interactions* **Xiaoping Liu**, Zhi-Ping Liu, Xing-Ming Zhao and Luonan Chen	*#35: A new model: Accelerating Processing Speed in Pathway Research Based on GPU* Bo Liao, **Ting Yao** and Xiong Li
17:40-18:00	*#85: Discovery of cell-type specific regulatory elements in the human genome by differential chromatin modification analysis* **Chen Chen**, Shihua Zhang and Xiang-Sun Zhang	*#1: A novel P-Wave Detection algorithm in ECG Signal* **Li Yongting and Li Ran**

18:00-20:00	**Welcome Reception**

August 24 (Saturday) Technical Sessions

Location	Conference Room Taoyuan 桃源厅	Conference Room Jianjiang 渐江厅
08:30-10:10	**Session A3: Complex Diseases I** **Chair: Lei Li**	**Session B3: Proteomics I** **Chair: Kang Ning**
08:30-08:50	*#15: Network analysis reveals roles of inflammation factors in different phenotypes of kidney transplantation patients* **Duojiao Wu**, Xiaoping Liu, Zhiping Liu, Luonan Chen and Tongyu Zhu	*#14: Prediction of Enzyme Catalytic Sites on Protein Using a Graph Kernel Method* **Benaragama Sanjaka** and Changhui Yan
08:50-09:10	*#29: A Co-expression Modules Based Gene Selection for Cancer Recognition* Xinguo Lu, **Yong Deng** and Bo Liao	*#22: Bi-factor analysis based on noise-reduction (BFANR): A New Algorithm for searching coevolving amino acid sites in proteins* Juntao Liu, Xiaoyun Duan, Jianyang Sun, Guojun Li, Lushan Wang and **Bingqiang Liu**
09:10-09:30	*#34: A key network approach reveals new insight in Alzheimer's disease* **Jan Schlüsener**, Xiaomei Zhu, Hermann Schlüsener, Gao-Wei Wang and Ping Ao	*#24: Apoptosis proteins subcellular localization prediction based on the knowledge mining of amino acid index database* Zhuoxing Shi, **Yuhua Yao** and Bo Liao
09:30-09:50	*#40: A Novel HMM for Analyzing Chromosomal Aberrations in Heterogeneous Tumor Samples* Hong Xia, Yuanning Liu, Minghui Wang, Huanqing Feng and **Ao Li**	*#27: Prediction of hot spots in protein interfaces using extreme learning machine* Wen-Juan Zhang and **Lin Wang**
09:50-10:10	*#67: Multiclass Classification of Sarcomas using Pathway Based Feature Selection Method* **Jianlei Gu**, Yao Lu, Cong Liu and Hui Lu	*#28: Rank-based interolog mapping for predicting protein-protein interactions between genomes* Yu-Shu Lo, Chun-Chen Chen, Kai-Cheng Hsu and **Jinn-Moon Yang**
10:10-10:30	**Coffee break**	
10:30-12:10	**Session A4: Complex Diseases II** **Chair: Fengfeng Zhou**	**Session B4: Proteomics II** **Chair: Xing-Ming Zhao**
10:30-10:50	*#43: Temporal order of somatic mutations during tumorigenesis based on Markov chain model* **Hao Kang**, Tao Zeng and Luonan Chen	*#57: Predicting the non-compact conformation of amino acid sequence by particle swarm optimization* **Yuzhen Guo** and Yong Wang
10:50-11:10	*#45: Anti-cancer Effect of Aloe-Emodin on Breast Cancer Cell Line, MCF-7* **Indah Mohd Amin**, Siti H Sheikh Abdul Kadir, Nik Mohd Mazuan Nik Mohd Rosdy, Rosfaiizah Siran and Narimah Ah Hasani	*#36: Mining Literature of Protein Phosphorylation using Dependency Trees* Mang Wang, Hong Xia, Dongdong Sun, Huanqing Feng, **Minghui Wang** and Ao Li
11:10-11:30	*#50: On Knowledge Discovery for Pancreatic Cancer Using Inductive Logic Programming* **Yushan Qiu**, Wai-Ki Ching, Kazuaki Shimada, Nobuyoshi Hiraoka, Kensei Maeshiro, Kiyoko F Aoki-Kinoshita and Koh Furuta	*#32: Proteome Compression via Protein Domain Compositions* **Morihiro Hayashida**, Peiying Ruan and Tatsuya Akutsu
11:30-11:50	*#60: Module network based cross-tissue analysis of Type 1 diabetes mellitus* **Tao Zeng**, Chuan-Chao Zhang, Juan Liu and Luonan Chen	*#56: Electrostatics and Structural Analysis of DNA-binding Sites in SSBs and DSBs* **Wei Wang**, Juan Liu and Lida Zhu

11:50-12:10	*#7: Induction of Apoptosis Associated with ER Stress and TP53 in MCF-7 cells by the Nanoparticle [Gd@C82(OH)22]n: A Systems Biology Study* **Lin Wang**, Jie Meng, Weipeng Cao, Qizhai Li, Yuqing Qiu, Xingjie Liang, Baoyun Sun, Yuliang Zhao and Lei Li	*#52: Discriminating Native Protein-DNA Complexes From Decoys Using Spatial Specific Scoring Matrices* Wen Cheng and **Changhui Yan**

12:30-13:30	**Lunch**

Location	Conference Room Taoyuan 桃源厅	Conference Room Jianjiang 渐江厅
14:00-15:40	**Session A5: Network Biology I** **Chair: Yu Xue**	**Session B5: Systems Biology I** **Chair: Ruiqi Wang**
14:00-14:20	*#9: Mining Disease Associated Biomarker Networks from PubMed* **Zhong Huang**	*#5: Dynamical complexity in a predator-prey eco-epidemical system* **Min Su** and Zhen-Shan Lin
14:20-14:40	*#41: Two Programmed Replicative Lifespans of Saccharomyces cerevisiae Formed by Endogenous Molecular-Cellular Network* **Jie Hu**, Xiaomei Zhu, Xinan Wang, Ruoshi Yuan, Wei Zheng, Minjuan Xu and Ping Ao	*#16: Global stability of the SEIR epidemic model with infectivity in both latent period and infected period* **Yu Zhang** and Ze-Zhu Ren
14:40-15:00	*#58: Meta-Analysis on Gene Regulatory Networks Discovered by Pairwise Granger Causality* **Gary Tam**, Ys Hung and Chunqi Chang	*#44: Towards Kinetic Modeling of Metabolic Networks with Incomplete Parameters* **Wei Zheng**, Xiaomei Zhu, Yongcong Chen, Paohung Lin and Ping Ao
15:00-15:20	*#63: The Residue Interaction Network Analysis of Dronpa and a DNA clamp* **Guang Hu**, Wenying Yan, Jianhong Zhou and Bairong Shen	*#38: Dynamical behaviour of an anti-HBV infection therapy model with time-delayed immune response* **Xinjian Zhuo** and Yongmei Su
15:20-15:40	*#64: The construction of tissue specificity phosphorylation network.* **Yin-Ying Wang**, Chenglei Sun, Luonan Chen and Xing-Ming Zhao	*#51: A multi-scale approach for simulating time-delay biochemical reaction systems* **Yuanling Niu**, Kevin Burrage and Chengjian Zhang
15:40-16:00	**Coffee break**	
16:00-17:40	**Session A6: Network Biology II** **Chair: Yong Wang**	**Session B6: Systems Biology II** **Chair: Guang Hu**
16:00-16:20	*#30: Exploring the interaction patterns in seasonal marine microbial communities with network analysis* **Shao-Wu Zhang**, Ze-Gang Wei, Chen Zhou, Yu-Chen Zhang and Ting-He Zhang	*#42: Cell Commitment Motif Composed of progenitor-specific TF and Fate-Decision Motif* **Tongpeng Wang**, Peipei Zhou and Ruiqi Wang
16:20-16:40	*#13: Colored Petri Nets for Multiscale Systems Biology – Current Modeling and Analysis Capabilities in Snoopy* **Fei Liu**, Monika Heiner and Ming Yang	*#59: EdgeSVM: a method for identifying differentially correlated gene pairs as edge markers.* **Wanwei Zhang**, Tao Zeng and Luonan Chen
16:40-17:00	*#65: Inferring Gene Regulatory Networks from Integrative Omics Data via LASSO-type regularization methods* **Jing Qin**, Yao Hua Hu, Feng Xu and Junwen Wang	*#68: Detect taxonomy-specific pathway associations with environmental factors using metagenomic data* **Xue Tian**, Fuzhou Gong and Shihua Zhang
17:00-17:20	*#23: Antirheumatic effects of Tripterygium wilfordii Hook F in a network perspective* Haiyang Fang, Tinghong Yang, Yichuan Wang, Yang Ga, Yi Zhang, Weidong Zhang and **Jing Zhao**	*#62: Overshooting in biological systems modeled by Markov chains* **Chen Jia** and Minping Qian

17:20-17:40	*#61: Effective identification of essential proteins based on priori knowledge, network topology and gene expressions* **Min Li**, Ruiqing Zheng, Hanhui Zhang, Jianxin Wang and Yi Pan	*#2: Cholesterol Oxidase from Staphylococcus epidermidis and its Application in Determination of Cholesterol Content of Egg Yolk* **Hamed El-Shora** and Mai Taha

18:00-20:00	**Banquet**

August 25 (Sunday) Tour

08:00-18:00	One day tour to Yellow Mountain (One-day-tour-tickets are needed)
08:00-09:00	Gathering in front of the hotel and depart to Yellow Mountain
09:00-10:00	Take park bus and cable car to the top of mountain
12:30-13:30	Lunch at the park restaurant
16:00-17:00	Down to the park gate
17:00-18:00	Take bus back to the hotel
18:00-19:00	**Dinner at the hotel**

Note: Those who leave Huangshan on August 25 may check out the hotel before 8:00 and deposit luggage at the reception desk. Do not leave your luggage in bus because the buses may be used by other passengers during the day.

