The 8th International Conference on Systems Biology
The 4th Translational Bioinformatics Conference
(ISB/TBC 2014)

Organizers

October 24-27, 2014
Qingdao, China

ISB/TBC 2014 Schedule

October 24 Friday	8:00-22:00	Registration (hotel lobby at Huiquan Dynasty Hotel)		
	14:00-15:40	Workshop WA1(Room N1MR)	Workshop WB1(Room HH)	
		Translational biomedical and clinical informatics Paper IDs: 137, 139, 111, 140, 92 Chair: Qian Zhu	Network Biology Paper IDs: 59, 80, 94, 91, 88 Chair: Xing-Ming Zhao	
	15:40-16:00	Coffee break		
	16:00-18:00	Workshop WA2(Room N1MR)	Workshop WB2(Room HH)	
		Metagenomics and its applications in medical research Paper IDs: 71, 78, 70, 119,76 Chair: Kang Ning	Computational Systems Biology Paper IDs: 20, 73, 55, 108 Chair: Zhi-Ping Liu	
	18:00-20:00	Dinner		
	20:00-21:30	Board member meeting of ORSC-CSB (Room N1MR)		
October 25	08:10-08:30	Opening Session (Chair: Luonan Chen)		
	08:30-10:30	Plenary Session P1 (Chair: Luonan Chen) (Room CICC)		
	10:30-10:50	Coffee break		
	10:50-12:30	Session A1(Room CICC)	Session B1(Room N1MR)	Highlight Session H1(Room HH)
		Network Medecine Paper IDs: 13, 56, 66, 60 , 57 Chair: Ho-Jin Choi	Genome Wide Association Study Paper IDs: 10, 68, 69, 107, 89 Chair: Haipeng Li	Bioinformatics Chair: Junwen Wang
	12:30-14:00	Lunch		
		Session A2(Room CICC)	Session B2(Room N1MR)	Highlight Session H2(Room HH)

ISB/TBC 2014 Schedule

Saturday	14:00-15:40	Translational Bioinformatics Paper IDs: 48, 77, 101, 67, 90 Chair: Xuefeng Bruce Ling	Next Generation Data Analysis Paper IDs: 16, 28, 72 , 37, 79 Chair: Bairong Shen	Bioinformatics Chair: LingYun Wu
	15:40-16:20	Coffee break & Poster session		
	16:20-18:00	Session A3(Room CICC)	Session B3(Room N1MR)	Highlight Session H3(Room HH)
		Translational Bioinformatics Paper IDs: 49, 58, 74, 46, 32 Chair: Yan Zhang	Network Biology Paper IDs: 23, 30, 33, 41,25 Chair: Fengfeng Zhou	Complex Disease Chair: Ya-Ping Tian
	18:00-20:00	Reception		
	20:00-21:30	The launch meeting for three sub-society under ORSC-CSB (In Chinses)		
October 26 Sunday	08:30-10:30	Plenary Session P2 (Chair Ju Han Kim) (Room CICC)		
	10:30-10:50	Coffee break		
	10:50-12:30	Session A4(Room CICC)	Session B4(Room N1MR)	Highlight Session H4(Room HH)
		Computational Systems Biology Paper IDs: 96, 82, 116,112, 121 Chair: Guanyu Wang	Computational Systems Biology Paper IDs:43, 50, 53, 2, 61 Chair: Minping Qian	Complex Disease Chair: Lin Gao
	12:30-14:00	Lunch		
	14:00-15:40	Session A5(Room CICC)	Session B5(Room N1MR)	Session C1(Room HH)
		Computational Systems Biology Paper IDs: 9, 52, 47, 100, 42 Chair: Ruiqi Wang	Bioinformatics Paper IDs: 65, 99, 104, 19 Chair: Hai-Peng Li	Complex Disease Paper IDs: 81,114, 95, 109 Chair: Xiufen Zou

ISB/TBC 2014 Schedule

	15:40-16:20	Coffee break & Poster session		
	16:20-18:00	Session A6(Room CICC)	Session B6(Room N1MR)	
		Computational Systems Biology Paper IDs: 21, 36, 83, 39 Chair: Jingzhi Lei	Bioinformatics Paper IDs: 75,97,102, 29 Chair: Jin Wang	
	18:30-20:00	Banquet		
October 27 Monday	8:00-18:00	One day excursion in Qingdao area (TBD). Departure at 8:00 from lobby.		

Room CICC: ConferenceDynasty International Conference Center(王朝国际会议中心)
Room N1MR: No.1 Meeting Room(1号会议室)
Room HH: Huiquan Hall(汇泉厅)

ISB/TBC 2014 Program

October 24-26, Qingdao, Shandong, China

October 24 (Friday) Registration and Workshops

08:00-22:00 Registration, Participants arrival in Qingdao, check in Huiquan Dynasty Hotel, and registration package pick up (Hotel Lobby at Huiquan Dynasty Hotel).

14:00-15:40 Workshop WA1 (No.1 Meeting Room(1号会议室))

Topic: Translational biomedical and clinical informatics

Chair: Qian Zhu

14:00-14:20 *Evidence Based Disease Network Construction towards Drug Repositioning*
Liwei Wang, Jiabei Wang and Qian Zhu
Department of Medical Informatics, Jilin University, Changchun, China
Paper ID: 140

14:20-14:40 *Network-based Analysis of Time Series RNA-Seq Gene Expression Data by Integrating the Interactome and Gene Ontology Information*
Yuji Zhang
University of Maryland School of Medicine, USA
Paper ID: 137

14:40-15:00 *Evidence based computational drug repositioning candidate screening pipeline design: Case Study*
Qian Zhu, Yuji Zhang, Hongfang Liu and Jiabei Wang
University of Maryland, Baltimore County, USA
Paper ID: 139

15:00-15:20 *PRECISE:PRivacy-prEserving Cloud-assisted quality Improvement Service in hEalthcare*
Feng Chen, Shuang Wang, Noman Mohammed, Samuel Cheng and **Xiaoqian Jiang**
School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, USA
Paper ID: 111

15:20-15:40 *Network cluster analysis of protein–protein interaction network identified biomarker for type 2 diabetes*
Zhonghui Li, Zijun Qiao, Wenli Ma and Wenling Zheng
Southern Medical University, Institute of Genetic Engineering, Guangzhou, China
Paper ID: 92

14:00-15:40 Workshop WB1 (Huiquan Hall(汇泉厅))

Topic: Network Biology

Chair: Xingming Zhao

14:00-14:20 A Novel Markov Chain Modeling Method for Identifying Differential Pathways
Zhirui Zhang and Hong-Qiang Wang
Institute of Intelligent Machines, Chinese Academy of Science, China
Paper ID:59

14:20-14:40 *Network-based detection of Disease Modules and Potential Drug Targets in Intractable Epilepsy*
Hongwei Chu, Changkai Sun, Xuezhong Zhou, Guangming Liu, Lin Liu, Minghui Lv, Xiaofeng Zhou, Yiwei Wang, Xing Li, Pin Sun and Yizhun Zhu
Liaoning Provincial Key Laboratory of Cerebral Diseases, Institute for Brain Disorders Dalian Medical University

Paper ID: 80

14:40-15:00 *Detection of SNP-SNP Interaction based on the Generalized Particle Swarm Optimization Algorithm* ***Move to Poster Session***
Ma Changyi, Shang Junliang, Li Shengjun and Sun Yan
School of Information Science and Engineering, Qufu Normal University, China
Paper ID: 94

14:40-15:00 *cLP: Linear Programming with Biological Constraints and its Application in Classification Problems*
Manli Zhou, Youxi Luo, Guoqin Mai and Fengfeng Zhou
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, P.R. China
Paper ID:91

15:00-15:20 *Comparative genomics reveals a global map of selenium utilization and evolution in prokaryotes*
Ting Peng, Jie Lin and Yan Zhang
Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
Paper ID: 88

15:40-16:00 Coffee break

16:00-18:00 Workshop WA2 (No.1 Meeting Room(1 号会议室))

Topic: Big-data science for Bioinformatics

Chair: Kang Ning

16:00-16:20 *Dissecting the obesity disease landscape: identifying gene-gene interactions that are highly associated with Body Mass Index*
Rishika De, Shefali S. Verma, Michael V. Holmes, Folkert Asselbergs, Jason H. Moore, Brendan Keating, Marylyn D. Ritchie and Diane Gilbert-Diamond
Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA
Paper ID: 76

16:20-16:40 *Next-Generation Sequencing Data Analysis on Cloud Computing*
Taesoo Kwon, Won Gi Yoo, Won-Ja Lee, Won Kim and Dae-Won Kim
Korea Center for Disease Control and Prevention, Korea
Paper ID: 71

16:40-17:00 *Robust high-order gene-gene interaction analysis in Genome Wide Association Studies*
Yongkang Kim and Taesung Park
Seoul national university, Korea
Paper ID: 78

17:00-17:20 *Improving Mental Health using Sentiment Analysis on a Social Network*
Giryong Choi, Hyo Jin Do and Ho-Jin Choi
Korea Advanced Institute of Science and Technology, Korea
Paper ID: 70

17:20-17:40 *The Correlation and Regression Analysis on Aerosol Optical Depth, Ice Cover and Cloud Cover in Greenland Sea*
Bo Qu, Albert Gabric, Peijuan Gu and Meifang Zeng
Nantong University, China
Paper ID: 119

16:00-18:00 Workshop WB2 (Huiquan Hall(汇泉厅))

Topic: Computational Systems Biology

Chair: Zhiping Liu

16:00-16:20 *Incorporating feature reliability in false discovery rate estimation improves statistical power to detect differentially expressed features*
Elizabeth Chong, Yijian Huang, Hao Wu, Tianwei Yu, Dean Jones, Arshed Quyyumi, Karan Uppal and Nima Ghasemzadeh
Department of Biostatistics and Bioinformatics, Emory University, USA
Paper ID: 108

16:20-16:40 *Bioinformatic Inference of Changes in Levels of Reactive Oxygen Species and Their Carcinogenic Effects in Papillary Thyroid Carcinoma with Hashimoto Thyroiditis*
Jin Wook Yi, Sang Huyk Kwak, Jo-Heon Kim, Eun Kyung Paik, Ji-Youn Sung, Jihan Yu, Ji Hyun Chang, Sang Yun Ha, Kyu Eun Lee, Yeo-Kyu Youn and Ju Han Kim
Department of Surgery, Seoul National University Hospital, Seoul, Korea
Paper ID: 20

16:40-17:00 *In silico analysis of mutations in PITX3 gene* ***Move to Session B6***
Abida Arshad, Rashda Abbasi, Christian Sieber, Muhammad Arshad and Nafees Ahmad
Department of Zoology, PMAS Arid Agriculture University, Rawalpindi, Pakistan
Paper ID: 29

17:00-17:20 *A New Staging Framework by Fusion Molecular and Clinical Variables through CART Model* ***Move to Session A6***
Hongmin Cai and Ying Jin
School of Computer Science & Engineering, South China University of Technology Guangzhou, China
Paper ID: 39

16:40-17:00 *Evolution Analysis for HA Gene of Human Influenza A H3N2 Virus (1990 - 2013)*
Su-Li Li, Meng-Zhe Jin and Zhao-Hui Qi
College of Information Science and Technology, Shijiazhuang Tiedao University Shijiazhuang, China
Paper ID: 55

17:00-17:20 *Selecting Representative Topics in Biomedical Research Articles using MeSH Descriptors*
Chae-Gyun Lim, Byeong-Soo Jeong and Ho-Jin Choi
Kyung-Hee University, Korea
Paper ID: 73

18:00-20:00 Dinner

20:00-21:30 Board member Meeting for Computational Systems Biology Society of ORSC (No.1 Meeting Room(1 号会议室))

October 25 (Saturday) Technical sessions

07:30-11:30 Registration for late arrivals *(*Hotel Lobby at Huiquan Dynasty Hotel*)*

08:10-08:30 Opening Session for ISB/TBC2011 (Dynasty International Conference Center(王朝国际会议中心))

Chair: Luonan Chen

8:30-10:30 ISB/TBC Plenary Session P1 (Dynasty International Conference Center(王朝国际会议中心))

Chair: Luonan Chen

8:30-09:10 *Proteomics and Bioinformatics – Diagnostic Approaches to the Diagnosis and Understanding of Kawasaki Disease*
Harvey J. Cohen
Children's Hospital, Stanford University School of Medicine, USA

09:10-09:50 *Multi-Dimensional Cancer Genomics Studies in Systems Bioinformatics Perspective*
Sanghyuk Lee
Department of Life Science and Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 120-750, South Korea

09:50-10:30 *TBD*
Raul Rabadan
Department of Biomedical Informatics, Center for Computational Biology & Bioinformatics, Columbia University College of Physicians & Surgeons, USA

10:30-10:50 Coffee break

10:50-12:30 ISB/TBC Session A1 (Dynasty International Conference Center(王朝国际会议中心))

Topic: Network Medicine

Chair: Ho-Jin Choi

10:50-11:10 *The Translational Disease Network—from Protein Interaction to Disease Co-occurrence*
Hyunjung (helen) Shin, Yonghyun Nam, Dong-Gi Lee and Sunjoo Bang
Department of Industrial Engineering, Ajou University, South Korea
Paper ID: 13

11:10-11:30 *Studying the Genetics of Complex Diseases with Ethnicity-Specific Human Phenotype Networks: The Case of Type 2 Diabetes in East Asian Populations*
Jingya Qiu, Jason H. Moore and Christian Darabos
Institute for Quantitative Biomedical Sciences, Dartmouth College
Paper ID: 56

11:30-11:50 *Inferring drug-disease associations based on known protein complexes*
Liang Yu, Jianbin Huang, Zhixin Ma, Jing Zhang, Yapeng Zou and Lin Gao
School of Computer Science and Technology, Xidian University
Paper ID: 60

11:50-12:10 *Select and Label (S&L): a Task-Driven Privacy-Preserving Data Synthesization Framework*
Zhanglong Ji, **Xiaoqian Jiang**, Haoran Li, Li Xiong and Lucila Ohno-Machado
University of California, San Diego
Paper ID: 66

12:10-12:30 *A semi-tensor product approach for Probabilistic Boolean network*
Xiaoqing Cheng, Wai-Ki Ching and Yushan Qiu

The University if HongKong
Paper ID: 57

10:50-12:30 ISB/TBC Session B1 (No.1 Meeting Room(1号会议室))

Topic: Genome Wide Association Study

Chair: Haipeng Li

10:50-11:10 *Functional dyadicity and heterophilicity of gene-gene interactions in statistical epistasis networks*
Ting Hu, Angeline Andrew, Margaret Karagas and **Jason Moore**
Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover NH, USA
Paper ID: 10

11:10-11:30 *Novel therapeutics for coronary artery disease from genome-wide association study data*
Mani P. Grover, Sara Ballouz and Merridee Wouters
School of Medicine, Deakin University, Geelong, Victoria, Australia.
Paper ID: 68

11:30-11:50 *Detection and Analysis of Disease-associated Single Nucleotide Polymorphism Influencing Post-translational Modification*
Yul Kim, Chiyong Kang, Bumki Min and Gwan-Su Yi
Dept. of Bio and Brain Engineering, KAIST, South Korea
Paper ID: 69

11:50-12:10 *Detecting Gene-Gene Interactions Using a Permutation-based Random Forest Method*
Jing Li, James Malley and Jason Moore
Dartmouth College, USA
Paper ID: 107

12:10-12:30 *SUMORESLER, a bioinformatics approach for identifying SUMOylation sites by combining sequence, structural and functional features*
Jinlei Zhang, Yang Zhang, Fuyi Li, Mingjun Wang, Geoffrey Webb, Chen Li, Jiangning Song
College of Information Engineering, Northwest A&F University, Yang100, China
Paper ID: 89

10:50-12:30 ISB/TBC Highlight Session H1 (Huiquan Hall(汇泉厅))

Topic: Bioinformatics

Chair: Junwen Wang

10:50-11:10 *Analysis of Stochastic cell dynamics from single cell data by Flow Cytometry (FCM)*
Minping Qian and Guanglu Gong

10:10-11:30 *Two novel formulations for biochemical reaction networks*
Tianshou Zhou
Sun Yat-Sen University

11:30-11:50 *Raison d'être of insulin resistance: the adjustable threshold hypothesis*
Guanyu Wang
Department of Biology, South University of Science and Technology of China

11:50-12:10 *Dynamical Analysis on a 2-D Disease Model with Convex Incidence Rate*
Pei Yu
Department of Applied Mathematics, Western University, London, Ontario, Canada

12:10-12:30 *Robust Period of Mammalian Circadian Oscillator from Amplitude Balance between Feedback Loops*
Jie Yan, Guangsen Shi, Zhihui Zhang, Xi Wu, Zhiwei Liu, Lijuan Xing, Zhipeng Qu, Zhen Dong, **Ling Yang** and Ying Xu
Center for Systems Biology, Soochow University, China

12:30-14:00 Lunch break

14:00-15:40 ISB/TBC Session A2 (Dynasty International Conference Center(王朝国际会议中心))