Plenary Sessions

Mining important agronomic trait genes by evolutionary genomics

Wen Wang

Stake Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China

Domestic organisms evolved distinct traits rapidly under artificial selection. Traditionally mapping of quantitative trait loci (QTL) has been the main approach to identify agronomic genes. However, QTL mapping usually is time consuming and labor demanding. Conceivably all agronomic trait related genes should have been selected by human, and thus detecting selection signals in domestic species' genomes or identifying tag SNPs in improved breeds could be a highly efficient approach to mine agronomic trait genes. Based on this hypothesis, we comprehensively compared cultivated and wild rice through population genomics, and identified thousands of genes located in selected regions of cultivated rice, which includes those reported domestication genes. Using this large quantity of cultivated and wild rice genome variation dataset as control, we further identified those elite variety tag functional SNPs (ETASs) in 6 elite rice varieties. Guided by this preliminary ETAS analysis, we thoroughly characterized one protein-altering ETAS in the 9-cis-epoxycarotenoid dioxygenase gene (Nced) of the upland rice variety, IRAT104. This ETAS displayed a drastic frequency difference between upland and irrigated rice, and a selective sweep was observed around it. Functional analysis showed that in upland rice, this ETAS is associated with significantly higher ABA levels and denser lateral roots, suggesting its association with upland rice suitability. These results indicate evolutionary genomics can be a efficient strategy to mine agronomically important genes.

Hierarchical Multilabel Classification in Disease Diagnosis Using Public Gene Expression Data

Haiyan Huang

Department of Statistics

Interdepartmental Group in Biostatistics

Graduate Group in Computational and Genomic Biology

University of California, Berkeley

http://www.stat.berkeley.edu/~hhuang

The rapid accumulation of gene expression data has offered unprecedented opportunities to study human diseases. The National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) is currently the largest database that systematically documents the genome-wide molecular basis of diseases. This talk introduces an effort to turn the NCBI GEO expression repository into an automated disease diagnosis database, such that a query gene expression profile can be assigned to one or multiple disease concepts. As hierarchical multi-label classification (HMC) is a natural formulation of a disease diagnosis question, this talk also discusses some statistical issues involved in HMC.

Genome-Phenome Association Analysis under Complex Structures

Eric Xing

School of Computer Science at Carnegie Mellon University.

Genome-wide association (GWA) mapping have recently become a popular approach for identifying genetic loci that are responsible for increased disease susceptibility. In the presence of rich side-information about the structure of the data, such as population structures revealed by ancestral inference and heredity analysis, genome structures exposed in linkage disequilibrium studies, trait structures captured by expression networks or phenotype clusters, to name a few, traditional approaches such as p-value based pair-wise

SNP-trait association tests, PCA, or lasso have difficulties in admitting (the constraints induced by) and exploiting (the benefits offered by) such information. In this talk, I will present a class of new models, algorithms, and theories that go beyond the traditional approach and enable effective use of structural information for GWA mapping in large-scale and high-dimensional setting given whole genome and phenome data. This approach builds on the sparse structured regression method in statistics, enjoying strong statistical guarantee and scalability, and can be flexibly configured to handle different structural information. I will demonstrate application of this approach to a number of complex GWA scenarios, including associations to trait networks or cluster-tree, to dynamic traits, under admixed populations, and with epistatic effects.

Highlight Sessions

#53

Identifying disease genes and module biomarkers by differential interactions

Xiaoping Liu

Institute of Industrial Science, University of Tokyo

A complex disease is generally caused by the mutation of multiple genes or by the dysfunction of multiple biological processes. Systematic identification of causal disease genes and module biomarkers can provide insights into the mechanisms underlying complex diseases, and help develop efficient therapies or effective drugs. In this paper, we present a novel approach to predict disease genes and identify dysfunctional networks or modules, based on the analysis of differential interactions between disease and control samples, in contrast to the analysis of differential gene or protein expressions widely adopted in existing methods. As an example, we applied our method to the study of three-stage microarray data for gastric cancer. We identified network modules or module biomarkers that include a set of genes related to gastric cancer, implying the predictive power of our method. The results on holdout validation data sets show that our identified module can serve as an effective module biomarker for accurately detecting or diagnosing gastric cancer, thereby validating the efficiency of our method. We proposed a new approach to detect module biomarkers for diseases, and the results on gastric cancer demonstrated that the differential interactions are useful to detect dysfunctional modules in the molecular interaction network, which in turn can be used as robust module biomarkers.

#54

How do the evolutionary events of gene and protein domain contribute to the increasing biological complexity?

Dong Yang

State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine

Some evolutionary events, including new domain emergence, domain shuffling and gene duplication, substantially contribute to the increasing of genome complexity. We focused on the question how these genomic complexity information are utilized at given physiological states, contributing to the increasing of phenotypic complexity during evolution. Based on our previous studies1,2, we further comprehensively analyzed the relationship among the genomic complexity factors and explored the composition characteristics of certain-state proteomes (CSPs, i.e., all the proteins expressed at certain physiological states) at both qualitative and quantitative levels, taking the characters of protein domain number, age, distribution frequency and gene duplication history into protein classification. First, at genome level, we found that, compared to singletons, paralogs tend to contain more types of protein domain, especially the younger domains,

and the domains shared by more than one gene and family. These results revealed that the bio-complexity-causing evolutionary events at both gene and protein domain levels interplay and facilitate each other. Second, at proteome level, paralogs, multi-domain proteins, proteins only containing older domains, or containing gene- or family-specific domains are significantly over-represented in CSPs compared to GEP (all the proteins encoded by the genome), and vice versa. Interestingly, younger domain-containing paralogs and the proteins containing both younger domain and gene- or family-specific domains are significantly over-represented. These results indicate, at given physiological states, biological complexity is more dependent on gene duplicates, diverse domain organization than the new types of domain and domain sharing among genes and families. In the view of protein abundance, we found that bio-complexity-related proteins, including paralogs, multi-domain proteins, proteins containing younger domains, and the sharing domains, all tend to be with lower abundance, and vice versa. Our work provides new insights into the mechanisms involved in the increasing of genome complexity and its realization at molecular phenotypic level.

#81

The lysine modifications: data resources and computational analysis

Yu Xue

Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

The lysine residues in proteins are “hotspot” sites for a number of post-translational modifications (PTMs), such as well-known ubiquitin and ubiquitin-like conjugation, acetylation, and newly identified succinylation and crotonylation. First, we developed a family-based database for ubiquitin and ubiquitin-like conjugation, which is one of the most important PTMs responsible for regulating a variety of cellular processes, through a similar E1 (ubiquitin-activating enzyme)–E2 (ubiquitin-conjugating enzyme)–E3 (ubiquitinprotein ligase) enzyme thioester cascade. From the scientific literature, 26 E1s, 105 E2s, 1,003 E3s and 148 deubiquitination enzymes (DUBs) were collected and classified into 1, 3, 19 and 7 families, respectively. To computationally characterize potential enzymes in eukaryotes, we constructed 1, 1, 15 and 6 hidden Markov model (HMM) profiles for E1s, E2s, E3s and DUBs at the family level, separately. Moreover, the ortholog searches were conducted for E3 and DUB families without HMM profiles. Then the UUCD database was developed with 738 E1s, 2,937 E2s, 46,631 E3s and 6,647 DUBs of 70 eukaryotic species. The detailed annotations and classifications were also provided. More recently, we collected 66,826 known sites in 18,522 unique proteins for 11 types of lysine modifications. Based on the data resource, a human acetylation network was constructed, while the site-specific predictors for these modifications were being constructed. We anticipate our data resources and analysis can be helpful for better understanding the lysine modifications.

Reference

1. Gao T, Liu Z, Wang Y, Cheng H, Yang Q, Guo A, Ren J, Xue Y. (2013) UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation. Nucleic Acids Res., 41:D445-51

#82

Two straws make a jewel: screening the big data for disease diagnosis signature

Fengfeng Zhou

Shenzhen Institutes of Advanced Technology, and Key Laboratory of Health Informatics, Chinese Academy of Sciences

As the rapid innovation and development of new high-throughput biomedical data production technologies, massive amount of biomedical data

are generated at an accelerated speed. Computer scientists and statisticians focus on optimizing the supervised or unsupervised clustering of the biomedical data, whereas the OMIC biologists try to screen for biologically meaningful knowledge from the big health data, for the purpose of guiding the next-step low-throughput experimental validations. Sometimes the computational scientists generate a clustering model with a group of features, that have little associations with the available biological knowledge, or whose number is too large to be experimentally verified considering the limited resources. This work tries to screen for a feature list, that is small in the feature number and has strong association with the available knowledge. We believe that this is essential for both the computational scientists and biomedical researchers.

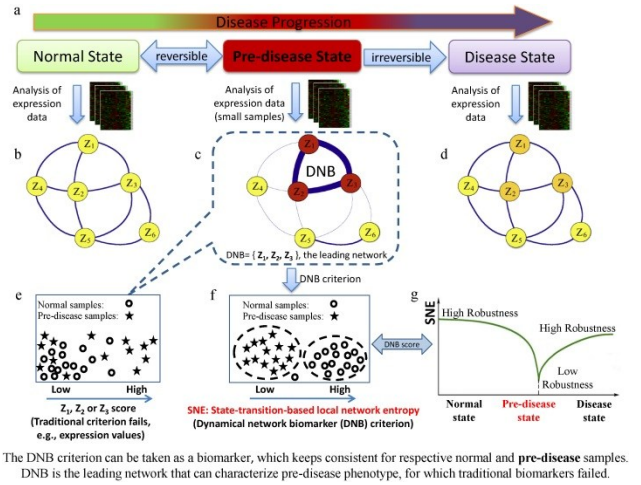
#83

Identifying the critical transition for complex diseases based on a small number of samples

Rui Liu

Department of Mathematics, South China University of Technology

Detecting the sudden deterioration or the critical transition of a complex disease is very hard, especially when there are only small samples available. This is not only because during the progression of a complex disease, the pre-disease state just before the critical transition is actually a limit of the normal state which results that the state of the system may show little apparent change before the critical tipping point is reached, but there is a great obstacle of detecting the pre-disease state, i.e., the small sample problem arising in clinical early diagnosis. In this talk, we will introduce both a novel computational approach and a composite scoring index for detecting the pre-disease state from only a small number of high-throughput samples. This approach is theoretically based on the dynamical network biomarker (DNB) theorem, and achieves the small sample (or even only a single sample) based early diagnosis which was validated by numerical simulation as well as experimental data. We will also give some comparisons between traditional biomarkers and DNB.



The DNB criterion can be taken as a biomarker, which keeps consistent for respective normal and **pre-disease** samples. DNB is the leading network that can characterize pre-disease phenotype, for which traditional biomarkers failed.