Topic: Translational Bioinformatics

Chair: Xuefeng Bruce Ling

14:00-14:20 *Identification of epigenetic modifications that contribute to pathogenesis in therapy-related AML: Effective integration of genome-wide histone modification with transcriptional profiles*
Xinan Yang, Bin Wang and John Cunningham
The University of Chicago, USA
Paper ID: 48

14:20-14:40 *Interpretation of personal genome sequencing data in terms of disease ranks based on mutual information*
Young-Ji Na, Kyung-Ah Sohn and Ju Han Kim
University of Pennsylvania, USA
Paper ID: 77

14:40-15:00 *Identifying network biomarkers by protein-protein interaction affinity derived from law of mass action*
Jingxue Xin, Xianwen Ren, Luonan Chen and Yong Wang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
Paper ID: 101

15:00-15:20 *VaccineWatch : a monitoring system of vaccine messages from social media data*
Somrak Numnark, Supawadee Ingsriswang and Duangdao Wichadakul
National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand
Paper ID: 67

15:20-15:40 *Differentially Private SNP selection in Genome-Wide Association Studies*
Haoran Li, **Xiaoqian Jiang**, Zhanglong Ji and Li Xiong
Emory University, USA
Paper ID: 90

14:00-15:40 ISB/TBC Session B2(No.1 Meeting Room(1号会议室))

Topic: Next Generation Data Analysis

Chair: Bairong Shen

14:00-14:20 *PDEGEM: Modeling non-uniform read distribution in RNA-seq data*
Yuchao Xia, Fugui Wang, Minping Qian, Zhaohui Qin and Minghua Deng
The Center of Quantitive biology Peking University, China
Paper ID: 16

14:20-14:40 *The NGS markup language(NGSML): a general medium for representation and exchange of NGS data*
Chunjiang Yu, Wentao Wu and Bairong Shen
Center for Systems Biology, Soochow University, 215006, Suzhou,China
Paper ID: 28

14:40-15:00 *Application of Meta-Mesh on the analysis of microbial communities from human associated-habitats*
Xiaoquan Su, Gongchao Jing, Shi Huang, Jian Xu and Kang Ning
Bioinformatics Group of Single-Cell Center, Qingdao Institute of Bioenergy and Bioprocess

Technology, Chinese Academy of Sciences
Paper ID: 37

15:00-15:20 *Bi-objective Optimization of a Continuous Biological Process*
Gongxian Xu, Ying Liu, Chao Yu and Dan Su
Department of Mathematics, Bohai University
Paper ID: 72

15:20-15:40 *Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer*
Dokyoon Kim, Ruowang Li, Scott Dudek and Marylyn Ritchie
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, USA
Paper ID: 79

14:00-15:40 ISB/TBC Highlight Session H2 (Huiquan Hall(汇泉厅))

Topic: Bioinformatics

Chair: Ling-Yun Wu

14:00-14:20 *MetalExplorer, a bioinformatics tool for the improved prediction of eight types of metal-binding sites using a random forest algorithm with two-step feature selection*
Jiangning Song
Monash University

14:20-14:40 *NeSSM: A Next-Generation Sequencing Simulator for Metagenomics*
Ben Jia and **Chaochun Wei**
Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University

14:40-15:00 *Novel Bioinformatics Method for Identification of Genome-Wide Non-Canonical Spliced Regions Using RNA-Seq Data*
Yongsheng Bai
The Center for Genomic Advocacy (TCGA), Department of Biology, Indiana State University

15:00-15:20 *SMAL: A Resource of Spontaneous Mutation Accumulation Lines*
Wen Wei, Lu-Wen Ning, Yuan-Nong Ye, Shi-Jie Li, Hui-Qi Zhou, Jian Huang, and **Feng-Biao Guo***
Center of Bioinformatics and Key Laboratory for NeuroInformation of the Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China.

15:20-15:40 *Compendium of Protein Lysine Modifications: from acetylation, ubiquitination to new modifications*
Zexian Liu and Yu Xue
Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

15:40-16:20 Coffee break

16:20-18:00 ISB/TBC Session A3 (Dynasty International Conference Center(王朝国际会议中心))

Topic: Translational Bioinformatics

Chair: Yan Zhang

16:20-16:40 *WebDISCO: a Web service for DIStributed COx model learning without patient-level data sharing*
Chia-Lun Lu, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, **Xiaoqian Jiang** and Lucila

Ohno-Machado
Division of Biomedical Informatics, University of California San Diego, USA
Paper IDs: 49

16:40-17:00 *Cross-Platform and Cross-Device Pedometer System Designed for Healthcare Services*
Yongjin Kwon, Rebekah Jiyoung Cha, Kyuchang Kang and Changseck Bae
Electronics and Telecommunications Research Institute, Korea
Paper IDs: 58

17:00-17:20 *A New Approach for Estimating the Progression of Pancreatic Cancer*
Shuhao Sun and Tianhai Tian
School of Mathematical Sciences, Monash University, Melbourne, Australia
Paper IDs: 74

17:20-17:40 *Prediction of trans-acting siRNAs in human brain*
Xiaoshuang Liu, Guangxin Zhang, Changqing Zhang and Jin Wang
Nanjing University, China
Paper IDs: 46

17:40-18:00 *Comparison of Multi-Sample Variant Calling Methods for Whole Genome Sequencing*
Kwangsik Nho, John West, Huian Li, Robert Henschel, Apoorva Bharthur, Michel Tavares and Andrew Saykin
Indiana University School of Medicine, USA
Paper ID: 32

16:20-18:00 ISB/TBC Session B3(No.1 Meeting Room(1号会议室))

Topic: Network Biology
Chair: Fengfeng Zhou

16:20-16:40 *Crosstalk between pathways enhances the controllability of signaling networks*
Dingjie Wang, **Suoqin Jin** and Xiufen Zou
School of Mathematics and Statistics, Wuhan University
Paper IDs: 23

16:40-17:00 *A Tensor-Based Markov Chain Method for Module Identification from Multiple Networks*
Chenyang Shen, Shuqin Zhang and Michael Kwok-Po Ng
Department of Mathematics, Hong Kong Baptist University
Paper IDs: 30

17:00-17:20 *Testing Multiple Hypotheses through IMP Weighted FDR Based on a Genetic Functional Network with Application to a New Zebrafish Transcriptome Study*
Jiang Gui, Casey Greene, Con Sullivan, Walter Taylor, **Jason Moore** and Carol Kim
Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH, US
Paper IDs: 33

17:20-17:40 *Parallelization of Enumerating Tree-like Chemical Compounds by Breadth-first Search Order*
Morihiro Hayashida, Jira Jindalertudomdee, Yang Zhao and Tatsuya Akutsu
Bioinformatics Center, Institute for Chemical Research, Kyoto University
Paper ID: 25

17:40-18:00 *Extracting discriminatively interpretable features of gene network by combining gene expression, variance and covariance*
Xiangtian Yu, Tao Zeng, Guojun Li and Luonan Chen
School of Mathematics, Shandong University, Jinan 250100, China
Paper IDs: 41

16:20-18:00 ISB/TBC Highlight Session H3 (Huiquan Hall(汇泉厅))

Topic: Complex disease

Chair: Ya-Ping Tian

16:20-16:40 *Precision medicine: translating genomics to clinical applications using networks and ontologies*
Yves A. Lussier
The University of Arizona

16:40-17:00 *MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes*
Xiaoqi Zheng, Qian Zhao, Hua-Jun Wu, et al.
Shanghai Normal Univerisity, China

17:00-17:20 *Prediction, Prevention and Treatment of CNS Metastases*
Xuefeng Bruce Ling
Stanford University, USA

17:20-17:40 *Development and validation of a novel computational approach to identify epigenetic biomarkers associate with cancer prognosis*
Li Xu, Xue Xiao and Shanguang Chen
Harbin Institute of Technolog

17:40-18:00 *Breast tumor subgroups reveal diverse clinical prognostic power*
Zhaoqi Liu, Xiang-Sun Zhang & Shihua Zhang
National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

18:00-20:00 Welcome Reception

20:00-21:30 The launch meeting for three sub-societies under Computational Systems Biology Society of ORSC (No. 1 Conference Room in Huiquan Dynasty Hotel(一号会议室))

October 26 (Sunday) Technical sessions

08:30-10:20 ISB/TBC Plenary Session P2 (Dynasty International Conference Center(王朝国际会议中心))

Chair: Ju Han Kim

8:30-09:10 *An Emerging Paradigm: New genes drive evolution of expression networks and phenotypes*
Manyuan Long
Edna K. Papazian Distinguished Service Professor, Department of Ecology and Evolution, The University of Chicago, USA

09:10-09:50 *Translational Bionformatics: Past, Present, and Future*
Jessica Tenenbaum
Duke Translational Medicine Institute, USA

09:50-10:30 *Using blood multi-biomarkers could improve clinical significance*
Yaping Tian
Department of Clinical Biochemistry, Chinese PLA General Hospital, Beijing 100853

10:30-10:50 Coffee break

10:50-12:30 ISB/ Session A4 (Dynasty International Conference Center(王朝国际会议中心))

Topic: Computational Systems Biology

Chair: Guanyu Wang

10:50-11:10 *Graph Pyramid Approach for Protein Classification*
Tushar Sandhan, Youngjoon Yoo, Jin Young Choi and Sun Kim
Seoul National University, Korea
Paper ID: 96

11:10-11:30 *Relating Hepatocellular Carcinoma Tumor Samples and Cell Lines Using Gene Expression Data in Translational Research*
Bin Chen, Marina Sirota, Hua Fan-Minogue, Dexter Hadley and Atul Butte
Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, USA
Paper ID: 82

11:30-11:50 *Concordance between ex vivo PBMC and in vivo human infections confirmed by N-of-1-pathways analysis of single-subject transcriptome*
Vincent Gardeux, Anthony Bosco, Jianrong Li, Fernando D. Martinez and Yves A Lussier
Department of Medicine, University of Arizona, Tucson, AZ, USA
Paper ID: 116

11:50-12:10 *Neural fate decisions mediated by oscillatory and sustained Hes1*
Shanshan Li, Zengrong Liu and Ruiqi Wang
Institute of systems biology, Shanghai University
Paper ID: 121

12:10-12:30
Identify Critical Genes in Development with Consistent H3K4me2 Patterns across Multiple Tissues
Nan Meng, Raghu Machiraju and **Kun Huang**
The Ohio State University, USA
Paper ID: 112

10:50-12:30 ISB/TBC Session B4 (No.1 Meeting Room(1号会议室))

Topic: Computational Systems Biology

Chair: Minping Qian

10:50-11:10 *An independent filter for gene set testing based on spectral enrichment*
H. Robert Frost, Zhigang Li, Folkert Asselbergs and Jason Moore
Institute for Quantitative Biomedical Sciences,Geisel School of Medicine, Lebanon,
Paper ID: 43

11:10-11:30 *Predicting Golgi-resident proteins in plants by incorporating N-terminal transmembrane domain information in the general form of Chou's pseudo-amino acid compositions*
Yasen Jiao, Xiaoquan Su and Pufeng Du
School of Computer Science and Technology, Tianjin University, Tianjin, China
Paper ID: 50

11:30-11:50 *Sparse Electrocardiogram Signals Recovery Based on Solving a Row Echelon-Like Form of System*
Pingmei Cai, Guinan Wang, Hongjuan Zhang, Shuxue Ding and Zikai Wu
Department of Mathematics, Shanghai University ong, China
Paper ID: 53

11:50-12:10 *Topological Characterization of Housekeeping Genes in Human Protein-Protein Interaction Network*
Pei Wang, Yuhuan Zhang, Jinhu Lu and Xinghuo Yu
Henan University & RMIT University, China
Paper ID: 2

12:10-12:30 *Combined analysis of gene regulatory network and SNP information enhances identification of potential gene markers in mouse knockout studies with small number of samples*
Benjamin Hur, Heejoon Chae and Sun Kim
Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea
Paper ID: 61

10:50-12:30 ISB/TBC Highlight Session H4 (Huiquan Hall(汇泉厅))

Topic: Complex disease

Chair: Lin Gao

10:50-11:10 Transforming Trillions of Points of Data into Diagnostics, Therapeutics, and New Insights into Disease
Atul Butte
Department of Pediatrics, Stanford University School of Medicine

11:10-11:30 *The systematice approach to cancer chemoresisance for better mechanistic understanding and DNA methylation diagonstics, a personal journey bigining with the discoveries made from the integrative multi-omic analysis to the robust assays fit to the clinical practice*
Jingde Zhu
Anhui Cancer Hospital Hefei and Shanghai Cancer Institute, Shanghai, China

11:30-11:50 *A Systems Kinetic Metbabolic Model for Xiamenmycin Biosynthetic Pathway*
Minjuan Xu, Yong-Cong Chen, Xiao-Mei Zhu, Jun Xu and Ping Ao
Shanghai Jiao Tong University

11:50-12:10 *Stochastic modelling of biochemical systems of multi-step reactions using a simplified two-variable model*
Qianqian Wu, Kate Smith-Miles, Tianshou Zhou and Tianhai Tian
Monash University

12:10-12:30 *Characterizing and controlling the inflammatory network during influenza A virus*

infection
Suoqin Jin, Yuanyuan Li, Ruangang Pan, Xiufen Zou,
School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China.

12:30-14:00 Lunch break

14:00-15:40 ISB/TBC Session A5 (Dynasty International Conference Center(王朝国际会议中心))

Topic: Bioinformatics

Chair: Ruiqi Wang

14:00-14:20 *Evolutionary Pressures on the Yeast Transcriptome*
Dominique Chu and Anton Salykin
University of Kent, School of Computing
Paper ID: 9

14:20-14:40 *Systematic identification of local structure binding motifs in protein-RNA recognition*
Zhi-Ping Liu
Shandong University
Paper ID: 52

14:40-15:00 *RCARE: RNA Sequence Comparison and Annotation for RNA Editing*
Sooyoun Lee, Je Gun Joung, Chan Hee Park, Ji Hye Park and Ju Han Kim
Seoul National University College of Medicine, Korea
Paper ID: 47

15:00-15:20 *Measuring the Similarity of Protein Structures Using Image Local Feature Descriptors SIFT and SURF*
Morihiro Hayashida, Hitoshi Koyano and Tatsuya Akutsu
Bioinformatics Center, Institute for Chemical Research, Kyoto University
Paper ID: 100

15:20-15:40 *The mathematical model and simulationg of predicting the non-compact conformations on triangle lattice*
Yuzhen Guo, Yong Wang and Zikai Wu
Department of Mathematics , Nanjing University of Aeronautics and Astronautics
Paper ID: 42

14:00-15:40 ISB/TBC Session B5 (No.1 Meeting Room(1号会议室))

Topic: Bioinformatics

Chair: Hai-Peng Li

14:00-14:20 *Identifying Prognostic Features by Bottom-up Approach and Correlating to Drug Repositioning*
Wei Li, Jian Yu, Baofeng Lian, Han Sun, Jing Li, Menghuan Zhang, Qian Liu, Yixue Li and **Lu Xie**
Shanghai Center for Bioinformation Technology, Shanghai Institutes of Biomedicine, Shanghai Academy of Science and Technology, Shanghai 201203, P. R. China
Paper ID: 65

14:20-14:40 *DMET-Miner: Efficient Learning of Association Rules from Genotyping Data for Personalized Medicine*
Pietro Hiram Guzzi, Mario Cannataro and **Giuseppe Agapito**
Informatics and Biomedical Engineering, University "Magna Græcia" of Catanzaro, Italy
Paper ID: 99

14:40-15:00 *Drug Name Recognition Using Conditional Random Fields with Word Embeddings*
Shengyu Liu, Buzhou Tang, Qingcai Chen, Xiaolong Wang and Bin Tang
Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School
Paper ID: 104

15:00-15:20 *Improving common lines detection in protein 3D structure reconstruction from the cryo-EM by a new optimized denosing filter*
Biao Zhang, Qiyu Jin, Jinhua Yang and Hong-Bin Shen
Shanghai Jiaotong University, China
Paper ID: 19

15:20-15:40 *TBD*
Chunbao Miao
Sugon Information Industry Co., Ltd., China

14:00-15:40 ISB/TBC Session C1 (Huiquan Hall(汇泉厅))

Topic: Complex disease

Chair: Xiufen Zou

14:00-14:20 *Centrality of complex disease genes unveiled by eQTL associations*
Haiquan Li, Nima Pouladi, Vincent Gardeux, Qi Luo, Qike Li, Jianrong Li, Fernando Martinez, Joe Garcia and Yves Lussier
University of Arizona, USA
Paper ID: 81

14:20-14:40 *A novel knowledge-based three-body potential for transcription factor binding site prediction*
Wenyi Qin, Guijun Zhao, Caiyan Jia and Hui Lu
University of Illinois at Chicago, USA
Paper ID: 114

14:40-15:00 *Network-based Prediction and Knowledge Mining Of Disease Genes*
Matthew Carson and Hui Lu
University of Illinois at Chicago, USA
Paper ID: 95

15:00-15:20 *Identification association of drug-disease by using functional Gene module for breast cancer*
Lida Zhu and Fuxi Zhu
Computer School of Wuhan University
Paper ID: 109

15:20-15:40 *Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks*
Lin Wang, Yong Wang, Qinghua Hu and Shao Li
Tianjin University of Science and Technology