Figure 1: A schematic illustration of dynamical features for disease progression from a normal state to a disease state through a pre-disease state. (a) Three states during progression of a disease. (b) The normal state is a steady state, where the system generally has high resilience and robustness to perturbations. (c) The pre-disease state is defined as a limit of the normal state and situated before the imminent phase transition point is reached. At this stage, the system is with low resilience and robustness even to small perturbations but still reversible to the normal state when appropriately interfered. {Z1, Z2, Z3} is the dominant group or the DNB. (d) The disease state is the other steady state, at which the system is usually irreversible to the normal state due to its high resilience and robustness. (e) Traditional biomarkers failed to distinguish the pre-disease samples from normal samples. (f) SNE (DNB score) is effective in distinguishing the pre-disease samples. (g) The SNE is the conditional entropy of the previous state, which does not change significantly during the normal state but it drops sharply during the pre-disease state. By contrast, the SNE converges to another steady value during a disease state. The SNE drops drastically whenever the system approaches a critical transition point, so it can provide an effective early-warning signal for identifying the pre-disease state and the leading network that makes the first move toward a disease state.

References

[1]. Chen L, Liu R, Liu Z, Li M, Aihara K. Detecting

early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Scientific Reports, 2012; 2: 342; DOI: 10.1038/srep00342.
[2]. Liu R, Li M, Liu Z, Wu J, Chen L, Aihara K. Identifying critical transitions and their leading networks for complex diseases. Scientific Reports, 2013; 2: 813; DOI: 10.1038/srep00813.
[3]. Liu R, Aihara K, Chen L. Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes. Quantitative Biology, 2013; DOI:10.1007/s40484-013-0008-0.
[4]. Liu R, Wang X, Aihara K, Chen L. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. Medicinal Research Reviews, 2013; DOI: 10.1002/med.21293.

#84

Stochastic modelling of biochemical systems with multi-step reactions and inference of model parameters

Qianqian Wu
School of Mathematical Sciences
Monash University

The advances in systems biology have raised the importance of mathematical modelling for studying complex biological systems. One of the fundamental problems in systems biology is how to design simplified models for describing the dynamics of complex real-life biochemical systems. In particular, obtaining accurate description of chemical events with multi-step chemical reactions is still regarded as a challenge in both chemistry and biophysics, though a number of modelling approaches have been attempted to tackle this issue in recent years. During my PhD studies, I have proposed a two-variable model to describe chemical events with multi-step chemical reactions. The innovation of this research is the introduction of a new concept that represents the location of molecules in the multi-step reactions, and we use it as the second indicator of the system dynamics. I have also proposed a simulation algorithm to compute the probability of firing of the last step reaction in the multi-step event and this probability function is further evaluated using a deterministic model with ordinary differential equations and a stochastic model with a stochastic simulation algorithm. The efficiency of the proposed two-step model is evaluated through a simplified mRNA degradation process based on the experimentally measured data sets. Numerical results suggested that the proposed new two-step model with two variables could generate simulations that match experimental data very well.

Another fundamental problem in systems biology is the inference of network structure and unknown model parameters. This problem is confronting because it is difficult to estimate these parameters in stochastic regulatory networks according to a limited amount of experimental data. To address this issue, I have proposed a new method to estimate parameters in stochastic models using the frequency distribution in the framework of the approximate Bayesian computation (ABC). Two stochastic models are used to demonstrate the efficiency and effectiveness of the proposed method. Simulation results suggest that the usage of frequency distribution improves the accuracy of the estimates. In addition, I have examined the approaches to measure the accuracy of simulations. When the error is measured over every pair of two consecutive observations, the estimated parameters have better accuracy than those obtained by measuring the error of the simulation over the entire observation time period.

References

[1] Wu Q, Smith-Miles K and Tian T., *A two-variable model for stochastic modelling of chemical events with multi-step reactions*, Proceedings of 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM2012), 270-275, IEEE Press, 2012.
[2] Wu Q., Smith-Miles K., Zhou T. and Tian T., *Stochastic modelling of chemical events with multi-step reactions using a simplified two-variable model*, to appear in BMC Systems Biology.
[3] Wang J.*, Wu Q*. and Tian T., *An integrated*

approach to infer dynamic protein-gene interactions – a case study of human P53 protein, submitted for publication. (*The authors make the same contribution).
[4] Wu Q, Smith-Miles K and Tian T., *Approximate Bayesian Computation using the simulated likelihood density for estimating rate constants in biochemical reaction systems*, submitted for publication.

#85

Discovery of cell-type specific regulatory elements in the human genome by differential chromatin modification analysis

Chen Chen

National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Chromatin modifications have been comprehensively illustrated to play important roles in gene regulation and cell diversity in recent years. Given the rapid accumulation of genome-wide chromatin modification maps across multiple cell types, there is an urgent need for computational methods to analyze multiple maps to reveal combinatorial modification patterns and define functional DNA elements, especially those are specific to cell types or tissues. In this current study, we developed a computational method using differential chromatin modification analysis (dCMA) to identify cell-type-specific genomic regions with distinctive chromatin modifications. We then apply this method to a public data set with modification profiles of nine marks for nine cell types to evaluate its effectiveness. We found cell-type specific elements unique to each cell type investigated. These unique features show significant cell-type-specific biological relevance and tend to be located within functional regulatory elements. These results demonstrate the power of a differential comparative epigenomic strategy in deciphering the human genome and characterizing cell specificity.

References

[1] Chen Chen*, Shihua Zhang*#, Xiang-Sun Zhang (*Co-first authors; #Corresponding author). Discovery of cell-type specific regulatory elements in the human genome by differential chromatin modification analysis. Nucleic Acids Research. (2013), in press.

#86

Finite element simulation of ion channel systems: continuum model, numerical method, and software platform

Bin Tu

Academy of Mathematics and Systems Science, CAS

As it is hard to apply all atomic model to simulate the whole process of ion permeation in ion channel, we use continuum electrodiffusion description for ion flow in the channel system. Electrodiffusion process exists in many apparently different physical objects such as electrolyte cell, nanofluidic device, charged porous media, and ion channel in biology. Real 3D ion channel is particularly difficult to simulate due to the multiscale nature of the transport process, the complex geometry/boundary of the channel protein system, and the singular charge distribution inside the channel protein(s). For this reason, there are so far only a very few softwares publicly available in this important area of biology. We will show our recent relevant works and plan to build up such a platform. In the first part, we'll talk about the continuum models and numerical works. They include the Poisson-Boltzmann equation, the Poisson-Nernst-Planck equations and their improved forms, and some efficient algorithms we developed for the solution of these equations. In the second part, we will describe the molecular meshing problem which is essential for finite/boundary element modelings. We recently developed a novel and robust mesh generation tool TMSmesh that can handle complex and arbitrarily large biomolecular

system. In the third part, I will give a brief introduction to an undergoing project of designing a visualization system, MMV, to facilitate researches in this area. Finally, we will show applications using our parallel finite element solver to compute properties such as current-voltage characteristics (curves) to a few channel systems. The results agree well with those obtained with Brownian Dyanmics simulations and experiments.

#87

GWAS3D: detecting human regulatory variants by integrative analysis of genome wide associations, chromosome interactions, and histone modifications

Junwen Wang

The University of Hong Kong

Interpreting the genetic variants located in the regulatory regions, such as enhancer and promoter, is an indispensable step to understand molecular mechanism of complex traits. Recent studies show that SNPs detected by genome wide association study (GWAS) are significantly enriched in the regulatory regions. Therefore, detecting, annotating and prioritizing of genetic variants affecting gene regulation are critical to our understanding of genotype-phenotype relationships. Here, we developed a web server GWAS3D to systematically analyze the genetics variants that could affect regulatory elements, by integrating annotations from cell type specific chromatin states, epigenetic modifications, sequence motif, and cross species conservation. The regulatory elements are inferred from the genome-wide chromosome interaction data, chromatin marks in 16 different cell types, and 73 regulatory factors motifs from the ENCODE project. Furthermore, we used these function elements, as well as risk haplotype, binding affinity, conservation, and the *P*-values reported from the original GWA study to reprioritize the genetic variants. Using studies from low-density lipoprotein cholesterol (LDL-C), we demonstrated that our reprioritizing approach is effective and cell type specific. In conclusion, GWAS3D provides a comprehensive annotation and visualization tool to help users interpreting their results. The web server is freely available at http://jjwanglab.org/gwas3d.

#88

HaploShare: Identification of extended haplotypes shared by cases and evaluation against controls

Wanling Yang

The University of Hong Kong

Recent founder mutations may play important roles in complex diseases and Mendelian disorders. Detecting shared haplotypes that are identical by descent (IBD) could facilitate discovery of these mutations. Several programs address this, but are usually limited to detecting pair-wise shared haplotypes and not providing a comparison of cases and controls. Here we present a novel algorithm and software package (HaploShare) that detects extended haplotypes that are shared by multiple individuals, and allows comparisons between cases and controls. A catalog of haplotypes is first generated from healthy controls from the same population and used for phasing genotypes in cases. By accounting for all possible haplotype pairs that could explain the genotypes for each individual in a given haplotype block and possible transitions between blocks, the effect of phase uncertainty on detection power is minimized. In cases, haplotypes shared by pairs are identified and used to detect sharing of these haplotypes by different pairs. A likelihood ratio of a shared haplotype being due to IBD or chance is estimated for each extended haplotype. Controls are used similarly through many rounds of simulations to obtain an empirical null distribution of the largest likelihood ratios of shared haplotypes, to give statistical estimates of shared haplotypes detected in cases that may be associated with an underlying disease. Extensive testing on simulated and real cases demonstrated significant improvements in detection power and

reduction of false positive rate by HaploShare relative to other available programs.