15:40-16:20 Coffee break

16:20-18:20 ISB/TBC Session A6 (Dynasty International Conference Center(王朝国际会议中心))

Topic: Computational Systems Biology

Chair: Jinzhi Lei

16:20-16:40 *Detection of Core Cancer Modules by Mutated Gene Network in Glioblastoma*
Feng Li, Lin Gao and Xiaofei Yang

Xidian University
Paper ID: 21

16:40-17:00 *MapIn: an interactive tool for mapping biological descriptors to ontologies*
Panwen Wang, Jun Li, Xiaorong Liu, Pak Sham and Junwen Wang
The University of Hong Kong
Paper ID: 36

17:00-17:20 *Mining Correlation Patterns of Taxa, Pathways and Environmental Factors with An Improved Weighted Network Community Detection Algorithm*
Xiao-Ying Yan, Shao-Wu Zhang and Ze-Gang Wei
College of Automation, Northwestern Polytechnical Univerwsity, China
Paper ID: 83

17:20-17:40 *TBD*
Ling Wang
CloudScientific Technology Co., Ltd , China

17:40-18:00 *A New Staging Framework by Fusion Molecular and Clinical Variables through CART Model*
Hongmin Cai and Ying Jin
School of Computer Science & Engineering, South China University of Technology Guangzhou, China
Paper ID: 39

16:20-18:00 ISB/TBC Session B6 (No.1 Meeting Room(1号会议室))

Topic: Bioinformatics

Chair: Jin Wang

16:20-16:40 *Joint identification of differentially expressed genes and phenotype-associated genes*
Samuel Sunghwan Cho, Minseok Seo, Su-Kyung Shin, Eun-Young Kwon, **Yun-Jung Bae**, Mi-Kyung Sung, Myung-Sook Choi and Taesung Park
Seoul National University
Paper ID: 75

16:40-17:00 *An Entropy-based Statistical Workflow Provides Noise-Minimizing Biological Annotation for Muscular Aging*
Theodoros Koutsandreas, Ioannis Valavanis, Eleftherios Pilalis and **Aristotelis Chatziioannou**
Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation (NHRF), Athens, Greece
Paper ID: 97

17:00-17:20 *Systematic Analysis of Specific Biomarkers and Pathological Mechanism for Non-alcoholic steatohepatitis Based on Gene Expression Profiling of Multiple Progressive Stages*
Weili Lin, Qingchen Zhang, Dingfeng Wu, Shaohua Yuan, Lixin Zhu and Ruixin Zhu
Department of Bioinformatics, Tongji University, Shanghai, P.R. China
Paper ID: 102

17:20-17:40 *Goes Open——Publish with Genomics, Proteomics & Bioinformatics*
Yuxia Jiao
GPB Editorial Offics, Beijing Institute of Genomics, Chinese Academy of Sciences

17:40-18:00 *In silico analysis of mutations in PITX3 gene*
Abida Arshad, Rashda Abbasi, Christian Sieber, Muhammad Arshad and Nafees Ahmad
Department of Zoology, PMAS Arid Agriculture University, Rawalpindi, Pakistan
Paper ID: 29

18:30-20:00 Banquet

October 27 (Monday) Technical Sessions

Social Program: Half-day tour

08:00-18:00 One day excursion in Qingdao. Departure at Huiquan Dynasty Hotel Lobby

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

Book of Abstracts

October 10, 2014

Plenary Session P1   08:30-10:30   October 25 (Saturday)

08:30-09:10 *Proteomics and Bioinformatics – Diagnostic Approaches to the Diagnosis and Understanding of Kawasaki Disease*

Harvey J. Cohen

Children's Hospital, Stanford University School of Medicine, USA

Abstract：Kawasaki disease (KD) is an acute inflammatory disease of childhood that can be difficult to distinguish from other systemic febrile illnesses (FI). The etiologic agent(s) and pathophysiology remain poorly characterized. The acute vasculitis can be effectively treated with intravenous immunoglobulin (IVIG), but treatment must be initiated early in the course of the illness in order to be effective. Diagnosis of KD is based on clinical criteria with no specific laboratory diagnostic test, which can result in delayed or missed diagnosis. We utilized both plasma based proteomic investigations and quantitative analyses of KD-associated patterns of demographic, clinical and laboratory data to both improve diagnostic accuracy and investigate the nature of the disease. Using mass spectroscopy, we identified a fragment of serum amyloid A2 as a potential biomarker and are investigating its role in the pathogenesis of KD. We used statistical learning methods to develop and validate KD diagnostic algorithms, which can be applied to existing and evolving information technologies to potentially create novel and inexpensive point-of-care tools for the diagnosis of acute KD. The combination of linear discriminant analyses and random forest techniques has resulted in being able to distinguish KD from FI in 95% of the patients tested.

09:10-09:50 *Multi-Dimensional Cancer Genomics Studies in Systems Bioinformatics Perspective*

Sanghyuk Lee

Korean BioInformation Center, Korea Research Institute of Bioscience and Biotechnology, Korea

Abstract：Cancer is a complex systems disease with a number of genetic and epigenetic aberrations involved. As can be seen in the recent papers from the TCGA (The Cancer Genome Atlas) projects, multi-dimensional approach of combining genomic, transcriptomic, proteomic, and epigenomic data has proved to be a powerful method of identifying causal variations among numerous passenger mutations. In this talk, I will present the recent progress of our cancer genomics projects on T-cell lymphoma, lung cancer, and gastric cancer. Study design will be explained with a brief introduction for each project. Proper choice of algorithms and methods is critical in every stage of data analysis including variant calling, transcriptome quantification, and systems interpretation. Strategies to identify driver mutations in cancer will be illustrated for diverse situations of multi-dimensional omics data. Then I will provide an example of integrative analysis to identify causal variations and to obtain insights on disease mechanisms. A short perspective on future cancer genomics research will be given as well.

09:50-10:30 *TBD*

Raul Rabadan

Department of Biomedical Informatics, Center for Computational Biology & Bioinformatics, Columbia University College of Physicians & Surgeons, USA

Abstract：

Plenary Session P2   08:30-10:30   October 26 (Sunday)

08:30-9:10 *An Emerging Paradigm: New genes drive evolution of expression networks and phenotypes*

Manyuan Long

Edna K. Papazian Distinguished Service Professor, Department of Ecology and Evolution, The University of Chicago, USA

Abstract：How do genetic systems evolve to drive changes of functions and phenotypes of organisms? A new research paradigm is emerging:the gene networks are evolving systems that break evolutionary robustness, driven by addition of new genes and their changes. These new genes can reshape the genetic expression networks and rewire the biological pathways and thus impact the phenotypes and functions, which can bring up different gene systems in closely related species. Since the early reports of new genes-driven evolution

of gene networks in Drosophila, similar processes have been recently reported in other different organisms, ranging from nematodes to plants and mammals. These studies also raised new research directions with interesting scientific questions and are changing the ways to understand biological systems and their impacts on molecular functions to phenotypes of organisms.

Ref: Long M et al, 2013. Annu Rev Genet.

Chen et al, 2013. Nature Rev Genet.

09:10-09:50 *Translational Bionformatics: Past, Present, and Future*

Jessica Tenenbaum
Duke Translational Medicine Institute, USA

Abstract：

09:50-10:30 *Using blood multi-biomarkers could improve clinical significance*

Yaping Tian
Department of Clinical Biochemistry, Chinese PLA General Hospital, Beijing 100853

Abstract：Both western medicine and Chinese traditional medicine were believed that the healthy condition is the balance of the biological system. They have developed variety measuring system to monitoring the abnormal of the system so that the disease could be found and diagnosed. Nowadays, patients in hospital have to testing tens to hundreds parameters and the doctors have to use their knowledge to understand the meaning of each tests and then made their medical decision. With the clinical using of genomic and proteomic technology, much more information could be provided for each patient. The doctors would met more difficulty for knowing the clinical significance of these data and bio-informatics tool would be necessary in future to help the doctors and clinical workers to make their medical decisions. We have primarily doing some studies and trying to use the blood multi-biomarkers to improve the specificity and sensitivity of the laboratory tests and the results showed good prospect which could definitely enhance the clinical significance of the individual tests.

Session A1 10:50-12:30 October 25 (Saturday)

10:50-11:10 *The Translational Disease Network—from Protein Interaction to Disease Co-occurrence*

Hyunjung(helen) Shin, Yonghyun Nam, Dong-Gi Lee and Sunjoo Bang
Department of Industrial Engineering, Ajou University, South Korea

Paper ID: 13

Abstract: **Background:** In spite of the recent advances in understanding the human disease network, there may yet be a tendency for medical professionals to regard it as just the research of biologists. This may be due to obstacles in the expedition of the “bench-to-bedside” approach—“translating” a newfound discovery in biology into diagnostic/prognostic tools in medicine. One main obstacle may be related to the manner of representing the final outcomes of the research. In most cases, a disease network is provided as a map of topologies between diseases. This format is difficult for physicians to use in practice to deduce the probability of the co-occurrence of diseases. It would be more convenient if the outcomes were given in the form of scores or probabilities, which represent the likelihoods of diseases co-occurring with a primary disease. And, if the confusion caused by different nomenclatures is alleviated and if the experimental results from the bench are validated based on medical experiences from the bedside, it will bring us a step closer to truly “translative” interdisciplinary research, bridging the gap between biological research outcomes and medical applications. **Proposed Method:** We focus on a class of metabolic diseases (or disorders) that are highly prevalent at the population level, but have had little progress in the analysis of disease networks because of the lack of information on the associations between diseases. We construct a disease network based on protein-protein interaction (PPI) data and provide a network based scoring algorithm measuring the probabilities of disease co-occurrence. The former provides a disease network to which the latter applies. The proposed method for network construction is a novel and systematic approach that can embrace the methodolgies of previous studies; it draws latent associations between diseases from the PPI network by employing the notion of the walk in graph theory. To provide probabilities of co-occurring diseases, a new scoring algorithm is proposed that collects information on the latent association between diseases spread over the network, utilizing the well-established traits of graph-based semi-supervised learning. The entries in our disease network are extracted from the taxonomy of Medical Subject Headings, and the scoring results are validated using medical reports from the disease comorbidity study. **Results:** The proposed disease network substantially increases the connectivity between metabolic diseases, overcoming the sparse associations that had been identified in previous studies. The comparative analysis states that because of the richer connections made available using this network, the likelihood that a co-occurrence of disease will inferred when using the proposed scoring algorithm is higher for most diseases, including rare diseases. The probabilities for co-occurring diseases, determined by this network approach, appear to be concordant with the existing studies on disease comorbidity. Taken together, the comparative analysis offers strong support for the structural relevance of our disease network, and for the functional relevance of our disease-scoring algorithm.

11:10-11:30 *Studying the Genetics of Complex Diseases with Ethnicity-Specific Human Phenotype Networks: The Case of Type 2 Diabetes in East Asian Populations*

Jingya Qiu, Jason H. Moore and Christian Darabos
Institute for Quantitative Biomedical Sciences, Dartmouth College
Paper ID: 56

Abstract: In recent years, genome-wide association studies (GWAS) have led to the discovery of 200+ genetic variations (SNPs) at 100+ loci associated with type 2 diabetes mellitus (T2D). It was also observed that East Asians develop T2D at a higher rate, younger age, and lower body mass index than their European ancestry counterparts. The reason behind this occurrence remains elusive. With comprehensive searches through the National Human Genome Research Institute GWAS catalog literature, we man- ually curated over 2,800 ethnicity-specific SNPs associated with T2D and 70 other related traits. The GWAS catalog reports data such as p-value, odds ratio, and risk allele frequency for each SNP found to be associated with T2D with no consideration of ancestry. Many of the values are derived from combining initial and repli- cation samples from mixed populations. Analysis of all-inclusive data can be misleading, as not all variants are transferable across diverse populations. The extraction of ethnicity data allowed us to construct three population-specific Human Phenotype Networks (HPN), centered around T2D to quantitatively analyze and visualize the disparities in genetic variants between different ethnic groups. We studied the global properties of the networks and their relationships to one-another. Our study of interethnic differences in the genetic variants associated with T2D suggests the possibility of different pathways involved in the pathogenesis of T2D amongst different populations. We identified 99 SNPs highly significant to T2D, most initially discovered in Europeans and replicated in East Asians, suggesting shared biological pathways. Of the 99 SNPs, however, 21 were specific to East Asian populations but impossible to replicate in other cohorts. Furthermore, many SNPs showed significant differences in p-value and risk allele frequencies in studies of comparable size. For example rs2237892 in locus KCNQ1, a critical gene in insulin-secreting INS-1 cells, proved to be highly significant in East Asian population but not in Europeans. We also reported opposite cases. Using the network models as a tool to generate new interaction hypotheses in a clinical context, we identified population-specific links in the East Asian HPN and studied the genetic relationship between T2D and myocardial infarction and ovarian cancer.

11:30-11:50 *Inferring drug-disease associations based on known protein complexes*

Liang Yu, Jianbin Huang, Zhixin Ma, Jing Zhang, Yapeng Zou and Lin Gao
School of Computer Science and Technology, Xidian University
Paper ID: 60

Abstract: Inferring drug-disease associations is critical in unveiling disease mechanisms, as well as discovering novel functions of available drugs, or drug repositioning. Previous work is primarily based on drug–gene–disease relationship, which throws away many important information since genes execute their functions through interacting others. To overcome this issue, we propose a novel methodology that discover the drug-disease association based on protein complexes. Firstly, the integrated heterogeneous network consisting of drugs, protein complexes, and disease are constructed, where we assign weights to the drug-disease association by using probability. Then, from the tripartite network, we get the indirect weighted relationships between drugs and diseases. The larger the weight, the higher the reliability of the correlation. We apply our method to mental disorders and hypertension, and validate the result by using comparative toxicogenomics database. Our ranked results can be directly reinforced by existing biomedical literature, suggesting that our proposed method obtains higher specificity and sensitivity. The proposed method offers new insight into drug-disease discovery.

11:50-12:10 *Select and Label (S&L): a Task-Driven Privacy-Preserving Data Synthesization Framework*

Zhanglong Ji, **Xiaoqian Jiang**, Haoran Li, Li Xiong and Lucila Ohno-Machado
University of California, San Diego
Paper ID: 66

Abstract: Privacy is a big concern to the public but data sharing has tremendous societal benefits, especially in biomedicine. Existing model perturbation methods can only support a limit number of exploratory model construction before the privacy budget is depleted. On the other hand, most data synthesization approaches are not model specific, which have limited utility for any specific task. We developed a novel differentially private data synthesization framework called select and label (S&L), which can generate synthetic data to meet the classification need. The basic idea is to synthesize ambiguous points near decision boundary (i.e., of the classification model) to be weighted and labeled by a differentially private procedure that is optimized for the classification model. We applied our framework to kernel SVM models and demonstrated superior performance than existing approaches.

12:10-12:30 *A semi-tensor product approach for Probabilistic Boolean network*

Xiaoqing Cheng, Wai-Ki Ching and Yushan Qiu
The University if Hong Kong
Paper ID: 57

Abstract: Modeling genetic regulatory networks is an important issue in systems biology. Various models and mathematical formalisms have been proposed in the literature to solve the capture problem. The main purpose in this paper is to show that the transition matrix generated under semi-tensor product approach (Here we call it the probability structure matrix for simplicity)and the traditional approach(Transition probability matrix) are similar to each other.And we shall discuss three important problems in Probabilistic Boolean Networks (PBNs): the dynamic of a PBN, the steady-state probability distribution and the inverse problem. Numerical examples are given to show the validity of our theory. We shall give a brief introduction to semi-tensor and its application. After that we shall focus on the main results: to show the similarity of these two matrices. Since the semi-tensor approach gives a new way for interpreting a BN and therefore a PBN, we expect that advanced algorithms can be developed if one can describe the PBN through semi-tensor product approach.

Session A2 14:00-15:40 October 25 (Saturday)

14:00-14:20 *Identification of epigenetic modifications that contribute to pathogenesis in therapy-related AML: Effective integration of genome-wide histone modification with transcriptional profiles*

Xinan Yang, Bin Wang and John Cunningham
The University of Chicago, USA
Paper ID: 48

Abstract: **Background:** Therapy-related, secondary acute myeloid leukemia (t-AML) is an increasingly frequent complication of intensive chemotherapy. This malignancy is often characterized by abnormalities of chromosome 7, including large deletions or chromosomal loss. A variety of studies suggests that decreased expression of the EZH2 gene located at 7q36.1 is critical in disease pathogenesis. The histone methyltransferase EZH2 has been implicated in repression of transcription through modification of a lysine residue on histone H3 (H3k27). However, the critical target genes of EZH2 and their regulatory roles remain unclear. **Method:** To characterize the subset of EZH2 target genes that might contribute to t-AML pathogenesis, we developed a novel computational analysis to integrate tissue-specific histone modifications and genome-wide transcriptional regulation. Initial integrative analysis utilized a novel “seq2gene” strategy to link epigenetic modification of regulatory regions within 150kb of a coding gene to the gene’s transcription. By coupling these results with our Phenotype-Genotype-Network (PGnet) algorithm, we identified several “biomodules” (a group of genes that share similar expression patterns and genomic or functional characteristics). **Results:** We identified SEMA3A (semaphoring 3A) as a novel oncogenic candidate that is regulated by EZH2-silencing, using data derived from both normal and leukemic cell lines as well as murine cells deficient in EZH2. A microsatellite marker in SEMA3A promoter has been associated with chemosensitivity and radiosensitivity. Notably, our subsequent studies in primary t-AML demonstrate an expected up-regulation of SEMA3A that is EZH2-modulated. Furthermore, we have identified three biomodules that are co-expressed with SEMA3A and up-regulated in t-AML, one of which consists of previously characterized EZH2 repressed gene targets. The other two biomodules include MAPK8 and TATA box targets. Together, our studies suggest an important role for EZH2 targets in t-AML pathogenesis that warrant further study. **Conclusions:** We introduced three computational systems biology strategies below. First, a new seq2gene strategy in chromatin immuneprecipitation sequencing (ChIP-seq) analysis can largely explore potential target genes. Second, a novel application of the PGnet algorithm profoundly enriches genes with similar expression profile and functions. Finally, an integrative analysis on ChIP-seq data, gene expression data and gene function information successfully discovers an independently validated biomarker in t-AML. These developed computational algorithms and strategies will enhance the knowledge discovery and hypothesis-driven analysis of multiple next generation sequencing data, for t-AML and other complex diseases.