#89

QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data

Xiaoquan Su

Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences

Next-generation sequencing (NGS) technologies have been widely used in life sciences. However, several kinds of sequencing artifacts, including low-quality reads and contaminating reads, were found to be quite common in raw sequencing data, which compromise downstream analysis. Therefore, quality control (QC) is essential for raw NGS data. However, although a few NGS data quality control tools are publicly available, there are two limitations: First, the processing speed could not cope with the rapid increase of large data volume. Second, with respect to removing the contaminating
reads, none of them could identify contaminating sources de novo, and they rely heavily on prior information of the contaminating species, which is usually not available in advance. Here we report QC-Chain, a fast, accurate and holistic NGS
data quality-control method. The tool synergeticly comprised of user-friendly tools for (1) quality assessment and trimming of raw reads using Parallel-QC, a fast read processing tool; (2) identification, quantification and filtration of unknown contamination to get high-quality clean reads. It was optimized based on parallel computation, so the processing speed is significantly higher than other QC methods. Experiments on simulated and real NGS data have shown that reads with low sequencing quality could be identified and filtered. Possible contaminating sources could be identified and quantified de novo, accurately and quickly. Comparison between raw reads and processed reads also showed that subsequent analyses (genome assembly, gene prediction, gene annotation, etc.) results based on processed reads improved significantly in completeness and accuracy. As regard to processing speed, QC-Chain achieves 7–8 time speed-up based on parallel computation as compared to traditional methods. Therefore, QC-Chain is a fast and useful quality control tool for read quality process and de novo contamination filtration of NGS reads, which could significantly facilitate downstream analysis. QC-Chain is publicly available at: http://www.computationalbioenergy.org/qc-chain.html.

Parallel Sessions

#1

A novel P-Wave Detection algorithm in ECG Signal

Li Yongting and Li Ran

Inner Mongolia University of Technology

Electrocardiogram (ECG) is characterized by a recurrent wave sequence of P and R-wave associated with each beat. The automatic detection of ECG waves is important to cardiac disease diagnosis. Lots of wave detection schemes and algorithms have been proposed, however, the majority of those are about R-wave detection. Although P-wave is also very important in ECG auto-diagnosis, there are a little kinds of method about P-wave detection. In this paper, a novel P-wave detection algorithm based on prony analysis is proposed. The proposed method applies prony analysis technique to separate DCT coefficients of ECG data. The method gets good result. Data in experiment are from MIT Arrhythmia Database. The reliability of this algorithm in accurately detecting P-wave is 97.8%.

#2

Cholesterol Oxidase from Staphylococcus epidermidis and its Application in Determination of Cholesterol Content of Egg Yolk

Hamed El-Shora
Mansoura University

Cholesterol oxidase (CO EC 1.1.3.6) activity from Staphylococcus epidermidis was investigated. Meal and yeast extract were the best nitrogen source whereas glycerol and galactose were the best carbon sources for enzyme production. The enzyme was induced by spermine, spermidine, putrescine, benzyl adenine and gibberellic acid. CO was purified with specific activity of 62 Umg-1 protein using 80 % ammonium sulfate, DEAE-cellulose and Sephadex G200. The Vmax and Km values were 23.8 Umg-1 protein and 0.26 mM, respectively. CO was activated by reduced glutathione (GSH) and dithiothreitol (DTT). Ca, Mg and Mn were activators whereas Al, Cu. Zn and Ba were inhibitors. K did not affect the enzyme activity. Modification of CO with citraconic anhydride and phthalic anhydride retained appreciable enzyme activity at 60 ○C. Polyols such as glycerol, sorbitol, mannitol and xylitol as well as glycol chitosan offered good stability at 60 ○C. Collagen, malto dextran, bovine serum albumin (BSA) and proline supported enzyme stability at 70 ○C. The results indicate that lysyl, tyrosyl, and tryptophenyl residues are taking part in enzyme catalysis. Urea protected lysyl and tryptophanyl against modification by dansyl chloride and N-bromosuccinimide but did not protect tyrosyl residue against modification by N-acetylimidazole. The solubilization was better than saponification in determination of cholesterol content of egg yolk by the purified enzyme.

#3

A systematic study on GPCR prototypes: did they really evolve from prokaryotic genes?

Jiayan Wu
Beijing Institute of Genomics, Chinese Academy of Sciences

G-protein couple receptors (GPCR) only represent in eukaryotes and they are essential protein superfamilies in cellular signaling. Numerous identification methods and classification systems for GPCR have been employed. Several deductions have already presumed their evolution while the ancestor of GPCR is seldom further studied and here, we investigated structure variances and domain distributions of each GPCR subclasses in different eukaryotes. Our result indicates that only all metabotropic glutamate receptor family exists in most ancient eukaryotes like protists. Phylogenetic analysis shows 7-transmembrane of metabotropic glutamate receptor family is closer to bacteriorhodopsin than rhodopsin-like and secretin receptor family. We presume one of metabotropic glutamate receptor subclasses is ancestor of GPCR, which possibly evolved from a compound of bacteriorhodopsin and periplasmic binding proteins. Our result also demonstrates that each type of GPCR subclasses has its own specific motifs and identical structures, which would help us for future studies on GPCR orphans' prediction and classification.

#5

Dynamical complexity in a predator-prey eco-epidemical system

Min Su
School of Mathematics, Hefei University of Technology

Effects of the relationship between species and environment on an eco-epidemiological system are investigated. And periodic variation is also added to the disharmony parameter. The dynamic behaviors of the system are simulated numerically. A variety of complex population dynamics including stable state, periodic resonance and chaos are obtained. The most important result is that harmony relationship between prey species and environment is benefit for

the controlling of disease. Our result reinforces the conjecture that the relationship between species and environment is crucial to transmission of infectious disease.

#6

DMS model Calibration Using Genetic Algorithm

Bo Qu

Nantong University

Recent researchers suggested Dimethyl sulphide (DMS) flux emission in Arctic Ocean plays an important role for the global warming. A Genetic Algorithm (GA) method was developed and used in calibrating the DMS model parameters in Barents Sea in Arctic Ocean (70-80N, 30-35E). Two-step GA calibrations were performed. First step was to calibrate the most sensitive parameters based on Chlorophyll_a (CHL) satellite SeaWIFS 8-day data. DMS model was then calibrated for another 5 most sensitive parameters. The best fitness was as good as -0.76 for CHL calibration in 1998-2002. The GA proved an efficient tool in the multiple-parameter calibration task. Model simulations indicate significant inter-annual variation in the CHL amount leading to significant inter-annual variability in the observed and modeled production of DMS and DMS flux in the study region in Arctic Ocean.

#7

Induction of Apoptosis Associated with ER Stress and TP53 in MCF-7 cells by the Nanoparticle $[Gd@C_{82}(OH)_{22}]_n$: A Systems Biology Study

Lin Wang

Academy of Mathematics and Systems Science

The nanoparticle gadolinium endohedral metallofullerenol [Gd@C82(OH)22]n is a new candidate for cancer treatment with low toxicity. However, its anti-cancer mechanisms remain mostly unknown. In this study, we took a systems biology view of the gene expression profiles of human breast cancer cells (MCF-7) and human umbilical vein endothelial cells (ECV304) treated with and without [Gd@C82(OH)22]n, respectively, measured by the Agilent Gene Chip G4112F. To properly analyze these data, we modified a suit of statistical methods we developed. For the first time we applied the sub-sub normalization to Agilent two-color microarrays. Instead of a simple linear regression, we proposed to use an one-knot SPLINE model in the sub-sub normalization to account for nonlinear spatial effects. The parameters estimated by LTS- and S-estimators show similar normalization results. We made several kinds of inferences by integrating the expression profiles with the bioinformatic knowledge in KEGG pathways, Gene Ontology, JASPAR, and TRANSFAC. In the transcriptional inference, we modified the BASE method so that a transcription factor's up-regulation and down-regulation activities are inferred separately. Overall, [Gd@C82(OH)22]n induces more differentiation in MCF-7 cells than in ECV304 cells, particularly in the reduction of protein processing such as protein glucosylation, folding, targeting, exporting, and transporting. Among the KEGG pathways, the ErbB signaling pathway is up-regulated, whereas protein processing in endoplasmic reticulum (ER) is down-regulated. CHOP, a key pro-apoptotic gene downstream of the ER stress pathway, increases to nine folds in MCF-7 cells after treatment. These findings indicate that ER stress may be one important factor that induces apoptosis in MCF-7 cells after [Gd@C82(OH)22]n treatment. The expression profiles of genes associated with ER stress and apoptosis are statistically consistent with other profiles reported in the literature, such as those of HEK293T and MCF-7 cells induced by the miR-23a~27a~24-2 cluster. Furthermore, one of the inferred regulatory mechanisms comprises the apoptosis network centered around TP53, whose effective regulation of apoptosis is somehow reestablished after [Gd@C82(OH)22]n treatment. These results elucidate the application and

development of [Gd@C82(OH)22]n and other fullerene derivates.

#8

Imputing missing values for Genetic Interaction Data

Minghua Deng
Peking University

Background: Epistatic Miniarray Profiles (EMAP) enables the research of genetic interaction as an important method to construct large-scale genetic interaction network. However, high proportion of missing values frequently poses problems in the EMAP data analysis since they can hinder useful information of the datasets. While there have been some imputation approaches available to EMAP data, we adopted an improved SVD modeling procedure to impute the missing values in EMAP data, which results in the highest accuracy rate comparing with existent methods.

Results: The improved imputation method adopting an effective soft-threshold to SVD approach which has been showed to be the best method to impute the genetic interaction data comparing with a number of advanced imputation methods. Furthmore, the imputation can also improve the result of clustering on EMAP dataset, where more meaningful modules, known pathways and protein complexes could be detected after apply our imputation method on EMAP dataset.

Conclusion: The results demonstrate that, while missing data are complicating unavoidably in EMAP data, we can complete the original dataset by an efficient method to detect genetic information more exactly.

#9

Mining Disease Associated Biomarker Networks from PubMed

Zhong Huang
Drexel University

Disease related biomarker discovery is the critical step to realize the future personalized medicine and has been an important research area. With exponential growing of biomedical knowledge deposited in PubMed database, it is now an essential step to mine PubMed for biomarker-disease associations to support the laboratory research and clinical validation. We constructed list of human diseases that are most frequently associated with biomarker in literatures by text mining. Top ranked neurology diseases were then used to extract associated genes from PubMed using context sensitive information retrieval methods. Associated genes were then integrated into pathways and subject to network biomarker analysis. Our approach identifies both known and potential biomarkers for 3 neurodegenerative diseases.

#13

Colored Petri Nets for Multiscale Systems Biology – Current Modeling and Analysis Capabilities in Snoopy

Fei Liu
Harbin Institute of Technology

Systems biology has introduced a number of multiscale challenges, which, however, can be tackled by colored Petri nets, but not by traditional approaches like ordinary differential equations or Petri nets. In this paper, after a brief covering of multiscale challenges of systems biology, we report the modeling and analysis capabilities of colored Petri nets, which Snoopy by now offers, and describe how these capabilities are used to address those multiscale challenges. In doing so, we aim to attract more researchers to use the powerful capabilities of colored Petri nets to model and analyze multiscale biological systems.

#14

Prediction of Enzyme Catalytic Sites on Protein Using a Graph Kernel Method

Benaragama Sanjaka
North Dakota State University

Structural Genomics projects are producing structural data for proteins at an unprecedented speed. The functions of many of these protein structures are still unknown. To decipher the functions of these proteins and identify functional sites on their structures have become an urgent task. In this study, we developed an innovative graph method to represent protein surface based on how amino acid residues contact with each other. Then, we implemented a shortest-path graph kernel method to measure the similarities between graphs. We tried three variants of the nearest neighbor method to predict enzyme catalytic sites using the similarity measurement given by the shortest-path graph kernel. The prediction methods were evaluated using the leave-one-out cross validation. The methods achieved accuracy as high as 77.1%. We sorted all examples in the order of decreasing prediction scores. The results revealed that the positive examples (catalytic site residues) were associated with higher prediction scores and they were enriched in the region of top 10 percentile. Our results showed that the proposed methods were able to capture the structural similarity between enzyme catalytic sites and would provide a useful tool for catalytic site prediction.

#15

Network analysis reveals roles of inflammation factors in different phenotypes of kidney transplantation patients

Duojiao Wu
Zhongshan Hospital，Fudan University

Background: Inflammation induced by immunologic rejection is an important impediment to the long-term renal allograft survival.Systems-level characterization of inflammation in kidney transplantation remains incomplete. Stratifying kidney transplantation patients based on phenotypes, the present study aimed at identifying the role of inflammation proteins in disease progress and assessing potential biomarkers for allograft monitoring.

Methods：Kidney transplantation patients with different phenotypes were collected: stable renal function (ST), impaired renal function (Injury), acute rejection (AR) ,chronic rejection(CR). We stratified the patients into 3 levels according to their symptom and pathogenesis. Serum protein concentration was measured by using quantitative protein array. All differentially expressed proteins were analyzed by PPI to highlight proteins interactions in different levels of transplantation patients. By identifying level-related proteins, using Support Vector Machine (SVM) regression, subsequently with the analyzing ROC curve and the area under ROC curve (AUC), we evaluate the classification efficiency of these biomarkers based on leave-one-out validation. The level-related proteins were also annotated by KEGG enrichment analysis .

Results：On the hypothesis of common proteins and their up or down-regulation might induce disease progress, we obtained 12 common proteins and 11 level specific proteins among the phenotype-related PPI network. The level specific proteins could be potential biomarkers with diagnostic value. The classifying potency of 11 level specific proteins including IL1R1, IL16, TIMP1, CSF3, CXCL9, IL11, CXCL13, TNFRSF1B, CCL24,CCL1, IL6R were better than the performance of using all 40 proteins. The common proteins were annotated for KEGG enrichment: 1 Cytokine-cytokine receptor interaction; 2 Hematopoietic cell lineage; 3 Jak-STAT signaling pathway; 4 Allograft rejection; 5 T cell receptor signaling pathway.

#16

Global stability of the SEIR epidemic model with infectivity in both latent period and infected period

Yu Zhang
Harbin University of Commerce

An epidemic model with infectivity and recovery in both latent and infected period is introduced. Utilizing the LaSalle invariance principle and Bendixson criterion,the basic reproduction number is found, we prove that the disease-free equilibrium is globally asymptotically stable when the basic reproduction number is less than one. The disease-free equilibrium is unstable and the unique positive equilibrium is globally asymptotically stable when the basic reproduction number is greater than one.Numerical simulations support our conclusions.

#18

Tissue Significances Tests on DNA Binding Sequence Motifs for Human Genes

Xiujun GONG
Tianjin University

DNA binding sequence motifs are becoming increasingly important in the analysis of gene regulation, disease diagnosis and drug design. Although so far there are amount of tools available to discover these kinds of motifs, little was done to identify the biological functions, especially in tissue or cell type specific contributions, of those motifs. In this paper we used an integrated pipeline to discover sequences motifs for the promoter regions of human genes. Then we distinguished two types of motifs: tissue rich motifs (TRM) and tissue even motifs (TEM), using hypotheses test approaches including Bayesian hypothesis, Binomial distribution and traditional z-test. We finally got 233 overlapped TRMs and 56 TEMs. Most of those motifs are validated against JASPAR databases.

#22

Bi-factor analysis based on noise-reduction (BFANR): A New Algorithm for searching coevolving amino acid sites in proteins

Bingqiang Liu
School of Mathematics, Shandong University

The statistical analysis had shown that protein has correlated evolution between amino acid sites with certain interaction pattern. Although distributing in distant positions in the primary protein sequence, these interacting amino acids closely connected in the third-dimensional protein structure, composing specific structural elements. These elements are relatively independent in structure, function and evolution between each other. However, the sites within the same element have significant correlation in evolution. Thus, systemic studies on these structural elements inside a protein will contribute to the clarification of protein function. This paper proposed a new algorithm, noise-reduction (BFANR) based bi-factor analysis, to extracts significant structural elements from the statistical noise in amino acid sequences. We did internal correlation test, statistical independence test, evolutionary rate analysis and evolution independence analysis to evaluate the prediction of new algorithm. The results showed that the amino acids within each predicted structural element are closely related, while different structural elements are significantly statistical independent. The results also indicated that the structural elements have specific evolution directions. In addition, the evaluation showed that new algorithm was more robust under the influence of noise sites and could extract non-random protein structural elements from the large amount of noise sites.

#23

Anti-rheumatic effects of Tripterygium wilfordii Hook F in a network perspective

Jing Zhao
Department of Mathematics
Logistical Engineering University
Chongqing, China

Rheumatoid arthritis (RA) is a chronic disease that affects the joints, often those in a person's wrists, fingers, and feet. In contrast to FDA-approved anti-RA drugs, Tripterygium wilfordii Hook F (TwHF), a traditional Chinese medicine (TCM), featured as multi-targeting, have been acknowledged with notable anti-RA effects although the pharmacology is unclear. In this work, we investigated the therapeutic mechanisms of TwHF at protein network level. First, RA-associated genes, the protein targets of FDA approved anti-RA drugs and TwHF were collected. Then we mapped the protein targets of TwHF on the drug-target network of FDA approved anti-RA drugs and KEGG RA pathway, based on these information and resources. Furthermore, we quantitatively analyzed the anti-rheumatic effect of TwHF and compared it with those of FDA approved anti-RA drugs by a network based anti-rheumatic effect score. Our study suggests that TwHF may function as a combination of disease-modifying anti-rheumatic drug and non-steroidal anti-inflammatory drug and its anti-rheumatic power could be comparable with that of anti-inflammatory agents. This study may facilitate our understanding of the RA treatment by TwHF from the perspective of network systems and it may suggest new approach for the study of TCM pharmacology.

#24

Subcellular localization prediction of apoptosis proteins based on the data mining for amino acid index database

Yuhua Yao
Zhejiang Sci-Tech University

In this work, based on the ACF model and the SVM classifier, succeeded on trials mining information that it's more effective to analyze the subcellular localization prediction of apoptosis proteins when adopting hydrophobicity property. This information is obtained in three benchmark datasets by using the ACF model and SVM to scan the AAindex database, which contains 544 kinds of amino acids. The contribution of this work is that it first did a comprehensive research on the effectiveness of the amino acid index for the subcellular localization of apoptosis proteins.

#27

Prediction of hot spots in protein interfaces using extreme learning machine

Lin Wang
Tianjin University of Science and Technology

The identification of hot spots, a small subset of protein interfaces that account for the majority of binding free energy, is becoming increasingly important for the research on protein-protein interaction and drug design. For each interface residue or target residue to be predicted, we extract hybrid features which incorporate a wide range of information of the target residue and its spatial neighbor residues, i.e. the nearest contact residue in the other face (mirror-contact residue) and the nearest contact residue in the same face (intra-contact residue). We present a novel extreme learning machine (ELM) model to effectively integrate these hybrid features for predicting hot spots in protein interfaces. The training set includes 318 alaninemutated interface residues derived from 20 protein complexes in Alanine Scanning Energetics Database (ASEdb). The independent test set contains 125 alanine-mutated interface residues in 18 protein complexes deposited in Binding Interface Database (BID). Our method can achieve accuracy (ACC) of 81.1% in the training set,

and ACC of 78.4% and Matthew's correlation coefficient (MCC) of 0.448 in the independent test set. Compared to SVM model with the same hybrid features, ELM model achieves competitive accuracy in the training set and higher accuracy in the test set. Furthermore, performance of our ELM model exceeds other existing methods, such as Robetta, FOLDEF, KFC, KFC2, MINERVA and HotPoint in the independent test set.

#28

Rank-based interolog mapping for predicting protein-protein interactions between genomes

Jinn-Moon Yang

National Chiao Tung University

As rapidly increasing number of sequenced genomes, the methods for predicting protein-protein interactions (PPIs) from one organism to another is becoming important. Best-match and generalized interolog mapping methods have been proposed for predicting (PPIs). However, best-match mapping method suffers from low coverage of the total interactome, because of using only best matches. Generalized interolog mapping method may predict unreliable interologs at a certain similarity cutoff, because of the homologs differed in subcellular compartment, biological process, or function from the query protein. Here, we propose a new "rank-based interolog mapping" method, which uses the pairs of proteins with high sequence similarity (E-value<10-10) and ranked by BLASTP E-value in all possible homologs to predict interologs. First, we evaluated "rank-based interolog mapping" on predicting the PPIs in yeast. The accuracy at selecting top 5 and top 10 homologs are 0.17, and 0.12, respectively, and our method outperformed generalized interolog mapping method (accuracy=0.04) with the joint E-value<10-70. Furthermore, our method was used to predict PPIs in four organisms, including worm, fly, mouse, and human. In addition, we used Gene Ontology (GO) terms to analyzed PPIs, which reflect cellular component, biological process, and molecular function, inferred by rank-based mapping method. Our rank-based mapping method can predict more reliable interactions under a given percentage of false positives than the best-match and generalized interolog mapping methods. We believe that the rank-based mapping method is useful method for predicting PPIs in a genome-wide scale.