14:20-14:40 *Interpretation of personal genome sequencing data in terms of disease ranks based on mutual information*

Young-Ji Na, Kyung-Ah Sohn and Ju Han Kim
University of Pennsylvania, USA
Paper ID: 77

Abstract: The rapid advances in genome sequencing technologies have resulted in an unprecedented number of genome variations being discovered in humans. However, there has been very limited coverage of interpretation of the personal genome sequencing data in terms of diseases. In this paper we present the first computational analysis scheme for interpreting personal genome data by simultaneously considering the functional impact of damaging variants and curated disease–gene association data. This method is based on mutual information as a measure of the relative closeness between the personal genome and diseases. We hypothesize that a higher mutual information score implies that the personal genome is more susceptible to a particular disease than other

diseases. The method was applied to the sequencing data of 50 acute myeloid leukemia (AML) patients in The Cancer Genome Atlas. The utility of associations between a disease and the personal genome was explored using data of healthy (control) people obtained from the 1000 Genomes Project. The ranks of the disease terms in the AML patient group were compared with those in the healthy control group using “Leukemia, Myeloid, Acute” (C04.557.337.539.550) as the corresponding MeSH disease term. The mutual information rank of the disease term was substantially higher in the AML patient group than in the healthy control group, which demonstrates that the proposed methodology can be successfully applied to infer associations between the personal genome and diseases. Overall, the area under the receiver operating characteristics curve was significantly larger for the AML patient data than for the healthy controls. This methodology could contribute to consequential discoveries and explanations for mining personal genome sequencing data in terms of diseases, and have versatility with respect to genomic-based knowledge such as drug–gene and environmental-factor–gene interactions.

14:40-15:00 *Identifying network biomarkers by protein-protein interaction affinity derived from law of mass action*

Jingxue Xin, Xianwen Ren, Luonan Chen and Yong Wang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
Paper ID: 101

Abstract: Identifying effective biomarkers for complex diseases is an important but challenging task in biomedical research today. Molecular level data is abundant since microarray and high-throughput sequencing technologies makes mRNA expression data easy to be obtained. However the gap remains in phenotype level, i.e., cancer classification and diagnoses with high accuracy from these data sources are still challenging problems. In this paper, we propose a novel method for identifying network biomarkers based on protein-protein interaction affinity (PPIA). The basic philosophy is asking the new question that “which protein-protein interaction are differentially connected”. We firstly approximate PPIAs by estimating the concentration of protein complex based on mass action law. Then a small and non-redundant group of protein-protein interactions are selected based on their PPIA to maximize the ability to discerning case and control samples. This problem is mathematically formulated as a linear programming problem, which is computationally fast and guarantees a globally optimal solution. The proposed method exploits the activities of protein complexes which integrate the static protein-protein interaction information with dynamical gene expression data. In addition, our new method directly takes into account protein-protein interactions in biological processes to form effective network biomarker. We tested our method by several experimental datasets in breast cancer. Extensive results demonstrate the effectiveness and efficiency of the proposed method for identifying a group of protein-protein interactions as more biologically interpreted network biomarker with comparable accuracy.

15:00-15:20 *VaccineWatch : a monitoring system of vaccine messages from social media data*

Somrak Numnark, Supawadee Ingsriswang and Duangdao Wichadakul
National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand
Paper ID: 67

Abstract: To exploit social media data in vaccine-related areas, we proposed VaccineWatch, a monitoring system with visualizations and analytics of significant vaccine information from Twitter and RSS feeds. The system was designed and implemented as a web application with following distinguished features. First, it comes with graphical user interfaces that visualize perspectives of vaccine-related information mined from social media data. Second, it provides a set of filters allowing users to focus on their diseases, vaccines, countries, and/or companies of interest. Third, it includes the helper tools for the management of social media data collection and backend processes such as Twitter and RSS crawlers. The prototype of VaccineWatch is available at www.vacciknowlogy.org/VaccineWatch

15:20-15:40 *Differentially Private SNP selection in Genome-Wide Association Studies*

Haoran Li, Xiaoqian Jiang, Zhanglong Ji and Li Xiong
Emory University, USA
Paper ID: 90

Abstract: Privacy preserving data sharing in genome-wide association study (GWAS) has recently received considerable attention. Differential privacy has emerged as one of the most rigorous privacy guarantees. It protects the situation that an adversary knows complete information about all persons in the data set except one single individual. In this paper, we propose two advanced Laplace mechanisms based on genotype and allelic test contingency tables to select most relevant SNPs while guaranteeing differential privacy. Instead of directly adding Laplace noise to $\chi2$-statistics, we inject Laplace noise to cell counts of the allelic test contingency table, and use these noisy counts to compute the perturbed $\chi2$or allelic test statistics. We prove the privacy guarantee of both algorithms and analyze the convergence properties of the two perturbed test statistics when the number of samples tends to be infinity. Experiments using real and simulated data sets demonstrate that our advanced Laplace algorithms generate accurate top relevant SNPs with better utility than state-of-the-art techniques under different parameter setting.

Session A3 16:20-18:00 October 25 (Saturday)

16:20-16:40 *WebDISCO: a Web service for DIStributed COx model learning without patient-level data sharing*

Chia-Lun Lu, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, **Xiaoqian Jiang** and Lucila Ohno-Machado
Division of Biomedical Informatics, University of California San Diego, USA
Paper IDs: 49

Abstract: Abstract—The Cox proportional hazards model (a.k.a. Cox model) is one of the most popular statistical methods for studying survival data. Survival studies usually require large amounts of data to achieve sufficient statistical power. In the biomedical context, data sharing among institutions could provide this added power. However, privacy concerns can block efforts to collect data in a central repository. In this article, we propose a web service for distributed Cox model learning (WebDISCO) in which data does not leave its origin. WebDISCO processes sensitive data locally, exchanging only statistics to build a global Cox model. We conducted experiments using clinical registry data and compared model coefficients learned from the proposed WebDISCO model with those from the Cox model based on a central repository. Our results show that both models calculate near-identical model coefficients with differences in the range of $\mathbf{10^{-12}}$ to $\mathbf{10^{-15}}$, suggesting that the distributed version performs as well as the centralized one. WebDISCO provides an easy-to-use and computationally efficient web service for biomedical researchers to conduct distributed survival analysis. This web service is available at https://webdisco.ucsd-dbmi.org:8443/cox/

16:40-17:00 *Cross-Platform and Cross-Device Pedometer System Designed for Healthcare Services*

Yongjin Kwon, Rebekah Jiyoung Cha, Kyuchang Kang and Changseck Bae
Electronics and Telecommunications Research Institute, Korea
Paper IDs: 58

Abstract: Physical activity is closely related to one's health status. Especially the intensity of physi-cal activity is more important than other features for health benefits, which can be com-puted by the number of steps. With the advent of mobile devices, pedometer system can be implemented on mobile devices with their built-in sensors. However, due to the variety of types of platforms and devices, it is hard to ensure the consistency of step counting. In this paper, we propose a robust pedometer system for healthcare services, which ensures the consistent results of step counting upon heterogeneous platforms and multiple mobile devices. Based on the proposed system, we present the actual implementation of pedome-ter applications for different platforms and devices. We examine our implementation to verify that it is useful in real life with respect to the accuracy of step counting and battery consumption.

17:00-17:20 *A New Approach for Estimating the Progression of Pancreatic Cancer*

Shuhao Sun and Tianhai Tian
School of Mathematical Sciences, Monash University, Melbourne, Australia
Paper IDs: 74

Abstract: Cancer of the pancreas is a highly lethal disease and has an extremely poor prognosis. It is the fourth leading cause of death from cancer in the US and the twelfth worldwide. There are currently only few therapeutic options for patients with pancreatic cancer. Hence new insights into the pathogenesis of this lethal disease are urgently needed. In recent years, extensive biological research has been conducted to study the mechanisms that control the initiation and progression of pancreas cancer. Mathematical models have also been used to present quantitative analysis and predict reasonable time schemes for the progression of pancreatic cancer. However, in those published articles, it was assumed that the mutation rate was constant, which is not realistic. In this work, we present a new approach using nonconstant mutation rate and hence reveal several important biological parameters of cancer progression, such as initial mutation rate as well as doubling time (or selective advantage coefficients) in different stages, and eventually present a better time scheme. Under more realistic assumptions regarding gene mutation and a more reasonable mutation rate, the averaged values of doubling time and selective advantage coefficient generated by our model are consistent with the predictions made by the published models.

17:20-17:40 *Prediction of trans-acting siRNAs in human brain*

Xiaoshuang Liu, Guangxin Zhang, Changqing Zhang and Jin Wang

Nanjing University, China
Paper IDs: 46

Abstract: Endogenous small non-coding RNAs have been found to play pivotal roles in regulating gene expression in eukaryotes. While huge interests are put into the function and molecular mechanism of microRNAs in the development and disease of various organisms via the repression of mRNA of protein coding gene, new discoveries indicate that they may trigger the generation of another type of small RNAs, the trans-acting siRNAs (ta-siRNAs). This implies a new mode of RNA function, the regulation between small RNAs which gives rise to an even more complicate and elaborate pattern of RNA regulation mechanism for gene expression. We proposed a method for mining ta-siRNA sequences according to the generating process of this type of small RNA. The performance of the method was evaluated on Arabidopsis data. Using the human brain small RNA and degradome data, 155 small RNAs were found that satisfy the ta-siRNA characteristics. Furthermore, DRAXIN and ATCAY genes which preferentially expressed in brain were predicted to be the targets of 12 potential ta-siRNAs.

17:40-18:00 *Comparison of Multi-Sample Variant Calling Methods for Whole Genome Sequencing*

Kwangsik Nho, John West, Huian Li, Robert Henschel, Apoorva Bharthur, Michel Tavares and Andrew Saykin
Indiana University School of Medicine, USA
Paper ID: 32

Abstract: Rapid advancement of next-generation sequencing (NGS) technologies has facilitated the search for genetic susceptibility factors that influence disease risk in the field of human genetics. In particular whole genome sequencing (WGS) has been used to obtain the most comprehensive genetic variation of an individual and perform detailed evaluation of all genetic variation. To this end, sophisticated methods to accurately call high-quality variants and genotypes simultaneously on a cohort of individuals from raw sequence data are required. On chromosome 22 of 818 WGS data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), which is the largest WGS related to a single disease, we compared two multi-sample variant calling methods for the detection of single nucleotide variants (SNVs) and short insertions and deletions (indels) in WGS: (1) reduce the analysis-ready reads (BAM) file to a manageable size by keeping only essential information for variant calling (“REDUCE”) and (2) call variants individually on each sample and then perform a joint genotyping analysis of the variant files produced for all samples in a cohort (“JOINT”). JOINT identified 515,210 SNVs and 60,042 indels, while REDUCE identified 358,303 SNVs and 52,855 indels. JOINT identified many more SNVs and indels compared to REDUCE. Both methods had concordance rate of 99.60% for SNVs and 99.06% for indels. For SNVs, evaluation with HumanOmni 2.5M genotyping arrays revealed a concordance rate of 99.68% for JOINT and 99.50% for REDUCE. REDUCE needed more computational time and memory compared to JOINT. Our findings indicate that the multi-sample variant calling method using the JOINT process is a promising strategy for the variant detection, which should facilitate our understanding of the underlying pathogenesis of human diseases.

Session A4 10:50-12:30 October 26 (Sunday)

10:50-11:10 *Graph Pyramid Approach for Protein Classification*

Tushar Sandhan, Youngjoon Yoo, Jin Young Choi and Sun Kim
Seoul National University, Korea
Paper ID: 96

Abstract: Uncovering the hidden organizational characteristics and regularities among biological sequences, is the key issue for detail understanding of an underlying biological phenomenon. Conventional homology based protein function prediction via classification approaches mostly rely on the Global Features (GF) by considering only strong protein similarity matches. Here we construct the Protein-Protein Similarity (PPS) network, which captures the subtle properties of the protein families. The proposed method considers the Local Features (LF) as well as the GF, by considering the interactions among weakly interacting proteins in PPS network and by using the hierarchical graph analysis via graph pyramid. Different underlying properties of the protein families are uncovered by operating the proposed graph based features at various pyramid levels. With each correctly classified test sequence, the fast incremental learning ability of the proposed method further boost up the training model. Experimental results show that the proposed hierarchical voting algorithm using graph pyramid, helps to improve computational efficiency as well as the protein classification accuracy.

11:10-11:30 *Relating Hepatocellular Carcinoma Tumor Samples and Cell Lines Using Gene Expression Data in Translational Research*

Bin Chen, Marina Sirota, Hua Fan-Minogue, Dexter Hadley and Atul Butte
Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, USA
Paper ID: 82

Abstract: Cancer cell lines are used extensively to study cancer biology and to test hypotheses in translational research. The relevance of cell lines is dependent on how closely they resemble the tumors being studied. Relating tumors and cell lines, and recognizing their similarities and differences are thus very important for translational research. Rapid advances in genomics have led to the generation of large volumes of genomic and transcriptomic data for a diverse set of primary cancer samples, normal tissue samples and cancer cell lines. Hepatocellular Carcinoma (HCC) is one of the most common tumors worldwide, with high occurrence in Asia and sub-Saharan regions. The current effective treatments of HCC remain limited. In this work, we compared the gene expression measurements of 200 HCC tumor samples from The Cancer Genome Atlas and over 1000 cancer cell lines including 25 HCC cancer cell lines from Cancer Cell Line Encyclopedia. We showed that the HCC tumor samples correlate closely with HCC cell lines in comparison to cell lines derived from other tumor types. We further demonstrated that the most commonly used HCC cell lines resemble HCC tumors, while we identified nearly half of the cell lines that do not resemble primary tumors. Interestingly, a substantial number of genes that are critical for disease development or drug response are either expressed at low levels or absent among highly correlated cell lines; additional attention should be paid to these genes in translational research. Our study will be used to guide the selection of HCC cell lines and pinpoint the specific genes that are differentially expressed in either tumors or cell lines.

11:30-11:50 *Concordance between ex vivo PBMC and in vivo human infections confirmed by N-of-1-pathways analysis of single-subject transcriptome*

Vincent Gardeux, Anthony Bosco, Jianrong Li, Fernando D. Martinez and Yves A Lussier
Department of Medicine, University of Arizona, Tucson, AZ, USA
Paper ID: 116

Abstract: **Background.** Understanding individual patient host-response to viruses is key to designing optimal personalized therapy. Unsurprisingly, in vivo human experimentation to understand individualized dynamic response of the transcriptome to viruses are rarely studied because of the obviously limitations stemming from ethical considerations of the clinical risk. In this rhinovirus study, we first hypothesize that ex vivo human cells response to virus can serve as proxy for otherwise controversial in vivo human experimentation. Of note, comparing the fold change of a few paired measures is the state of the art in human ex vivo assays, which does not scale up to genomics measurements due to excess false positive results. We further hypothesized that the N-of-1-pathways framework, previously validated in cancer, can be effective in understanding the more subtle individual genomic response to viral infection. N-of-1-pathways framework could provide such insight as it is designed to identify deregulated pathways from ontology-anchored gene sets in two paired samples of genome-scale measurements. Finally, we also developed a novel visualization method, similarity Venn Diagram, that provides the similar results between two sets of qualitative measures that can be compared by similarity metrics (e.g. ontology, information theoretic distance, etc). **Method.** N-of-1-pathways computes a significance score for a list of given genesets, using the 'omics profiles of a mere two samples as input (e.g. normal/tumoral, pre/post-treatment, infected vs non infected cells). We extracted the peripheral blood mononuclear cells (PBMC) of four human subjects, aliquoted in two paired samples one subjected to ex vivo rhinovirus infection. Their deregulated genes and pathways were compared quantitatively and qualitatively as a group to those of 9 human subjects prior and after intranasal inoculation "in vivo" with rhinovirus. We then clustered individual N-of-1-pathways scores to demonstrate that these profiles recapitulated the phenotypes of asymptomatic and symptomatic patients. Additionally, we developed the Similarity Venn Diagram, an efficient and deceptively simple method for comparing results expressed in an ontology organized as a directed acyclic graph. **Results.** We compared the N-of-1-pathways results using two established cohort-level methodologies: GSEA and enrichment of differentially expressed genes. Methodologically, we have extended contingency tables and odds ratio calculation to calculating the significance of Similarity Venn Diagrams. Results are biologically relevant and similar between in vivo and ex vivo studies, both at the genes and enriched pathways levels. Individual patient ROC curves demonstrate that deregulated pathways identified by N-of-1-pathways in PBMC cells of each single subject infected ex vivo recapitulate the biologically relevant pathways observed in vivo in a whole cohort (p=0.004). Further, a principal component analysis of N-of-1-Pathways Scores discriminates asymptotic patients from symptomatic infected patients in vivo (PBMC expression). **Conclusion.** There are less than five published transcriptomes of human viral infections in vivo. We show the first evidence that a novel transcriptome analysis of ex vivo essays has the potential to predict individualized response to infectious disease without the clinical risks otherwise associated to in vivo challenges.