#29

A Co-expression Modules Based Gene Selection for Cancer Recognition

Deng Yong

Hunan university

Gene expression profiles are used to recognize patient samples for cancer diagnosis and therapy. Gene selection is the crucial for high recognition performance. In usual gene selection methods the genes are considered as independent individuals and the correlation among genes are not used efficiently. In this description, a co-expression modules based gene selection method for cancer recognition is proposed. First, in the cancer dataset a weighted correlation network is constructed according to the correlation between each pairs of genes. Second, different modules from this network are identified and the significant modules are selected for following exploring. Then based on these informative modules information gain is applied to select the feature genes for cancer recognition. At last using LOOCV the experiments with different classification algorithms are conducted and the results show the proposed method makes better classification accuracy than traditional gene selection methods.

#30

Exploring the interaction patterns in seasonal marine microbial communities with network analysis

Shao-Wu Zhang
Northwestern Polytechnical University

With the development of high-throughput and low-cost sequencing technology, a large amount of marine microbial sequences is generated. So, it is possible to research more uncultivated marine microbes. The interaction patterns of marine microbial species and marine microbial diversity are hidden in these large amount sequences. Understanding the interaction pattern and structure of marine microbe have a high potential for exploiting the marine resources. Yet, very few marine microbial interaction patterns are well characterized even with the weight of research effort presently devoted to this field. In this paper, based on the 16S rRNA tag pyrosequencing data taken monthly over 6 years at a temperate marine coastal sits in West English Channel, we employed the CROP unsupervised probabilistic Bayesian clustering algorithm to generate the operational taxonomic units (OTUs), and utilized the PCA-CMI algorithm to construct the spring, summer, fall, and winter seasonal marine microbial interaction networks. From the four seasonal microbial networks, we introduced a novel module detecting algorithm called as DIDE, by integrating the dense subgraph, edge clustering coefficient and local modularity, to detect the interaction pattern of marine microbe in four seasons. The analysis of network topological parameters shows that the four seasonal marine microbial interaction networks have characters of complex networks, and the topological structure difference among the four networks maybe caused by the seasonal environmental factors. The marine microbial interaction patterns detected by DIDE algorithm in four seasons show evidence of seasonally interaction pattern diversity. The interaction pattern diversity of fall and winter is more than that of spring and fall, which indicates that the seasonal variability might have the greatest influence on the marine microbe diversity.

#31

Codon-Based Encoding for DNA Sequence Analysis

Byeong-Soo Jeong
Computer Engineering Dept. Kyung Hee University

With the exponential growth of biological sequence data (DNA or Protein Sequence), DNA sequence analysis became an essential task for biologist to understand its features, function, structure, and evolution. Encoding DNA sequences is an effective method to extract the features from DNA sequences. It has been popularly used for visualizing DNA sequences and analyzing similarities/dissimilarities between different species or cells. Even though there have been many encoding approaches proposed for DNA sequence analysis, we still need more elegant approaches for higher accuracy. In this paper, we propose another encoding approach for measuring similarity/dissimilarity between different species. Our approach can preserve nucleotide's physiochemical property, positional information, and also codon usage bias. Extensive performance study shows that our approach can provide higher accuracy than existing approaches in terms of degree of similarity.

#32

Proteome Compression via Protein Domain Compositions

Morihiro Hayashida
Bioinformatics Center, Institute for Chemical Research, Kyoto University

In statistical-mechanical angle, a living individual is regarded to construct an open non-equilibrium thermodynamic system, and to organize its own self. A DNA base sequence is one of information that an

individual maintains, and has been mutated, substituted. In this paper, we focus on domain compositions of proteins generated from DNAs as such information. We suppose that a protein is a multiset of domains, and compress whole proteins in an organism for the sake of obtaining the entropy. Since gene duplication and fusion have occurred through evolutionary processes, the same domains and the same compositions of domains appear in multiple proteins, which enables us to compress a proteome by using references to proteins for duplicated and fused proteins. Such a network with references to at most two proteins is modeled as a directed hypergraph. We propose a heuristic approach by combining the Edmonds algorithm and an integer linear programming, and apply our procedure to six proteomes of E. coli, S. cerevisiae, D. melanogaster, C. elegans, A. thaliana, and M. musculus. The compression size using both of duplication and fusion was smaller than that using only duplication. In addition, we observed several fusion events for these organisms.

#33

GPU-Meta-Storms: Computing the similarities among massive microbial communities using GPU

Xuetao Wang

Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences

With the development of next-generation sequencing and metagenomic technologies, the number of metagenomic samples of microbial communities is increasing with exponential speed. The comparison among metagenomic samples could facilitate the data mining of the valuable yet hidden biological information held in the massive metagenomic data. However, current methods for metagenomic comparison are limited by their ability to process very large number of samples each with large data size.

In this work, we have developed an optimized GPU-based metagenomic comparison algorithm, GPU-Meta-Storms, to evaluate the quantitative phylogenetic similarity among massive metagenomic samples, and implemented it using CUDA (Compute Unified Device Architecture) and C++ programming. Our results have shown that with the optimization of the phylogenetic comparison algorithm, memory accessing strategy and parallelization mechanism on many-core hardware architecture, the GPU-Meta-Storms could compute the pair-wise similarity matrix for 1920 metagenomic samples in 4 minutes, which gained a speed-up of more than 1000 times compared to CPU version Meta-Storms on single-core CPU, and more than 100 times on 16-core CPU. Therefore, the high-performance of GPU-Meta-Storms in comparison of massive metagenomic samples could thus enable in-depth data mining from massive metagenomic data, and make the real-time analysis and monitoring of constantly-changing metagenomic samples possible.

#34

A key network approach reveals new insight in Alzheimer's disease

Jan Schlüsener

Division of Immunopathology of the Central Nervous System

Alzheimer's disease is a severe neurodegenerative disorder without curative treatment. Extensive research on pathological molecular processes accumulated over the last years. These data combined will allow a system biology approach to identify the key regulatory elements of the disease and the establish a model descriptive of the disease process and predictive for the development of therapeutic agents. In this paper we propose a key network that is closed and uses a set of nodes as key elements: APP, TAU, BACE, Glutamate, CDK5, PI3K and HIF1α, which have been shown to be of importance to the pathogenesis of Alzheimer's disease. The created network, in total 40 nodes, is capable of creating new insight into feedback loops that seem to be of importance for the progression of

the disease. Further, it indicates cross-talk between pathways and identifies suitable target combination for therapy of AD.

#35

Accelerating Processing Speed in Pathway Research Based on GPU

Ting Yao

Hunan University

Genome-wide association study (GWAS) has become an effective and successful method to identify disease loci by considering SNPs independently. However, it may be invalid for uncovering the disease loci that not reaching a stringent genome-wide significance threshold. As a result, multi-SNP GWAS is developing rapidly as a complement to traditional GWAS. However, the high computational cost becomes a major limitation for it. The graphical processing unit (GPU) is a programmable graphics processor which has powerful parallel computing ability. And with the development, GPUs have been feasible for many scientific studies. Hence, we are motivated to use GPUs for pathway-based GWAS to improve computational efficiency. The experiment results attained showed the speed-up ratio can reach up to more than 160.

#36

Literature Mining of Protein Phosphorylation Using Dependency Parse Trees

Minghui Wang

School of Information Science and Technology, University of Science and Technology of China

As one of the most ubiquitous post-translational modifications (PTMs), protein phosphorylation plays an important role in various biological processes, such as signaling transduction, cellular metabolism, differentiation, growth, regulation and apoptosis. Information of protein phosphorylation is of great value not only in illustrating the underlying molecular mechanisms but also treatment of diseases and design of new drugs. Recently, there is an increasing interest in automatically extracting phosphorylation information from biomedical literatures. However, it still remains a challenging task due to the tremendous volume of literature and circuitous modes of expression for protein phosphorylation. In this study, we propose a novel text-mining method for efficiently retrieving and extracting protein phosphorylation information from literature. By employing natural language processing (NLP) technologies, this method transforms each sentence into dependency parse trees that can precisely reflect the intrinsic relationship of phosphorylation-related key words, from which detailed information regarding substrates, kinases and phosphorylation sites is extracted based on syntactic patterns. Compared with other existing tools, the proposed method demonstrates significantly improved performance measured by precision and recall, suggesting it is a powerful bioinformatic approach for retrieving phosphorylation information from a large amount of literature.

#38

Dynamical behaviour of an anti-HBV infection therapy model with time-delayed immune response

Xinjian Zhuo

Beijing University of Posts and Telecommunications

Mathematical models have been used to under-stand the factors that govern infectious disease progression in viral infections. In this paper, based on the standard mass action incidence, an anti-HBV therapy model with time-delayed immune response is set up. The time-delay is used to describe the period of time for antigenic stimulation to generate CTLs. The globally asymptotically stable analysis of the infection-free equilibrium is given in the paper. Some conditions for Hopf

bifurcation around endemic equilibrium to occur are also obtained by using the time delay as a bifurcation parameter.

#40

A Novel HMM for Analyzing Chromosomal Aberrations in Heterogeneous Tumor Samples

Ao Li

School of Information Science and Technology,University of Science and Technology of China

Comprehensive detection and identification of copy number and LOH of chromosomal aberration is required to provide an accurate therapy of human cancer. As a cost-saving and high-throughput tool, SNP arrays facilitate analysis of chromosomal aberration throughout the whole genome. The performance of previous approaches has been limited to several critical issues such as normal cell contamination, aneuploidy and tumor heterogeneity. For these reasons we present a Hidden Markov Model (HMM) based approach called TH-HMM (Tumor Heterogeneity HMM), for simultaneous detection of copy number and LOH in heterogeneous tumor samples using data from Illumina SNP arrays. Through adopting an efficient EM algorithm, our method can correctly detect chromosomal aberration events in tumor subclones. Evaluation on simulated data series indicated that TH-HMM could accurately estimate both normal cell and subclone proportions, and finally recovery the aberration profiles for each clones.