Software: http://Lussierlab.org/publications/N-of-1-pathways
Supplement data and files: http://Lussierlab.org/publications/Ex-vivo-ViralAssay

11:50-12:10 *Neural fate decisions mediated by oscillatory and sustained Hes1*

Shanshan Li, Zengrong Liu and Ruiqi Wang
Institute of systems biology, Shanghai University

Paper ID: 121

Abstract: During central nervous system (CNS) developing, Hes1 shows short period oscillations in progenitor cells, while stable low levels in neurons. The reason why diverse expression modes of Hes1 exist remains unknown. Here, we develop a mathematical model involving Hes1 and BM88, with the aim of understanding the complex molecular mechanism that orchestrates the processes of neural fate decision. Our simple but fundamental model can account for both Hes1 oscillations observed in neural progenitors and Hes1 regulation to BM88 in differentiation progress. Our results suggest that a relatively simple network is capable of accounting for some fundamental principles in progenitor maintenance and differentiation.

12:10-12:30 *Identify Critical Genes in Development with Consistent H3K4me2 Patterns across Multiple Tissues*

Nan Meng, Raghu Machiraju and **Kun Huang**
The Ohio State University, USA
Paper ID: 112

Abstract: Histone modification is an important epigenetic event which plays essential roles in cell differentiation and tissue development. Recent studies show that a unique dimethylation of lysine 4 residue on histone 3 (H3K4me2) distribution pattern around transcription starting sites (TSS) of genes marks tissue specific genes in human CD4+ T cells and mouse nervous tissue cells. However, existence of this pattern has not been widely tested in other tissue types and the implication of this pattern remains unclear. In this paper, we study the H3K4me2 distribution patterns across six different cell lines from five major tissue types (including muscular tissue, nervous tissue, non-blood connective tissue, blood, and epithelial tissue) as well as embryonic stem cells. We define a metric Œtail length^1 to quantitatively describe H3K4me2 distribution patterns around the TSS. While we have confirmed the observation that genes with long H3K4me2 tails around TSS are enriched with tissue specific functions, we also identified a group of 217 genes with ubiquitous long-tail H3K4me2 patterns in all the tested tissues as well as the embryonic stem cells (ESC). Since we have observed that the long-tail H3K4me2 pattern is often associated with high transcription activity, we hypothesize that these genes are active in multiple developmental stages and are of significant importance in tissue differentiation and development. Functional enrichment analysis confirmed that these genes are critical for development. Further analysis shows that genes in this group are highly interactive with other tissue specific genes as evinced by protein-protein interaction networks, suggesting their critical regulatory functions. Our results suggest that rich information on gene functions and epigenetic events can be revealed using pattern recognition methods.

Session A5 14:00-15:40 October 26 (Sunday)

14:00-14:20 *Evolutionary Pressures on the Yeast Transcriptome*

Dominique Chu and Anton Salykin
School of Computing, University of Kent, CT2 7NF, Canterbury, United Kingdom
Paper 9

Abstract: Codon usage bias (CUB) is the well known phenomenon that the frequency of synonymous codons is unequal. This is presumably the result of adaptive pressures favouring some codons over others. The underlying reason for this pressure is unknown, although a large number of possible driver mechanisms have been proposed. According to one hypothesis, the decoding time could be such a driver. The standard model for decoding speeds is the Gromadski-Rodnina model according to which decoding speed is determined by the ratio of cognate and non-cognate tRNAs. Recently, there have been a number of contributions in the literature arguing to the effect that this conventional speed-model is not relevant. Here we present an analysis of the Saccharomyces cerevisiae genome to check for selection pressures on the sequences. We choose measures that are directly based on the Gromadski-Rodnina model, i.e. if this model were irrelevant we would expect not to see any consistent selection pressures given these measures. We compare yeast transcripts with randomly generated synonymous codon sequences and compare key-features relating to the CUB, including decoding speed, the propensity for traffic jams and sequence homogeneity. We find that the yeast transcriptome displays strong adaptive signatures with respect to these measures, corroborating the relevance of the Gromadski-Rodnina model. Especially, we show that over 70% of ORFs have been subject to a strong selection pressure for translation speed and that there is also a strong selection pressure for the avoidance of traffic jams. Finally, both homogeneous and very heterogeneous transcripts are over-represented.

14:20-14:40 *Systematic identification of local structure binding motifs in protein-RNA recognition*

Zhi-Ping Liu
Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China
Paper ID: 52

Abstract: Many critical biological processes are strongly related to protein-RNA interactions. Revealing the structure motifs of performing protein-RNA binding function will provide valuable information for deciphering their interaction mechanisms and benefit complementary structure designs in bioengineering. In this work, we provide a study of systematic identification of protein structure motifs of RNA-binding sites in form of pockets on protein surfaces by clustering these local structure patterns into similar groups. We also identify the crucial recognition patterns and the structural complementary features in the protein-RNA binding events.

14:40-15:00 *RCARE: RNA Sequence Comparison and Annotation for RNA Editing*

Sooyoun Lee, Je Gun Joung, Chan Hee Park, Ji Hye Park and Ju Han Kim
Seoul National University College of Medicine, Korea

Paper ID: 47

Abstract: The posttranscriptional sequence modification of transcripts through RNA editing is an important mechanism for regulating protein function and is associated with human disease phenotypes. The identification of RNA editing or RNA-DNA difference (RDD) sites is a fundamental step in the study of RNA editing. However, a substantial number of false-positive RDD sites have been identified recently. A major challenge in identifying RDD sites is distinguishing between the true RNA editing sites and the false positives. Furthermore, determining the location of condition-specific RDD sites and elucidating their functional roles will help toward understanding various biological phenomena that are mediated by RNA editing. The present study investigated the use of RNA-sequence comparison and annotation for RNA editing (RCARE) for searching, annotating, and visualizing RDD sites using thousands of previously known editing sites, which can be used for comparative analyses between multiple samples. RCARE also provides evidence for improving the reliability of identified RDD sites. RCARE is a web-based comparison, annotation, and visualization tool for RNA editing research. It provides rich biological annotations and useful summary plots as well as the evidence level of each RNA editing site. Sequence-based alignment files can be converted into VCF files using a Python script and uploaded to the RCARE server for further analysis. RCARE is available for free at http://www.snubi.org/software/rcare/

15:00-15:20 *Measuring the Similarity of Protein Structures Using Image Local Feature Descriptors SIFT and SURF*

Morihiro Hayashida, Hitoshi Koyano and Tatsuya Akutsu
Bioinformatics Center, Institute for Chemical Research, Kyoto University

Paper ID: 100

Abstract: Understanding of protein structures is important to find their functions. Many methods such as structural alignment, alignment-free similarity, and use of structural fragments have been developed for finding similar protein structures. In our previous study, we transformed protein structures into images each pixel of which represents the distance between the corresponding C-alpha atoms, and proposed similarity measures between two protein structures based on Kolmogorov complexity using image compression algorithms. In this paper, we examine efficient and effective image recognition techniques, SIFT (Scale Invariant Feature Transform) and SURF (Speeded Up Robust Features), which are invariant to image scaling, translation, and rotation, and partially invariant to affine or three-dimensional projection. We propose similarity based on SIFT and SURF, and apply it to classification of several protein structures. The results suggest that the similarity based on SURF outperforms several existing similarity measures including the compression-based similarity measures in our previous study, and that SIFT and SURF are useful for recognizing protein structures as well as objects in images.

15:20-15:40 *The mathematical model and simulationg of predicting the non-compact conformations on triangle lattice*

Yuzhen Guo, Yong Wang and Zikai Wu
Department of Mathematics , Nanjing University of Aeronautics and Astronautics

Paper ID: 42

Abstract: In this paper, we focus on a studied of two dimensional hydrophobic-polar on triangle lattice. Non-compact conformation tries to fold the amino acids sequence into a relatively larger triangle lattice, which is more biologically realistic and significant than the compact conformation. Here, we established a mathematical model and a heuristic algorithm to predict the non-compact conformations. First, the protein structure prediction problem was abstracted to match amino acids to lattice points. The problem was then formulated as an integer programming model and we transformed the biological problem into an optimization problem. Classical particle swarm optimization algorithm was extended by the single point adjustment strategy to solve this problem. Compared with conformations on the square lattice, conformations on triangle lattice are more flexible in several benchmark examples.

Session A6 16:20-18:00 October 26 (Sunday)

16:20-16:40 *Detection of Core Cancer Modules by Mutated Gene Network in Glioblastoma*

Li Feng, Gao Lin and Yang Xiaofei
Xidian University
Paper ID: 21

Abstract: Understanding of the pathogenesis of cancer based on genomic data can provide a guidance of clinical diagnostics, prognostics and therapeutics of cancer. An important challenge for this problem is to distinguish “driver mutations” from “passenger mutations”. Considering the heterogeneity of the mutations and the fact that driver mutations always target pathways together, it is necessary to explain the development and molecular mechanisms of cancer in pathway level. In this paper we introduce a network-based method by quantifying how much two genes are involved in the same module to construct a mutated gene network, which is used for identifying core cancer modules. We consider the genes involved in core modules to be driver genes. This measure relies on two properties of mutations in a driver pathway: high coverage and mutual exclusivity. Our method does not limit the size of modules and does not depend on protein-protein interaction network data. This method is applied to somatic mutations of glioblastoma and identifies significant core cancer modules that intersect with known pathways, such as p53, RB, MAPK, cell cycle, CTCF and PI(3)K signaling pathways. Compared with MEMo and Multi-Dendrix, our method discovers both reported and unreported cancer modules by these two methods. It also finds some new genes that are rarely reported related to cancer. Consequently, our method can provide new related driver cancer genes for biologists for further study.

16:40-17:00 *MapIn: an interactive tool for mapping biological descriptors to ontologies*

Panwen Wang, Jun Li, Xiaorong Liu, Pak Sham and Junwen Wang
The University of Hong Kong
Paper ID: 36

Abstract: Mapping biological descriptors to a set of related concepts (e.g., gene ontology) can benefit the classification, comparison, retrieving and analysis of the descriptors. However, to guarantee accuracy, mapping processes tend to require manual intervention, and is therefore challenging to completely automate. Here we developed an Interactive Mapping tool, MapIn, (available both online and as standalone application), for users to map biological descriptors to a set of related concepts semi-automatically. MapIn offers mapping suggestions via a built-in algorithm to calculate the similarities between the descriptors and the related concepts. If preferred, users can implement their own similarity calculation algorithms by our Application Programming Interface in the standalone version. Here we demonstrate the utility and efficiency of MapIn in mapping Genome-Wide Association Study Catalog traits to Human Phenotype Ontology and Medical Subject Headings. The web server and standalone program of MapIn is freely available at http://jjwanglab.org/mapin.

17:00-17:20 *Mining Correlation Patterns of Taxa, Pathways and Environmental Factors with An Improved Weighted Network Community Detection Algorithm*

Xiao-Ying Yan, Shao-Wu Zhang and Ze-Gang Wei
College of Automation, Northwestern Polytechnical Univerwsity, China
Paper ID: 83

Abstract: 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. Generally, the functional capability and taxa structure are highly related with environment factors in microbial communities, which are hidden in these large amount sequences. However, most works used the canonical correlation analysis (CCA) method to research the correlative relationship among taxa, pathways and environmental factors. CCA can be difficult to find which environmental factors are the major determinants of some special taxa and pathway. In this paper, we integrated 14 ocean metagenomes with geographical, meteorological and geophysicochemical data to construct the correlative weighted networks with Spearman correlation. By using an improved weighted network community detection algorithm, named as IWNCD, we find some special correlation patterns of taxa, pathways and environmental factors. Analysis of these patterns shows that the climatic factors such as temperature, sunlight, and correlated CO2, and the nutrients such as chlorophyII and primary production are the main determining factors of the functional community composition; The growth and development of some special taxa are dependent on some main environmental factors such as sunlight, temperature, CO2, primary production, dissolved oxygen, dissolved silicate; In addition, sampling sites more similar in geographic location have a greater tendency to be closer together based on their metabolic pathways.

17:20-17:40 *TBD*

Ling Wang
CloudScientific Technology Co., Ltd , China

Session B1 10:50-12:30 October 25 (Saturday)

10:50-11:10 *Functional dyadicity and heterophilicity of gene-gene interactions in statistical epistasis networks*

Ting Hu, Angeline Andrew, Margaret Karagas and **Jason Moore**
Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover NH, USA
Paper ID: 10

Abstract: The interaction effect among multiple genetic factors, i.e.~epistasis, plays an important role in explaining susceptibility on common human diseases and phenotypic traits. The uncertainty over the number of genetic attributes involved in interactions poses great challenges in genetic association studies and calls for advanced bioinformatics methodologies. Network science has gained popularity in modeling genetic interactions thanks to its structural characterization of large number of entities and their complex relationships. However, little has been done on functionally interpreting statistically inferred epistatic interactions using networks. In this study, we used Gene Ontology (GO) to functionally annotate genes as vertices in a statistical epistasis network, and quantitatively characterize the correlation between the distribution of gene functional properties and the network structure by measuring dyadicity and heterophilicity of each functional category in the network. These two parameters quantify whether genetic interactions tend to occur more likely on genes from the same functional category, i.e.~dyadic effect, or more likely on genes from across different functional categories, i.e.~heterophilic effect. By applying this framework to a population-based bladder cancer dataset, we were able to identify several GO categories that have significant dyadicity or heterophilicity associated with bladder cancer susceptibility. Thus, our informatics framework suggests a new methodology on embedding functional analysis in network modeling of statistical epistasis in genetic association studies.

11:10-11:30 *Novel therapeutics for coronary artery disease from genome-wide association study data*

Mani P. Grover, Sara Ballouz and Merridee Wouters
School of Medicine, Deakin University, Geelong, Victoria, Australia.
Paper ID: 68

Abstract: **Background:** Coronary artery disease (CAD), one of the leading causes of death globally, is influenced by both environmental and genetic risk factors. Gene-centric genome-wide association studies (GWAS) involving cases and controls have been remarkably successful in identifying genetic loci contributing to CAD. Modern in silico platforms, such as candidate gene prediction tools, permit a systematic analysis of GWAS data to identify candidate genes for complex diseases like CAD. Subsequent integration of drug-target data from drug databases with the predicted candidate genes for CAD can potentially identify novel therapeutics suitable for repositioning towards treatment of CAD. **Methods**: In previous work, we used Gentrepid (www.gentrepid.org) as a candidate gene prediction platform to identify 647 candidate genes for CAD using Wellcome Trust Case-Control Consortium GWAS data. Without the aid of the additional CAD cases, Gentrepid successfully predicted 55% of the candidate genes identified by the more powerful CARDIoGRAMplusC4D consortium meta-analysis. Hence, Gentrepid was capable of enhancing lower quality genotype-phenotype data of gene function, using a knowledgebase of existing biological data. Here, we extended our methodology by integrating drug data from three drug databases: the Therapeutic Target Database, PharmGKB and Drug Bank; with candidate gene predictions from Gentrepid. We utilized known CAD targets and the scientific literature as benchmarks to validate Gentrepid specific predictions for CAD. **Results:** Our analysis identified a total of 184 predicted candidate genes as novel therapeutic targets for CAD, and 981 novel therapeutics feasible for repositioning in clinical trials towards treatment of CAD. Two benchmarks showed that our results were significant ($p < 0.05$). **Conclusions:** We have demonstrated that currently available drugs may potentially be repositioned as novel therapeutics for the treatment of CAD. Drug repositioning can save valuable time and money spent on preclinical and phase I clinical studies.

11:30-11:50 *Detection and Analysis of Disease-associated Single Nucleotide Polymorphism Influencing Post-translational Modification*

Yul Kim, Chiyong Kang, Bumki Min and Gwan-Su Yi
Dept. of Bio and Brain Engineering, KAIST, South Korea
Paper ID: 69

Abstract: Post-translational modification (PTM) plays a crucial role in biological functions and corresponding disease developments. Discovering disease-associated non-synonymous SNPs (nsSNPs) altering PTM sites can help to estimate the various PTM candidates involved in diseases, therefore, an integrated analysis between SNPs, PTMs and diseases is necessary. However, only a few types of PTMs affected by nsSNPs have been studied without considering disease-association until now. In this study, we developed a new database called PTM-SNP which contains a comprehensive collection of human nsSNPs that affect PTM sites, together with disease

information. Total 179,325 PTM-SNPs were collected by aligning missense SNPs and stop-gain SNPs on PTM sites (position 0) or their flanking region (position -7 to 7). Disease-associated SNPs from GWAS catalogs were also matched with detected PTM-SNP to find disease associated PTM-SNPs. Our result shows PTM-SNPs are highly associated with diseases, compared with other nsSNP sites and functional classes including near gene, intron and so on. PTM-SNP can provide an insight about discovering important PTMs involved in the diseases easily through the web site. PTM-SNP is freely available at http://gcode.kaist.ac.kr/ptmsnp.

11:50-12:10 *Detecting Gene-Gene Interactions Using a Permutation-based Random Forest Method*

Jing Li, James Malley and Jason Moore
Dartmouth College, USA
Paper ID: 107

Abstract: Identifying gene-gene interactions is essential to understand disease susceptibility and to detect genetic architectures underlying complex diseases. Here, we aim at developing a permutation-based methodology relying on a machine learning method, random forest, to detect gene-gene interactions. We named our approach permuted random forest (pRF); it identifies the top interacting single nucleotide polymorphism (SNP) pairs by estimating how much power a random forest classification model is influenced by removing pairwise interactions. We systematically tested our approach on a simulation study with datasets possessing various genetic constraints including heritability, number of SNPs, sample size, etc. Our methodology showed high success rates for detecting the interaction SNP pair. We also applied our approach to two bladder cancer datasets, which showed consistent results with well-studied methodologies, such as multifactor dimensionality reduction (MDR) and statistical epistasis network (SEN). Furthermore, we built permuted random forest networks (PRFN), in which we used nodes to represent SNPs and edges to indicate interactions.

12:10-12:30 *SUMORESLER, a bioinformatics approach for identifying SUMOylation sites by combining sequence, structural and functional features*

Jinlei Zhang, Yang Zhang, Fuyi Li, Mingjun Wang, Geoffrey Webb, Chen Li, Jiangning Song
College of Information Engineering, Northwest A&F University, Yangling, 712100, China
Paper ID: 89

Abstract: SUMOylation is one of the important types of post-translational modifications (PTMs) in eukaryotic cells, which plays an essential role in myriad cellular processes ranging from protein folding and maturation to signal transduction. However, SUMOylation sites are commonly identified by experimental approaches, which are laborious and expensive. As an alternative, bioinformatics approaches are cost effective and can be used in a high-throughout manner to predict and prioritize potential SUMOylation substrates and sites. In this study, we propose SUMORESLER (SUMO RESidue LEarneR), a bioinformatics approach for identifying SUMOylation sites for three species H. sapiens, M. musculus and S. cerevisiaehave by integrating heterogeneous sequence, structural and functional features. We apply a two-step feature selection method to filter redundant and irrelevant features and select a condensed feature subset. Benchmarking experiments using five-fold cross-validation tests indicate that SUMORESLER achieves a competitive prediction performance compared with another two existing tools seeSUMO and GPS-SUMO. SUMORESLER is anticipated to be a useful approach for in silico identification of novel feSUMOylation substrates and sites.

Session B2 14:00-15:40 October 25 (Saturday)

14:00-14:20 *PDEGEM: Modeling non-uniform read distribution in RNA-seq data*

Xia Yuchao, Wang Fugui, Qian Minping, Qin Zhaohui and Deng Minghua
The Center of Quantitive biology Peking University, China
Paper ID: 16

Abstract: RNA-Seq is a powerful new technology to comprehensively analyze the transcriptome of any given cells. An important task in RNA-Seq data analysis is quantifying the expression levels of all transcripts. Although many methods have been introduced and much progress has been made, a satisfactory solution remains to be elusive. In this article, we borrow the idea from the Positional Dependent Nearest Neighborhood (PDNN) model, originally developed for analyzing microarray data, to model the non-uniformity of read distribution in RNA-seq data. We propose a robust nonlinear regression model named PDEGEM, a Positional Dependent Energy Guided Expression Model to estimate the abundance of transcripts. Using real data, we find that the PDEGEM fits the data better than mseq in all three real datasets we tested. We also find that the expression measure obtained using PDEGEM showed higher correlation with that obtained from alterative assays for quantifying gene and isoform expressions. Based on these results, we believe that our PDEGEM can improve the accuracy in modeling and estimating the transcript abundance and isoform expression in RNA-Seq data. Additionally, although the stacking energy and positional weight of the PDEGEM are relatively related to sequencing platforms and species, they share some common trends, which indicates thatthe PDEGEM could partly reflect the mechanism of DNA binding between the template strain and the new synthesized read.

14:20-14:40 *The NGS markup language(NGSML): a general medium for representation and exchange of NGS data*

Chunjiang Yu, Wentao Wu and Bairong Shen
Center for Systems Biology, Soochow University, 215006, Suzhou,China
Paper ID: 28

Abstract: **Background:** With the rapid development of NGS, more and more software and tools are produced to analyze these data. However, data exchange between these tools is very inconvenient for users. There is no general format for representation and exchange of NGS data now. At the same time, a variety of new databases have been built to store the increasing number of data. Since the majority of tools for next generation sequencing defined their own formats, data exchange between these tools is very inconvenient. If there is a general format that the different database can store in this format, then data exchange will be easier, and a tool can convert between NGS data format, then the existing software and tools can handle different format data conveniently. **Methods:** To design a general format that can store different NGS data, we collected 14 NGS databases from PubMed. We discussed 27 data formats used by NGS databases and analyze the most used formats. We absorbed the merits of these formats and design a new XML-based format (NGSML). We also developed software entitled MSGMLEditor, which is designed to create, edit and convert NGSML file. **Results:** We defined a general XML-based NGSML format for representation and exchange of NGS data. We also provided a user-friendly GUI software NGSMLEditor(NGS XML Editor) for creating, editing and converting NGSML file and a NGS-FC(Next Generation Sequencing Format Convert) tool integrated in NGSMLEditor to convert formats and retrieve next generation sequencing data from databases. **Conclusions:** The NGSML format which we designed can represent and exchange data for the most of NGS databases. Our format can be used to increase the amount of information that it can store and let the format more flexibility. NGSMLEditor make NGS data exchange more easily. The data, software and supplementary files for NGSML are deposited at http://sysbio.suda.edu.cn/NGSML.

14:40-15:00 *Application of Meta-Mesh on the analysis of microbial communities from human associated-habitats*

Xiaoquan Su, Gongchao Jing, Shi Huang, Jian Xu and Kang Ning
Bioinformatics Group of Single-Cell Center, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences
Paper ID: 37

Abstract: With the current fast accumulation of microbial community samples and related metagenomic sequencing data, data integration and analysis system is urgently needed for in-depth analysis of large scale metagenomic samples (also referred to as “microbial communities”) of interest. Although several existing databases have collected a large number of metagenomic samples, they mostly serve as data repositories with crude annotations, and offer limited functionality for analysis. Moreover, the few available tools for comparative analysis in the literature could only support the comparison of a few pre-defined set of metagenomic samples. To facilitate comprehensive comparative analysis on large amount of diverse metagenomic samples, we have designed a Meta-Mesh system for a variety of analyses including quantitative analysis of similarities among microbial communities and computation of the correlation between the meta-information of these samples. We have used Meta-Mesh for systematically and efficiently analyses on diverse sets of human associate-habitat microbial community samples. Results have shown that Meta-Mesh would serve as an efficient data analysis platform for discovery of clusters, biomarker and other valuable biological information from a large pool of human microbial samples.
.

15:00-15:20 *Bi-objective Optimization of a Continuous Biological Process*

Gongxian Xu, Ying Liu, Chao Yu and Dan Su

Department of Mathematics, Bohai University
Paper ID: 72

Abstract: This paper addresses the bi-objective optimization of continuous bio-dissimilation process of glycerol to 1, 3- propanediol. A bi-objective optimization model is firstly proposed to maximize the production rate of 1, 3-propanediol, simultaneously maximize the conversion rate of glycerol and ensure the bioprocess is operated under steady-state conditions. Then this bi-objective problem can be transformed into a sequence of single objective problems by using the weighted-sum and normal-boundary intersection methods respectively. Finally, these single objective problems are solved by an interior point method. The results show that the weighted-sum and normal-boundary intersection methods can obtain the approximate Pareto-optimal set of the proposed bi-objective optimization problem.

15:20-15:40 *Predicting censored survival data based on the interactions between meta-dimensional omics data in* *breast cancer*

Dokyoon Kim, Ruowang Li, Scott Dudek and Marylyn Ritchie
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, USA
Paper ID: 79

Abstract: **Background:** Evaluation of survival models to predict cancer patient prognosis is one of the most important areas of emphasis in cancer research. A binary classification approach has difficulty directly predicting survival due to the characteristics of censored observations and the fact that the predictive power depends on the threshold used to set two classes. In contrast, the traditional Cox regression approach has some drawbacks in the sense that it does not allow for the identification of interactions between genomic features, which could have key roles associated with cancer prognosis. In addition, data integration is regarded as one of the important issues in improving the predictive power of survival models since cancer could be dysregulated by multiple alterations through genome, epigenome, transcriptome, and proteome dimensions. **Methods:** Here we have proposed a new integrative framework designed to perform these three functions simultaneously: (1) predicting censored survival data; (2) integrating meta-dimensional omics data; (3) identifying interactions within/between meta-dimensional genomic features associated with survival. In order to predict censored survival data, martingale residuals were calculated as a new outcome and a new fitness function based on mean absolute difference of martingale residuals was implemented for the grammatical evolution neural network (GENN). **Results:** To test the utility of the proposed framework, a simulation study was conducted, followed by an analysis of meta-dimensional omics data including copy number, methylation, gene expression, and protein expression data in breast cancer retrieved from The Cancer Genome Atlas. According to the results from the simulation data, martingale residuals performed properly as a new continuous outcome in terms of finding true survival genes using GENN. On the basis of the results from breast cancer dataset, we were able to identify interactions not only within a single dimension of genomic data but also between meta-dimensional omics data that are associated with survival. Notably, the predictive power of our best meta-dimensional model was 73% which outperformed all of the other models conducted based on a single dimension of genomic data. **Conclusions:** Breast cancer is an extremely heterogeneous disease and the high degree of diversity within/between breast tumors could affect the risk of therapeutic responses and disease progression. Thus, identifying interactions within/between meta-dimensional omics data associated with survival in breast cancer is expected to provide guidance for improved meta-dimensional prognostic biomarkers and therapeutic targets.

Session B3 16:20-18:00 October 25 (Saturday)

16:20-16:40 *Crosstalk between pathways enhances the controllability of signaling networks*

Dingjie Wang, **Suoqin Jin** and Xiufen Zou
School of Mathematics and Statistics, Wuhan University
Paper IDs: 23

Abstract: The control problem of complex networks is one of the most challenging problems in biology and engineering fields. In this study, we explore the controllability and control energy of several signaling networks which consist of many interconnected pathways, especially networks with bow-tie architecture. Based on the theory of structure controllability and quantified bound of the required control energy that we derive, we show that the biological mechanisms, such as cross-pathway inhibition, compartmentalization, et al., make the networks easier to be fully controlled and the control energy of feed-forward networks is mainly determined by the decay rates of proteins. These results indicate the biological networks are optimally designed to achieve their normal functions from the viewpoint of control theory. Our work enables a comprehensive understanding of the impact of network structures and properties on controllability.

16:40-17:00 *A Tensor-Based Markov Chain Method for Module Identification from Multiple Networks*

Chenyang Shen, Shuqin Zhang and Michael Kwok-Po Ng
Department of Mathematics, Hong Kong Baptist University
Paper IDs: 30

Abstract: The interactions among different genes, proteins and other small molecules are becoming more and more significant and have been studied intensively nowadays. One general way that helps people understand these interactions is to analyze networks constructed from genes/proteins. In particular, module structure as a common property of most biological networks has drawn much attention of researchers from different fields. In most cases, biological networks can be corrupted by noise in the data and the corruption may cause mis-identification of module structure. Besides, some structure may be destroyed when improper experimental settings are built up. Thus module structure may be unstable when one single network is employed. In this paper, we consider employing multiple networks for consistent module detection in order to reduce the effect of noise and experimental setting. Instead of considering different networks separately, our idea is to combine multiple networks together by building them into tensor structure data. Then give any node as prior label information, tensor-based Markov chains are constructed iteratively for identification of the modules shared by the multiple networks. In addition, the proposed tensor-based Markov chain algorithm is capable of simultaneously evaluating the contribution from each network. It would be useful to measure the consistency of modules in the multiple networks. In the experiments, we test our method on two groups of gene co-expression networks from human beings. We also validate the modules identified by the proposed method.

17:00-17:20 *Testing Multiple Hypotheses through IMP Weighted FDR Based on a Genetic Functional Network with Application to a New Zebrafish Transcriptome Study*

Jiang Gui, Casey Greene, Con Sullivan, Walter Taylor, **Jason Moore** and Carol Kim
Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH, US
Paper IDs: 33

Abstract: In genome-wide studies, thousands of hypothesis tests are carried out at the same time. Bonferroni correction and False Discovery Rate (FDR) can effectively control type I error but often yield a high false negative rate. We aim to develop a more powerful method to detect differential expressed genes. We present an Weighted False Discovery Rate (WFDR) method that incorporate biological knowledge from genetic networks. We first identify weights using Integrative Multi-species Prediction (IMP) and then apply the weights in WFDR to identify differentially expressed genes through a IMP-WFDR algorithm. We conducted a simulation study to characterize the performance of this method. We performed genomic characterization to identify potential synergistic and antagonist interactions between the highly-conserved zebrafish cftr gene and the environmental toxicant arsenic, particularly in the context of a systemic infection with Pseudomonas aeruginosa. Zebrafish were exposed to arsenic at 10 parts per billion and/or infected with P. aeruginosa. Appropriate controls were included. We then applied IMP-WFDR during the analysis of differentially expressed genes. We compared the mRNA expression for each group and found over 200 differentially expressed genes and several enriched pathways including defense response pathways, arsenic response pathways, and the Notch signaling pathway.

17:20-17:40 *Parallelization of Enumerating Tree-like Chemical Compounds by Breadth-first Search Order*

Morihiro Hayashida, Jira Jindalertudomdee, Yang Zhao and Tatsuya Akutsu
Bioinformatics Center, Institute for Chemical Research, Kyoto University
Paper ID: 25

Abstract: Enumeration of chemical compounds greatly assists designing and finding new drugs, and determining chemical structures from mass spectrometry. In our previous study, we developed efficient algorithms, BfsSimEnum and BfsMulEnum for enumerating tree-like chemical compounds without and with multiple bonds, respectively. For many instances, our proposed algorithms were able to enumerate chemical structures faster than other existing methods. Latest processors consist of multiple processing cores, and are able to execute many tasks at the same time. In this paper, we develop three parallelized algorithms BfsEnumP1-3 by modifying BfsSimEnum in simple manners to further reduce execution time. BfsSimEnum constructs a family tree in which each vertex denotes a molecular tree. BfsEnumP1-3 divide vertices with some given depth of the family tree, and the subtree rooted at each vertex with the depth to be constructed is assigned to a processor. For evaluation, we perform experiments for several instances with varying the division depth and the number of processors, and show that BfsEnumP1-3 are useful to reduce the execution time for enumeration of tree-like chemical compounds. In addition, we show that BfsEnumP3 achieve more than 80% parallelization efficiency using up to 11 processors, and reduce the execution time using 12 processors to about 10% of that by BfsSimEnum.

17:40-18:00 *Extracting discriminatively interpretable features of gene network by combining gene expression, variance and covariance*

Xiangtian Yu, Tao Zeng, Guojun Li and Luonan Chen
School of Mathematics, Shandong University, Jinan 250100, China
Paper IDs: 41

Abstract: Feature extraction and selection is an important step for gene expression analysis, especially for differential gene expression analysis. In conventional analysis, it is assumed the control and case samples (e.g. normal and diseased) have great purity. However, recent studies reveal that the compositions of disease samples are more complicated than expected (e.g. heterogeneity of diseases). And even more, many disease genes are not always consistently up-/down-regulated, leading to be under-estimated. Although the (differential) expression variance and (differential) expression covariance can address such a problem in a network manner, such analyses always require multiple samples rather than one sample. To extract discriminatively interpretable features from gene expression data in one sample, a differential score (DEVC) based approach is proposed to combine the statistic measurements of gene expression, expression variance and expression covariance together. The

DEVC-based differential expression network (DEVC-net) has a bi-coloured topological structure, and is further implemented to enhance the power of differential analysis. To validate our theoretical method, a toy model is first shown. The DEVC-net approach then carried on two gene expression datasets about prostate cancer and diabetes, which consistently demonstrated: (1) differential expression variance is indeed a new informative source compared to differential average expression; (2) differential expression covariance can actually provide more discriminative gene-pairs rather than individual genes; (3) DEVC is effective to measure the expression state of genes and their network or modules in one sample. All of these results strongly support the universal advantage of DEVC, as effectively extracting discriminatively interpretable features of gene network of one sample when disease samples are heterogeneous.