#41

Two Programmed Replicative Lifespans of Saccharomyces cerevisiae Formed by Endogenous Molecular-Cellular Network

Jie Hu

Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University

Cellular replicative capacity is a therapeutic target for regenerative medicine as well as cancer treatment. The mechanism of replicative senescence and cell immortality is still unclear. We investigated the diauxic growth of Saccharomyces cerevisiae and proposed that there are two robust states with finite and infinite programmed cellular replicative lifespans formed by the endogenous molecular-cellular network of S. cerevisiae. For the state with the infinite lifespan, the cell cycle related gene expressions are continuously oscillating. In the other state, the cooperative effect of the mitogen-activated protein kinase (MAPK) signaling pathways with the cell cycle leads to a result that the cells stop dividing after several generations counting from the beginning of the post-diauxic growth. The number of dividing times is determined by both the endogenous network and the initial distribution of the corresponding gene expressions.

#42

Cell Commitment Motif Composed of progenitor-specific TF and Fate-Decision Motif

Tongpeng Wang

Institute of System Biology, Shanghai University

Mutual-inhibition motif is frequently-occuring motif in transcriptional regulatory networks for cell lineage commitment. Stable attractors represent cell commitment state. But how progenitor-specific transcription factors stabilize progenitor cells and commit them to different cell fates remains unexplained.In this paper we represent the motif for cell commitment, composed of mutual-inhibition motif and progenitor-specific transcription factor, and develop associated mathematical model. In the analysis of bifurcation and dynamical simulation, the model could exhibit multiple steady stable states and transition between them, cooresponding to

progenitor, committed cell state and different commitment processes. Furthermore, we demonstrate that different commitment patterns, for example that of hematopoitic stem cell and neural stem cell, could be represented with different bifurcation features.

#43

Temporal order of somatic mutations during tumorigenesis based on Markov chain model

Hao Kang

Key Laboratory of System Biology, Shanghai Institutes for Biological Science, Chinese Academy of Science

Tumors are developed and worsen with the accumulated mutations on DNA sequences during tumorigenesis. Identifying the temporal order of gene mutations will provide not only a new insight to study the tumorigenesis at the level of genome sequences, but also an effective tool to achieve early diagnosis as well as preventive medicine for patients. In this paper, we develop a Markov chain model based approach (TOMC) to accurately estimate the temporal order of gene mutations during tumorigenesis from genome sequencing data. We applied our method to analyze both simulated and real data-sets. Our approach has obvious advantages over the conventional methods by considering the effect of mutation dependence among genes. In our analysis, tumor suppressor genes (TSG) have been found to mutate ahead of oncogenes (OCG), which are considered as key events of functional loss and gain during tumorigenesis. Besides, our method provides a quantitative way to understand the development and progression of tumorigenesis based on high throughput sequencing data.

#44

Towards Kinetic Modeling of Metabolic Networks with Incomplete Parameters

Wei Zheng

Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China

Modeling is an important direction in systems biology. The target towards kinetic modeling for metabolic network is to develop a practical computational method which can handle incomplete parameters. In principle, we could start with a set of randomly chosen parameters; calculating fluxes and metabolites concentration and comparing with experiments; iterating until the best parameters are found. But the large parametric space may require billions of times of iterations. In order to overcome such a difficulty, we develop a method to obtain the structure of parametric space. We are able to discover the correlation between parameters and variables, which is helpful for us to estimate the possible value of parameters. Differ from previous method, the implicit relationship between parameter and variable are also provided directly by our method, which provides a potential for us to analyze the feature of metabolic network.

#45

Anti-Cancer Effect of Aloe Emodin on Breast Cancer Cells, MCF-7

Indah Mohd Amin

Universiti Teknologi MARA (UiTM) Malaysia

Phytochemicals of some plants are believed to have natural anti-proliferative properties to various cancer cells. Thus, they might have the potential as alternative choice for contemporary treatment as the latter are usually associated with many unpleasant side effects. The aim of this study is to investigate the possible anti-cancer effect of aloe emodin (AE; 1,8-Dihydroxy-3-hydro-xymethyl-anthraquinone) on estrogen-positive breast cancer cells, MCF-7. We

were able to demonstrate the efficiency of AE, an antraquinone derivatives which are present in Aloe Vera leaves, in limiting the proliferation effect of MCF-7 cells in a dose and time dependent manner using WST-1 assay. Our preliminary result suggests that AE could be a promising natural candidate for future pharmacological study, targeting in breast cancer prevention strategies.

#47

Reinitiation enhances reliable transcriptional responses in eukaryotes

Bo Liu

Institute of Industrial Science, The University of Tokyo

By establish a simple model that incorporates the reinitiation mechanism in eukaryotic transcription, we show by rigorous analysis based on the chemical master equations that reinitiation can enhance a reliable transcriptional response by reducing the noise intensity in the mRNA abundance though the following conditions: a high reinitiation rate, a stable reinitiation scaffold, and the direct coupling between the transcription the gene activation process.

#50

On Knowledge Discovery for Pancreatic Cancer Using Inductive Logic Programming

Yushan Qiu

The University of Hong Kong

Pancreatic cancer is a devastating disease and predicting the status of the patients becomes an urgent issue. In this paper, we explore the applicability of Inductive Logic Programming (ILP) method to show that the accumulated clinical laboratory data makes it possible to predict disease characteristics, which will contribute to the selection of therapeutic modalities of pancreatic cancer. The availability of a large amount of clinical laboratory data provides clues to aid in the knowledge discovery of diseases. In predicting the differentiation of tumor and the status of lymph node metastasis in pancreatic cancer, using our ILP model, we developed three rules that are consistent with descriptions in the literature. The rules we identified are useful to detect the differentiation of tumor and the status of lymph node metastasis in pancreatic cancer and therefore contributed significantly to the decision of therapeutic strategies. In addition, we also conduct a 5-fold cross-validation predictive accuracy to further validate the effectiveness and usefulness of the ILP model.

#51

A multi-scale approach for simulating time-delay biochemical reaction systems

Yuanling Niu

Central South University

This paper presents a multi-scale approach for simulating time-delay biochemical reaction systems when there are wide ranges of molecular numbers. We construct a new approach that can reduce the computational burden on the basis of the idea of a partitioned system and recent developments with stochastic simulation algorithm and the delay stochastic simulation method. It is shown that this algorithm is much more efficient than existing methods such as DSSA method and the modified next reaction method. Some numerical results are reported, confirming the accuracy and computational efficiency of the approximation.

#52

A Method For Discriminating Native Protein-DNA Complexes From Decoys Using Spatial Specific Scoring Matrices

Changhui Yan

north dakota state university

Decoding protein-DNA interactions is important to understanding gene regulation and has been investigated by worldwide scientists for a long time. However, many aspects of the interactions still need to be uncovered. The crystal structures of protein-DNA complexes reveal detailed atomic interactions between the proteins and DNA and are an excellent resource for investigating the interactions. In this study, we profiled the spatial distribution of protein atoms around six structural components of the DNA, which are the four bases, the deoxyribose sugar and the phosphate group. The resultant profiles not only revealed the preferred atomic interactions across the protein-DNA interface but also captured the spatial orientation of the interactions. The profiles are a useful tool for investigating protein-DNA interactions. We tested the strength of profiles in two experiments, discrimination of native protein-DNA complexes from decoys with mutant DNA and discrimination of native protein-DNA complexes from near-native docking decoys. The profiles achieved an average Z-score of 7.41 and 3.22 respective on benchmark datasets for the tests, both are better than other knowledge-based energy functions that model protein-DNA interaction based on atom pairs.

#55

Gene Ontology Based Housekeeping Gene Selection for RNA-seq Normalization

Tun-Wen Pai

Dept. of Computer Science and Engineering & Center of Excellence for the Oceans, Naional Taiwna Ocean University

RNA-seq analysis provides a powerful tool to reveal relationship between gene expression levels and biological function of proteins. However, prior to analyzing differential gene expression, selection of suitable housekeeping genes for calibration of multiple experimental datasets is still a challenging problem. This research proposes a novel method to facilitate biologists in selecting appropriate housekeeping genes for dataset normalization, which is mainly based on GO annotations and previously published transcriptome datasets. By integrating characteristics of GO term distance and cross-tissue / temporal stability of gene expression among various types of tissue / time points, the proposed method can enumerate a set of functionally related housekeeping genes to serve as a reference gene set for dataset normalization. Based on GO term distance measurement, the suggested housekeeping gene set possesses the most irrelevant relationship with user-defined keywords of experimental datasets. Testing RNA-seq datasets of black porgy have demostrated that selection of different housekeeping genes leads to strong impact on comparative results of differential gene expression analysis. The proposed methodology of houskeeping gene selection for inter-dataset normalization is extremely useful for RNA-seq and microarray related researches.

#56

Electrostatics and Structural Analysis of DNA-binding Sites in SSBs and DSBs

Wei Wang

School of computer, Wuhan University

Single-stranded DNA-binding proteins (SSBs) and double-stranded DNA-binding proteins (DSBs) play different roles in biological processes through binding stranded DNA-binding proteins or double-stranded DNA. However, the underlying binding mechanisms of SSBs and DSBs are not fully understood. Here, we employed binding

specificity sites in the SSBs and DSBs interfaces from known 3D structures, and extracted a set of novel features (electrostatic charge， secondary structure and spatial shape) based on the binding specificity sites to distinguish SSBs and DSBs. Using these features, we constructed a classifier to predict SSBs and DSBs on a non-homologous dataset. With 10-fold cross-validation, the classifier achieved an accuracy of 85.4%, F-measure of 0.86, MCC of 0.72 and AUC of 0.86 respectively. In conclusion, we found three new features from the binding specificity sites in the SSBs and DSBs interface which discriminate between the SSBs and DSBs. In addition, these features can also deepen our understanding of the proteins' specificity in the binding to ssDNA or dsDNA and assist us to do functional annotation for protein.

#57

Predicting the non-compact conformation of amino acid sequence by particle swarm optimization

Yuzhen Guo
Nanjing University of Aeronautics and Astronautics

Hydrophobic-hydrophilic (HP) model serves as a surrogate for the protein structure prediction problem to fold a chain of amino acids into a 2D square lattice. By the fact that the number of amino acids is equal to the number of lattice points or not, there are two types of folding conformations, i.e., the compact and non-compact conformations. Non-compact conformation tries to fold the amino acids sequence into a relatively larger square lattice, which is more biologically realistic and significant than the compact conformation. Here, we propose a heuristic algorithm to predict the non-compact conformations in 2D HP model. First, the protein structure prediction problem is abstracted to match amino acids to lattice points. The problem is then formulated as an integer programming model and we transform the biological problem into an optimization problem. Classical particle swarm optimization algorithm is extended by the single point adjustment strategy to solve this problem. Compared with existing self-organizing map algorithm, our method is more effective in several benchmark examples.