Session B4 10:50-12:30 October 26 (Sunday)

10:50-11:10 *An independent filter for gene set testing based on spectral enrichment*

H. Robert Frost, Zhigang Li, Folkert Asselbergs and Jason Moore
Institute for Quantitative Biomedical Sciences,Geisel School of Medicine, Lebanon,
Paper ID: 43

Abstract: Gene set testing has become an indispensable tool for the analysis of high-dimensional genomic data. An important motivation for testing gene sets, rather than individual genomic variables, is to improve statistical power by reducing the number of tested hypotheses. Given the dramatic growth in common gene set collections such as the Gene Ontology (GO), however, gene set testing is often performed with nearly as many gene sets as underlying genomic variables. Not only is power not improved in such cases, but large gene set collections typically contain many highly interdependent gene sets most of whose members are generated algorithmically without human review or experimental validation. To address the challenge posed by large, interdependent and low quality gene set collections, we have developed spectral gene set filtering (SGSF), a novel technique for independent filtering of gene set collections prior to standard gene set testing. The SGSF method uses as a filter statistic the p-value measuring the statistical significance of the association between each gene set and the principal components (PCs) of an empirical data set, taking into account the significance of the eigenvalue associated with each PC. The SGSF method is effective in any experimental context, e.g., analysis of cancer gene expression data, where the variance structure of genomic variables is associated with the experimental outcome of interest under HA. Because this filter statistic is independent of standard gene set enrichment test statistics under H0 but dependent under HA, the proportion of significantly enriched gene sets is increased without impacting the type I error rate. As shown using simulated gene sets with simulated data and MSigDB collections with microarray gene expression data, the SGSF algorithm accurately filters gene sets unrelated to the experimental outcome resulting in significantly increased gene set enrichment power.

11:10-11:30 *Predicting Golgi-resident proteins in plants by incorporating N-terminal transmembrane domain information in the general form of Chou's pseudo-amino acid compositions*

Yasen Jiao, Xiaoquan Su and Pufeng Du
School of Computer Science and Technology, Tianjin University, Tianjin, China
Paper ID: 50

Abstract: **Purpose:** Knowing the subcellular location of a protein is an important step in understanding its biological functions. Although a lot of works about this field have been done, it is still a challenging problem. So in this paper, we developed a new method to identify whether a protein is a Golgi-resident protein or not in plant cells. **Methods:** We proposed to incorporate transmembrane domain information and six different kinds of physicochemical properties of amino acids in the general form of Chou's pseudo-amino acid compositions including the amino acid compositions of the protein, which represents the occurrence frequencies of 20 different types of amino acids in the sequence, the di-peptide compositions of the protein, which represents the occurrence frequencies of 400 di-peptides and four parameters of the transmembrane domains: the average length of all transmembrane domains, the length of the first transmembrane domain, the number of transmembrane domains in the first 70 amino acids and the probability that the N-terminal of the protein is on the cytoplasmic side of the membrane. As we focused on predicting Golgi proteins in plants, we chose Arabidopsis thaliana as the model organism in this study and all the data was selected by several screening procedures strictly. **Results:** By using SVM based classifiers, our method achieved over 90% prediction accuracy in a 5-fold cross validation, which is much better than the other state-of-the-art methods.

11:30-11:50 *Sparse Electrocardiogram Signals Recovery Based on Solving a Row Echelon-Like Form of System*

Pingmei Cai, Guinan Wang, Hongjuan Zhang, Shuxue Ding and Zikai Wu
Department of Mathematics, Shanghai University ong, China

Paper ID: 53

Abstract: Sparse signal recovery has become a important part in the field of signal recovery. This paper proposed a two stage recovery algorithm for sparse signals based on the time domain. First, the dictionary(i.e. the mixing matrix) is estimated after the concentration subspaces are found. Next, in the signal recovery stage, we divide the time points into different layers according to the number of active sources at each time point. Then, by constructing some transformation matrices, these time points form a row echelon-like system. After these, the sources at each layer can be solved out explicitly by corresponding matrix operations. All these operations are conducted under a weak sparse condition that the number of active sources is less than the number of observations. Experiment results show that the method proposed has good performance for sparse signal recovery.

11:50-12:10 *Topological Characterization of Housekeeping Genes in Human Protein-Protein Interaction Network*

Pei Wang, Yuhuan Zhang, Jinhu Lu and Xinghuo Yu
Henan University & RMIT University, China

Paper ID: 2

Abstract: Human housekeeping genes (HKGs) are widely expressed in various tissues, which involve in cell maintenance or sustaining cell function, and are often taken as experimental control and normalization references in gene expression experiments. Based on literature curation and up-to-date databases, we construct a large-scale human protein-protein interaction network (HPIN) and a HKGs subnetwork. Through the topological features of HKGs in the HPIN, we characterize the topological features of human HKGs. Our results indicate HKGs are with very large average degree, k-shell, betweeness, semilocal and eigenvector centralities, clustering coefficient, closeness, PageRank and motif centrality, which are all higher than that of the HPIN. Among the nine indexes, HKGs are with the average betweeness about 7 times larger than that for the HPIN, but they are also with the largest coefficient of variant (CV). The closeness of HKGs is with the smallest CV and very large median. Based on ROC analysis, we find most of the indexes and their compositions can be used to predict HKGs, with prediction accuracy around 80%. Especially, the prediction accuracy of the closeness can achieve as high as 82.36%. The investigations shed some lights on the characterization and identification of human functional genes, which have potential implications in systems biology and networked medicine.

12:10-12:30 *Combined analysis of gene regulatory network and SNP information enhances identification of potential gene markers in mouse knockout studies with small number of samples*

Benjamin Hur, Heejoon Chae and Sun Kim
Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea

Paper ID: 61

Abstract: Even though RNA-sequencing data are widely used in order to find differential expressed genes among control and case studies, finding gene markers that represent phenotypic differences in small number of samples remains a challenging task. Finding gene markers using standard differential expressed gene methods produce too many candidate genes and the number of candidates varies at different threshold values. In addition, in small number of samples, it is known that the statistic power is too low to discriminate whether gene expressions were altered by genetic differences or not, which is a problem that RNA-sequencing data faces frequently. Hence, discovering gene markers remains a difficult task. In this study, we purpose a four-step filtering method that predicts gene markers from RNA-sequence data of mouse knockout studies by utilizing gene regulatory network constructed from omics data in the public domain, biological knowledge from curated pathways, and single-nucleotide polymorphism information. Our prediction method was not only able to reduce the number of candidate genes than the differential expressed gene-only filtered method, but also successfully predicted significant genes that are reported in research findings of the data contributors.

Session B5 14:00-15:40 October 26 (Sunday)

14:00-14:20 *Identifying Prognostic Features by Bottom-up Approach and Correlating to Drug Repositioning*

Wei Li, Jian Yu, Baofeng Lian, Han Sun, Jing Li, Menghuan Zhang, Qian Liu, Yixue Li and **Lu Xie**
Shanghai Center for Bioinformation Technology, Shanghai Institutes of Biomedicine, Shanghai Academy of Science and Technology, Shanghai 201203, P. R. China

Paper ID: 65

Abstract: **Background:** Traditionally top-down method was used to identify prognostic features in cancer research. That is to say, differentially expressed genes in cancer versus normal were identified to see if they possess survival prediction power. The problem is that prognostic features identified from one set of patient samples can rarely be transferred to other datasets. We apply bottom-up approach in this study: survival correlated or clinical stage correlated genes were selected first and prioritized by their network topology additionally, then a small set of features can be used as a prognostic signature. **Methods:** Gene expression profiles of a cohort of 221 hepatocellular carcinoma (HCC) patients were used as a training set, ‘bottom-up’ approach was applied to discover gene-expression signatures associated with survival in both tumor and adjacent non-tumor tissues, and compared with ‘top-down’ approach. The results were validated in a second cohort of 82 patients which was used as a testing set. **Results:** Two sets of gene signatures separately identified in tumor and adjacent non-tumor tissues by bottom-up approach were developed in the training cohort. These two signatures were associated with overall survival times of HCC patients and the robustness of each was validated in the testing set, and each predictive performance was better than gene expression signatures reported previously. Moreover, genes in these two prognosis signature gave some indications for drug-repositioning on HCC. Some approved drugs targeting these markers have the alternative indications on hepatocellular carcinoma. **Conclusion:** Using the bottom-up approach, we have developed two prognostic gene signatures with a limited number of genes that associated with overall survival times of patients with HCC. Furthermore, prognostic markers in these two signatures have the potential to be therapeutic targets.

14:20-14:40 *DMET-Miner: Efficient Learning of Association Rules from Genotyping Data for Personalized Medicine*

Pietro Hiram Guzzi, Mario Cannataro and **Giuseppe Agapito**
Informatics and Biomedical Engineering, University "Magna Græcia" of Catanzaro, Viale Europa (Località Germaneto), 88100 CATANZARO, ITALY

Paper ID: 99

Abstract: Recent developments of microarray technology enable the investigation of allelic variants that may be correlated to phenotypes. In particular the Affymetrix DMET (Drug Metabolism Enzymes and Transporters) platform enables the simultaneous investigation of all the genes that are related to drug absorption, distribution, metabolism and excretion (ADME). Recent studies demonstrated the effectiveness of the use of DMET data for studying drug response or toxicity in clinical studies. In a previous work we developed DMET-Analyzer, a platform able to automatize the statistical study of allelic variants, that has been validated in clinical studies. Nevertheless, main limitation of the underlying statistic analysis strategy of DMET-Analyzer is the focus on a single variant for each time, i.e. although it is able to correlate a single variant for each probe (related to a portion of a gene) through the use of the Fisher test, on the other hand it is unable to discover multiple associations among allelic variants. To overcome those limitations, here we propose an extension of DMET-Analyzer, named DMET-Miner, that is able to correlate the presence of a set of allelic variants by exploiting association rules mining and employing an Apriori-like discovery strategy. First, we formulate the problem of finding a set of candidate allelic variants correlated to the patients as the finding of Frequent Sets of allelic variants. Then, we developed DMET-Miner, a novel tool that is able to extensively mine DMET data and to discover multiple associations between allelic variants and patient’s conditions, using item sets and associations rules. Because of the huge DMET datasets produced in clinical studies, an issue in applying current association rules mining algorithms, such as Apriori, is the explosion of the search space and the production of an high number of frequent item sets. Thus, we applied a new efficient data structure (FP-tree) to implement a new version of Apriori based on FP-tree, that reduce the search space and the execution time. Preliminary experiments on a synthetic DMET dataset and on a real DMET dataset, show how DMET-Miner outperforms off-the-shelf data mining suites such as Weka in the mining of such datasets. To demonstrate the biological relevance of the extracted association rules and the effectiveness of the proposed approach from a medical point of view, some preliminary studies on a real clinical dataset are currently under medical investigation.

14:40-15:00 *Drug Name Recognition Using Conditional Random Fields with Word Embeddings*

Shengyu Liu, Buzhou Tang, Qingcai Chen, Xiaolong Wang and Bin Tang
Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School

Paper ID: 104

Abstract: Word embeddings that contain rich latent semantic information of words have been widely used in various natural language processing tasks recently and shown stable improvements. However, it has not been used in Drug name recognition (DNR). We propose a conditional random fileds (CRF)-based system using word embeddings. To investigate the effect of word embeddings on DNR, we compare it with the semantic features based on three public knowledge bases. The skip-gram model, a recently proposed algorithm, is used to induce word embeddings on about 17.3G unlabeled biomedical texts collected from MEDLINE. The CRF-based system is evaluated on the DDIExtraction 2013 corpus. Experimental results show that word embeddings significantly improve the performance of the CRF-based DNR system, and they are competitive with the semantic features based on the knowledge bases. Furthermore, the word embeddings are complementary to the semantic features based on the knowledge bases. When both the word embeddings and the semantic features based on the knowledge bases are added, our system achieves the best performance in F-score of 78.37%, which outperforms the best system of the DDIExtraction 2013 challenge by 6.87%.

15:00-15:20 *Improving common lines detection in protein 3D structure reconstruction from the cryo-EM by a new optimized denosing filter*

Biao Zhang, Qiyu Jin, Jinhua Yang and Hong-Bin Shen
Shanghai Jiaotong University, China
Paper ID: 19

Abstract: Reconstructing the three-dimensional structure of a macromolecule given its two-dimensional noisy projection images is the main task of the single particle reconstruction (SPR) of cryo-electron microscopy (cryo-EM). One of the biggest problems in the SPR process is that the direction of every projection image is unknown and no prior knowledge can be available. It becomes more difficult when the images are of low signal-to-noise ratio (SNR) and the image's common lines have to be estimated. In this paper, we propose a new denoising algorithm called NFOWEM based on the weighted average of the neighborhood observations. Its weights are optimized by minimizing a tight upper bound of mean square error. NFOWEM is more robust to high levels noise compared to other denoising algorithms. Experimental results on the 50S ribosomal subunit demonstrate that higher detection rates of common lines can be achieved on the noise filtered images by NFOWEM.

Session B6 16:20-18:00 October 26 (Sunday)

16:20-16:40 *Joint identification of differentially expressed genes and phenotype-associated genes*

Samuel Sunghwan Cho, Minseok Seo, Su-Kyung Shin, Eun-Young Kwon, **Yun-Jung Bae**, Mi-Kyung Sung, Myung-Sook Choi and Taesung Park
Seoul National University
Paper ID: 75

Abstract: Over the last decade, many analytical methods and tools have been developed for microarray data. The detection of differentially expressed genes (DEGs) among different treatment groups is often a primary purpose of microarray data analysis. In addition, association studies investigating the relationship between genes and a phenotype of interest such as survival time are also popular in microarray data analysis. Phenotype association analysis provides a list of phenotype-associated genes (PAGs). However, it is sometimes necessary to identify genes that are both DEGs and PAGs. We consider the joint identification of DEGs and PAGs in microarray data analyses. The first approach we used was a naïve approach that detects DEGs and PAGs separately and then identifies the genes in an intersection of the list of PAGs and DEGs. The second approach we used was a hierarchical approach that detects DEGs first and then chooses PAGs from among the DEGs or vice versa. In this study, we propose a new model-based approach for the joint identification of DEGs and PAGs. The proposed model-based methods were evaluated using experimental data and a few simulation studies. The proposed methods were used to analyze a microarray experiment in which the main interest lies in detecting genes that are both DEGs and PAGs, where DEGs are identified between two diet groups and PAGs are associated with four phenotypes reflecting the expression of leptin, adiponectin, insulin-like growth factor 1, and insulin. Model-based approaches provided a larger number of both DEGs and PAGs than do other methods. Simulation studies showed that they have more power than other methods. Through analysis of data from experimental microarrays and simulation studies, the proposed model-based approach was shown to provide a more powerful result than the naïve approach and the hierarchical approach. Since our approach is model-based, it is very flexible and can easily handle different types of covariates.

16:40-17:00 *An Entropy-based Statistical Workflow Provides Noise-Minimizing Biological Annotation for Muscular Aging*

Theodoros Koutsandreas, Ioannis Valavanis, Eleftherios Pilalis and **Aristotelis Chatziioannou**
Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation (NHRF), Athens, Greece
Paper ID: 97

Abstract: This study aims to expand the efficiency of the interpretation concerning the aging process, by exploring a broad gene set, derived from the analysis of an integrative transcriptomic microarray dataset. The dataset comprises human skeletal muscle samples, obtained from healthy males and females, that were used to derive a gene signature of a high informative content, with respect to its functional association with the aging phenotype. Towards this end, a multilayered computational workflow integrating advanced statistical methodologies for the derivation of reliable confidence measures, distribution-based entropy calculations to examine the informational content of the dataset, enrichment analysis, graph-theoretic methods and intuitive visualization was applied. Specifically, statistical testing revealed differentially expressed genes, while an uncertainty calculation algorithm,

exploiting Gene Ontology (GO) terms annotations, extended the list of significant genes from 254 to 2791, namely p-value threshold was increased from 0.0005 to 0.103, while keeping simultaneously noise measurements legitimately low. This rich gene set associated functionally the macroscopic phenotype of muscular aging with highly informative, stably correlated with each other, molecular annotations in the GO database. Finally, a set of 57 reliable genes was identified that comprise a gender-independent aging signature, after incorporating crucial information about genes pivotal regulatory role as inferred by the GO tree. The biological interpretation was highly assisted by the illustration of the functional mappings between genes, cellular location and biological processes through circle packing graphs

17:00-17:20 *Systematic Analysis of Specific Biomarkers and Pathological Mechanism for Non-alcoholic steatohepatitis Based on Gene Expression Profiling of Multiple Progressive Stages*

Weili Lin, Qingchen Zhang, Dingfeng Wu, Shaohua Yuan, Lixin Zhu and Ruixin Zhu
Department of Bioinformatics, Tongji University, Shanghai, P.R. China
Paper ID: 102

Abstract: Non-alcoholic fatty liver disease (NAFLD) covers both simple steatosis as a relatively benign non-progressive clinical course, and non-alcoholic steatohepatitis (NASH) as a severe pathological stage. Non-alcoholic steatohepatitis (NASH) is a progressive liver disease of worldwide significance with the dramatic rise of population of obesity and diabetes. However, the underlying pathogenesis of NASH is still ambiguous and yet to be fully elucidated. Simultaneously, specific candidate biomarkers for NASH are badly in need to facilitate the establishment of a non-invasive diagnosis for NASH. Taking advantage of well characterized high throughput data like microarray datasets of different clinically defined pathological groups including normal, obesity, steatosis and NASH, we performed a microarray-based gene expression profiling systematic analysis of three pathological stages in progression of NASH—healthy obesity, steatosis and NASH. For biomarker prediction, we defined a new method, "NASH—Counter discriminative pattern", to detect specific biomarkers for NASH. A NASH—Specific biomarker combination was identified. And differentially expressed genes (DEGs) in the combinations including AKR1B10, FABP4, SPP1, THY1, FABP5, FAT1 and MCM2 were demonstrated to be highly correlated with molecular pathological mechanisms, which underlie the development of NASH. The NASH—Specific biomarker combination we identified was validated in another independent dataset. The predictive accuracy for validation dataset reaches 91.1%. For studying the molecular pathological mechanism of development of NASH in different progressive stages, we performed functional enrichment analysis on three groups of DEGs responsible for three disease progressions including normal control to obesity, obesity to steatosis and steatosis to NASH. Through combining analysis of enriched pathways and biological processes, a hypothetical pathogenetic network of NASH was created and is consistent with multiple parallel hits hypothesis. Thus, our study provides a new method to investigate progressive disease with multiple pathological stages. In addition, a comprehensive and systematic understanding of the pathogenesis of NASH was revealed and may facilitate the diagnosis, prevention and treatment of NASH.