#58

Meta-Analysis on Gene Regulatory Networks Discovered by Pairwise Granger Causality

Gary Hak Fui Tam
Department of Electrical and Electronic Engineering, The University of Hong Kong

Identifying regulatory genes partaking in disease development is important to medical advances. Since gene expression data of multiple experiments exist, combining results from multiple gene regulatory network discoveries offers higher sensitivity and specificity. However, data for multiple experiments on the same problem may not possess the same set of genes, and hence many existing combining methods are not applicable. In this paper, we approach this problem using a number of meta-analysis methods and compare their performances. Simulation results show that vote counting is outperformed by methods belonging to the Fisher's chi-square (FCS) family, of which FCS test is the best. Applying FCS test to the real human HeLa cell-cycle dataset, degree distributions of the combined network is obtained and compared with previous works. Consulting the BioGRID database reveals the biological relevance of gene regulatory networks discovered using the proposed method.

#59

EdgeSVM: a method for identifying differentially correlated gene pairs as edge markers

Wanwei Zhang, Tao Zeng and Luonan Chen
Key Laboratory of Systems Biology, SIBS, CAS

Biomarker discovery is one of the major topics in high-throughput biological data analysis. Traditional methods focus on differentially expressed genes but ignore non-differentials. We assume that non-differentially expressed genes also contain rich functional information as the rewired interactions / edges in a biological network. Therefore, it is necessary to identify relevant interactions or gene pairs to distinguish different phenotypes in any case-control study. To address this issue, we proposed a new method, edgeSVM, to identify the differentially correlated pairs between non-differentially expressed genes as edge biomarkers, which has strong ability on distinguishing normal and disease samples. The merit of this study is to induce the exact mathematical model of previously used differential co-expression network from node space (i.e. genes) to edge space (i.e. interactions). In actual analysis on human cholangiocarcinoma dataset, the novel candidate edge biomarkers discovered by edgeSVM cast new insight into the pathogenesis of non-differentially expressed genes in cholangiocarcinoma.

#60

Module network based cross-tissue analysis of Type 1 diabetes mellitus

Tao Zeng

Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

There is no effective cure for advanced Type 1 diabetes mellitus (T1DM) nowadays, and thus it is crucial to detect and further treat T1DM in earlier stage. Different from analyses on individual genes, network-based studies can give biological hints on the molecular mechanism of T1DM initiation and progression at a system-wide level. To reveal dynamical organizations of these gene modules during T1DM development, we identify the pre-disease modules at pre-disease stage based on dynamical network biomarkers (DNBs); detect progressive modules at early stage by Progressive Module Network Model (PMNM), and further disease-responsive modules at advanced stage by cross-tissue gene expression analysis. In particular, using PMNM, we analysed the gene expression data of T1DM NOD mouse model. We found: (1) the critical transition point was identified and confirmed by the pre-disease modules or DNBs, which is considered as an earlier event during disease progression; (2) several highly ranked disease-responsive modules were significantly enriched on known T1DM associated genes with rewiring association networks, which are marks of later events during T1DM development; (3) tissue-specific progressive modules revealed several essential progressive genes, and a few of pathways representing the effect of environmental factors during T1DM early development.

#61

Effective identification of essential proteins based on priori knowledge, network topology and gene expressions

Min Li

School of Information Science and Engineering, Central South University

Identification of essential proteins is very important for understanding the minimal requirements for cellular life and also necessary for a series of practical applications, such as drug design. With the advances in high throughput technologies, a large number of protein-protein interactions are available, which makes it possible to detect proteins' essentialities from the network level. Considering that most species already have a number of known essential proteins, we propose a new priori knowledge-based scheme to discover new essential proteins from protein-protein interaction networks. Based on the new scheme, two essential protein discovery algorithms, CPPK and CEPPK, are developed. CPPK predicts new essential proteins based on network topology and CEPPK detects new

essential proteins by integrating network topology and gene expressions. The performances of CPPK and CEPPK are validated based on the protein-protein interaction network of Saccharomyces cerevisiae. The experimental results show that the priori knowledge of known essential proteins is effective for improving the predicted precision. The predicted precisions of CPPK and CEPPK clearly exceed that of the other ten previously proposed essential protein discovery methods: Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC), Subgraph Centrality(SC), Eigenvector Centrality(EC), Information Centrality(IC), Bottle Neck (BN), Density of Maximum Neighborhood Component (DMNC), Local Average Connectivity-based method (LAC), and Network Centrality (NC). Especially, CPPK achieves 40% improvement in precision over BC, CC, SC, EC, and BN, and CEPPK performs even better.

#62

Overshooting in biological systems modeled by Markov chains

Chen Jia
School of Mathematical Sciences, Peking University

A number of biological systems can be modeled by Markov chains, or equivalently, master equations. Recently, there has been an increasing concern about when biological systems modeled by Markov chains will perform a dynamic behavior called overshooting. In this article, we find that the steady-state behavior of the system will have a great effect on the occurrence of overshooting. We make it clear that overshooting in general can not occur in systems which will finally approach an equilibrium steady state. This explains why overshooting is only observed in systems with three or more states. We find that there are two types of overshooting, named as normal overshooting and oscillating overshooting. The previous type can be viewed as the limit of the latter type as the period of the oscillation tends to infinity. We also show that except for extreme cases, oscillating overshooting will occur if the system is far from equilibrium.

#63

The Residue Interaction Network Analysis of Dronpa and a DNA clamp

Guang Hu
Soochow University

Topology is an essential aspect in protein structure. The network paradigm is increasingly used to describe the topology and dynamics of proteins. In this paper, the effect of topology on residue interaction network was first investigated in the context of two types of proteins: Dronpa and a DNA clamp, which have cylindrical and toroidal topologies. Network properties including characteristic path lengths, clustering coefficients, diameters and centrality measures have been calculated to describe their small-world properties, density, as well as the detail topology of hydrophobic pocket in Dronpa and the communication path across the interface in the DNA clamp. We have also calculated residue centrality for both proteins and used them to predict residues, which are critical to the function of a given protein. The results are discussed in comparison with network properties of globular proteins and existing Elastic network model data. We hope that the relationship elucidated between residue interaction network and protein topology could be extended to other proteins.

#64

Construction of human tissue-specific phosphorylation networks with protein expression data

Yin-Ying Wang
Shanghai University

Phosphorylation is a post-translational modification process mediated by kinases through the addition of a covalently bound phosphate group, which plays important roles in a wide range of cellular progresses, such as signaling cascades and development. Over the past years, despite many phosphorylation sites have been determined with mass spectrometry techniques, it is not clear which kinase phosphorylates which proteins. Under the circumstance, we propose a new probabilistic model to identify the substrates phosphorylated by certain kinases. Furthermore, we construct three tissue-specific phosphorylation networks based on protein expression data. Investigating the constructed tissue specific networks, we find they are functionally consistent with the corresponding tissues, implying the effectiveness and biological significance of our proposed approach.

#65

Inferring Gene Regulatory Networks from Integrative Omics Data via LASSO-type regularization methods

Jing Qin

The University of Hong Kong

Inference of gene regulatory network from gene expression data at whole genome level is a grand challenge, especially in higher organisms, when the number of genes is large but the number of experimental samples is small. It is reported that the accuracy of current methods at genome scale dramatically drops from E. coli to S. cerevisiae due to the increase in the number of genes. This limits the applicability of current methods to the more complex genomes, like human and mouse. Least absolute shrinkage and selection operator (LASSO) is widely used for gene regulatory network inference from gene expression profiles. However, the accuracy of LASSO in large genome is not satisfactory. In this study, we apply two extended models of LASSO, the L0 and L1/2 regularization models, to infer gene regulatory network from both high-throughput gene expression data and transcription factor binding data in mouse embryonic stem cells (mESCs). We find that the L0 and L1/2 regularization models significantly outperform LASSO for network inference, and incorporating interactions between transcription factors and their targets remarkably improve the prediction accuracy. Our work demonstrates the efficiency and applicability of these two models for gene regulatory network inference from integrative omics data in large genome. The application of the two models will facilitate biologists to study the gene regulation of higher model organisms in a genome-wide scale.

#66

A novel discretization method for processing digital gene expression profiles

Jibin Qu

Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences

Discretization serves as an important preprocessing step for analyzing gene expression data and many algorithms have been proposed. However, most of the discretization methods were designed for microarrays. As a new technology, digital gene expression (DGE) profiles can overcome the limitation of microarrays and were applied in a widely range. In this paper, we proposed a novel discretization method for DGE data and the validations in a time-series gene expression dataset proved the efficiency of our method.

#67

Multiclass Classification of Sarcomas using Pathway Based Feature Selection Method

Hui Lv

Shanghai Jiao Tong University

Gene expression based prediction of disease states and prognosis is an important research area in

biomedical informatics. In general gene based prediction is the dominant method. Recently several pathway activity based feature selection methods, such as condition-responsive genes (CORGs) have been proposed. Currently these methods were limited to binary classification, while in many clinical problems a multiclass protocol is needed such as the classification of sarcomas. Here we built a multiclass CORGs method named mCORGs for the sarcomas classification. A k-nearest neighbor (KNN) classifier was implemented to evaluate the performance of gene-based and mCORGs feature selection method by independent test and cross validation. The performance demonstrated that the mCORGs method is a feasible and robust feature selection method for multi-class prediction problem; it has comparable discriminative power with reported manual gene selection method and more biological context of biomarkers in the multiclass sarcomas classification problem.

#68

Detect taxonomy-specific pathway associations with environmental factors using metagenomic data

Xue Tian

Academy of Mathematics and Systems Science, CAS

In microbial communities, the taxonomic structure and functional capability are highly related. We proposed a method by considering the combination of taxa and functional categories to explore the ecological mechanisms of microbial communities. Using GOS metagenomic samples, we tested this idea and its effectiveness. The combination of taxonomies and functional groups could reflect the difference between habitats and may help to explain the combination adaptability of microbes to environment.