17:20-17:40 *Goes Open——Publish with Genomics, Proteomics & Bioinformatics*

Yuxia Jiao
GPB Editorial Offics, Beijing Institute of Genomics, Chinese Academy of Sciences

Session C1 14:00-15:40 October 26 (Sunday)

14:00-14:20 *Centrality of complex disease genes unveiled by eQTL associations*

Haiquan Li, Nima Pouladi, Vincent Gardeux, Qi Luo, Qike Li, Jianrong Li, Fernando Martinez, Joe Garcia and Yves Lussier
University of Arizona, USA
Paper ID: 81

Abstract: The network property of disease genes, particularly the topological centrality, has been under investigation and long-term debates for many years. Whether complex disease genes are central or peripheral in the genetic networks is highly controversial. Furthermore, existing centrality studies are focusing on protein interaction networks, but rarely on genetic regulation and association networks, not to mention the cross-scale networks, such as eQTL (Expression Quantitative Trait Loci) network. In this paper, we surveyed the centrality of complex disease-related mRNAs and observed robust enrich centrality of these mNRAs under various eQTL strengths, across cell lines/tissues, and with stringent linkage disequilibrium control to remove confounders. The observations indicate that disease related genes are likely to be regulated by a larger number of SNPs than expected by chance, each of which may be of small and local effect. Our findings provide novel insights into the nature of complex disease.

14:20-14:40 *A novel knowledge-based three-body potential for transcription factor binding site prediction*

Wenyi Qin, Guijun Zhao, Caiyan Jia and Hui Lu

University of Illinois at Chicago, USA

Paper ID: 114

Abstract: A structure based statistical potential is developed for transcription factor binding site (TFBS) prediction. Besides the direct contact between amino acids from transcription factors and DNA bases, we also considered the influence of the neighboring base. This three body potential showed better discriminate powers than the two body potential. We validate the performance of the potential in TFBS identification, binding energy prediction, and binding mutation prediction. The results indicate that our current potential is among the best in existing energy functions for protein-DNA interaction predictions.

14:40-15:00 *Network-based Prediction and Knowledge Mining Of Disease Genes*

Matthew Carson and Hui Lu

University of Illinois at Chicago, USA

Paper ID: 95

Abstract: During the past few decades, the amount of biological data available for analysis has grown exponentially. Along with this vast amount of information comes the challenge to make sense of it all. The growing popularity of network analysis and data mining has allowed researchers to make connections between these data that was not possible in the past. It is now feasible to find not only direct interactions between biological elements such as proteins, DNA, and RNA, but distant relationships as well. Through the application of these relatively new tools, many are searching for the genetic causes of disease. We examined the human protein-protein interaction (PPI) network as it relates to human illness using the Disease Ontology and an alternating decision tree (ADTree) classifier in conjunction with 10 topological features. In addition, we created a bootstrapped tree to highlight conserved rules among multiple tree-building iterations. Subsequently, we analyzed false positive results for evidence of disease association not accounted for in the Disease Ontology-focused annotation of the human genome. Our classifier was able to predict dis-ease-related genes with 79% area under the ROC curve (AUC), which indicates the trade off between sensitivity and specificity and is a good predictor of how a classifier will perform on future data sets. We found that a combination of several net-work characteristics including degree centrality, disease neighbor ratio, eccentricity, and neighborhood connectivity help to distinguish between these two classes. Furthermore, the ADTree allowed us to see which combinations of strongly predictive attributes contributed most to protein-disease classification. In our post-processing evaluation, we found several examples of potential novel disease-related proteins and literature evidence for these. We use two proteins as examples to show that first and second neighbors in the PPI network can be used to identify likely disease associations.

15:00-15:20 *Identification association of drug-disease by using functional Gene module for breast cancer*

Lida Zhu and Fuxi Zhu

Computer School of Wuhan University

Paper ID: 109

Abstract: In oncology drug development, it is importance to develop low risk drugs efficiently, and computational methods have been paid more and more attention in drug discovery. However, the patient population heterogeneity complicates the prediction of the therapeutic efficiency. Here we introduce a novel method to identify repositioned drug against breast cancer by integrated the breast cancer survival data with the drug sensitivity information. Among the 140 drug candidates, we are able to filter 4 FDA approved drugs and could identify 2 breast cancer drugs among 4 known breast cancer therapeutic drug in total.

15:20-15:40 *Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks*

Lin Wang, Yong Wang, Qinghua Hu and Shao Li

Tianjin University of Science and Technology

Abstract: Drug targets and disease genes may work as driver factors in transcriptional level, which propagate signals through gene regulatory network and cause the downstream genes' differential expression. How to analyze transcriptional response data to identify meaningful gene modules shared by both drugs and diseases is still a critical issue for drug-disease associations and its molecular mechanism. In this article, we propose the drug-gene-disease coherent subnetwork concept to group the biological function related drugs, diseases, and genes. It was defined as the subnetwork with drug, gene, and disease as nodes and their interactions coherently crossing three data layers as edges. Integrating differential expression profiles of 418 drugs and 84 diseases, we develop a computational framework and

identify 13 coherent subnetworks such as inflammatory bowel disease and melanoma relevant subnetwork. The results demonstrate that our coherent subnetwork approach is able to identify novel drug indications and highlight their molecular basis.

Highlight track abstracts

Session H1：Highlight Track 10:30-12:30 October 25 (Saturday)

10:50-11:10 *Analysis of Stochastic cell dynamics from single cell data by Flow Cytometry (FCM)*

Minping Qian and Guanglu Gong
Peking University

Abstract: Single-cell and single-molecule experiments have shown that many important cellular processes, e.g. transcription, translation, replication, and gene regulation, are inherently stochastic. A new generation of Flow Cytometry (FCM) provides a way to see single cell behavior on up to 10 expression levels of genes, proteins and other items. This measurement provides not only the mean expression levels of genes for a population of cells, but also shows the stochastc behaviors of molecules in individual cell, such as phenotypes of cells and random switch between phenotypes. We model such stochastic behaviors by stochastic differential equations, which matches the regulatory law accepted by biologists and biophysists. Then we show the data from FCM, can be processed to see the the phenotypes and switch between them. The large deviation theory by Freidlin and Ventzell ([1]) is applied to explain our model on molecular level for cells reduced to Markov chain models for phenotype changing directly introduced on cell level as E. Lander's group did ([2]), when the scaling limit is taken [1]. It is indicated that the relationship between genes and proteins can be very different, i.e. genes of cells in one phenotype are positive correlated, while in another phenotypes they could be negatively correlated. Therefore, to mix data in different phenotypes in correlation analysis of gene expression may lead to wrong conclussions. We will also discuss the calculating methods for identifying phenotypes and switching between them.

[1] X Sunney Xie, Single-molecule approach to molecular biology in living bacterial cells. Annu. Rev. Biophys., 37:417–444, 2008.
[2] Freidlin MI, Szucs J, Wentzell AD (2012) Random perturbations of dynamical systems, vol.260 (Springer).
[3] Gupta PB, et al. (2011) Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell 146:633–644.

10:10-11:30 *Two novel formulations for biochemical reaction networks*

Tianshou Zhou
Sun Yat-Sen University

Abstract: Intracellular processes are noisy: in each cell, concentrations of molecules are subject to stochastic fluctuations due to the small numbers of these molecules. The chemical master equation (CME) in principle provides the most complete model for probabilistic behavior of a biochemical reaction network but the practical challenge is that its simulation is very time-consuming and even becomes infeasible in the case of multiple reactive species. This computational issue has remained unsolved over 70 years although many methods have been developed. Here we present two novel formulations for the CME, which can well solve this issue. These two schemes are established based on the CME, both identifying the structure of the reaction network. In contrast to the existing methods, our methods have more advantages, which will be specified in my talk. The overall formulations, theories and algorithms provides a paradigm for modeling, analysis and computation of biochemical networks with any finite reactions and reactive species.

11:30-11:50 *Raison d'être of insulin resistance: the adjustable threshold hypothesis*

Guanyu Wang
Department of Biology, South University of Science and Technology of China

Abstract: The epidemics of obesity and diabetes demand a deeper understanding of insulin resistance, for which the adjustable threshold hypothesis is formed in this paper. To test the hypothesis, mathematical modeling was used to analyze clinical data and to simulate biological processes at both molecular and organismal levels. I found that insulin resistance roots in the thresholds of the cell's bistable response. By assuming heterogeneity of the thresholds, single cells' all-or-none response can collectively produce a graded response at the whole-body level — conforming to existing data. The thresholds have to be adjustable to adapt to extreme conditions. During pregnancy, for example, the thresholds increase consistently to strengthen the mother's insulin resistance to meet the increasing glucose demand of the expanding fetal brain. I also found that hysteresis, a key element of the adjustable threshold hypothesis, can explain reactive hypoglycemia, which is characteristic of diabetes complications but remains poorly understood. Contrary to the common belief that insulin promotes glucose disposal, the results imply that insulin is the body's "ration stamp" to restricting glucose utilization by peripheral tissues and that insulin resistance is primarily a well-evolved mechanism. The hypothesis provides an intuitive and dynamical description of the previously formless insulin resistance, which may make the detection of pre-diabetes possible and may shed light on the optimal timing of therapeutic intervention. It also provides valuable clues to defining subtypes of type 2 diabetes that might respond differently to specific prevention and intervention strategies.

11:50-12:10 *Dynamical Analysis on a 2-D Disease Model with Convex Incidence Rate*

Pei Yu

Department of Applied Mathematics, Western University, London, Ontario, Canada

Abstract: Mathematical models in epidemiology and in-host disease share common features, dividing a population of individuals (epidemiology) or cells (in-host) into discrete classes relevant to the disease dynamics, and typically describing their dynamics with a system of ordinary differential equations (ODEs). A key feature of this system is the incidence function, which defines the spread of the infection to susceptibles.

In classical epidemiological models, the incidence rate is often assumed to take the form $\beta SI/N$, where $S(t)$ is the number of susceptible individuals, $I(t)$ is the number of infectives and β is a constant, the transmission rate [1]. When N, the population size, is constant, this incidence function is also simply written as βSI. Similarly, for in-host models, the rate at which uninfected cells become infected is often described as βxy, where $x(t)$ reflects the uninfected cell density and $y(t)$ denotes the density of infected cells [2]. Taking advantage of assuming that the incidence rate function $f(S,I,N)$ satisfies the concave condition $\partial 2f(S,I,N)/\partial I2 \leqslant 0$, Korobeinikov and Maini [3] derived elegant results for all concave incidence functions, showing that for the standard SIRS model [1] with a constant population size, the global asymptotic stability of the disease-free equilibrium when the basic reproduction number, $R0 \leqslant 1$, and global asymptotic stability of the endemic equilibrium when $R0 > 1$.

In contrast, we have recently analyzed a number of ODE models with convex incidence functions. If incidence is convex, or “synergistic”, the rate at which new infections occur can increase supralinearly with disease prevalence. This situation can arise in a number of realistic scenarios. For example, in in-host models of HIV, increasing the extent of the infection involves greater damage to the immune system, and can thus increase the incidence rate [4]. Similarly, in autoimmune disease, increases to the autoimmune response against self tissue can cause a positive feedback loop which will further increase the incidence rate [5]. While these two examples both arise in in-host disease modelling, catastrophic outbreak or pandemic conditions could also result in convex epidemiological incidence.

In this contribution, we analyze in detail the possible dynamical behaviors of a simple 2-dimensional disease model with a convex, or synergistic, incidence function. The system we analyze is a standard non-dimensionalized SI model which arises in both epidemiology and in-host modelling: it assumes a birth rate into the susceptible population, death rates for both populations, and an incidence rate between the two. The incidence function we study has an analytical form which has arisen in a number of models previously analyzed. In marked contrast to the powerful general conclusions obtained for concave incidence functions [3], we find that a wide range of dynamical behaviors are possible when incidence is synergistic. In particular, as previously analyzed in related higher-dimensional models [6, 7, 8], we note the appearance of recurrent infection, that is, cycles consisting of long periods close to the disease free equilibrium, punctuated by brief bursts of disease. This pattern of recurrence occurs in many diseases, including the intriguing pattern of “viral blips” in HIV, as well as the recurrent episodes characteristic of autoimmune diseases.

For the simple 2-dimensional model, we explore several mechanisms which can underly these physiologically relevant patterns of infection, finding that when the incidence function is convex, bistable equilibrium solutions, Hopf and generalized Hopf bifurcations and, in particular, homoclinic bifurcations may all contribute to disease recurrence.

12:10-12:30 *Robust Period of Mammalian Circadian Oscillator from Amplitude Balance between Feedback Loops*

Jie Yan, Guangsen Shi, Zhihui Zhang, Xi Wu, Zhiwei Liu, Lijuan Xing, Zhipeng Qu, Zhen Dong, **Ling Yang** and Ying Xu

Center for Systems Biology, Soochow University, China

Abstract: The mammalian circadian oscillation is driven by a delayed negative feedback loop (the primary loop). This system also includes a positive auxiliary loop: The nuclear receptor Rev-erba can promote itself by repressing the transcription of its inhibitor Cry1. Mathematical models reveal a key modulation mechanism rule: it’s the strength ratio between of negative feedback and the positive feedback, rather than the sole strength, determines the period. Additionally, a transmitting effect between the numerator and denominator of this intensity ratio ensures period robustness, and therefore provides an efficient means of correcting circadian disorders. Then quantitative RT-PCR and LumiCycle assays were used to confirm our predictions. This work reveals a “frequency modulation” construction with negative and positive feedback and provides an example of how to flexibly tune the period of an oscillator through the transcriptional regulations.

Session H2：Highlight Track 14:00-15:40 October 25 (Saturday)

14:00-14:20 *MetalExplorer, a bioinformatics tool for the improved prediction of eight types of metal-binding sites using a random forest algorithm with two-step feature selection*

Jiangning Song

Monash University

Abstract: Metalloproteins are highly involved in many biological processes, such as catalysis, recognition, transport,

transcription, and signal transduction. The metal ions they bind usually play key roles in mediating these diverse functional roles. Thus, the systematic analysis and prediction of metal-binding sites using sequence and/or structural information are crucial for understanding their sequence-structure-function relationships.

In this work, we develop MetalExplorer, a new machine learning-based method for predicting eight different types of metal-binding sites (Ca, Co, Cu, Fe, Ni, Mg, Mn, and Zn) in protein structures, which combines heterogeneous sequence-, structure-, and residue contact network-based features. The predictive performance of MetalExplorer was tested by cross-validation and independent tests using non-redundant datasets of known structures. This method applies a two-step feature selection approach based on the maximum relevance minimum redundancy and forward feature selection to identify the most informative features that contribute to the prediction performance. With a precision of 60%, MetalExplorer achieved high recall values, which ranged from 59% to 88% for the eight metal ion types in fivefold cross-validation tests. Moreover, the common and type-specific features in the optimal subsets of all metal ions were characterized and analyzed in terms of their contributions to the overall predictive performance. MetalExplorer compared favorably with a state-of-the-art tool SitePredict in terms of both the benchmark and independent datasets at the 60% precision control level. Thus, MetalExplorer is expected to be a powerful tool for the accurate prediction of potential metal-binding sites and it should facilitate the functional analysis and rational design of novel metalloproteins.

14:20-14:40 *NeSSM: A Next-Generation Sequencing Simulator for Metagenomics*

Ben Jia and **Chaochun Wei**
Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University

Abstract: Background: Metagenomics can reveal the vast majority of microbes that have been missed by traditional cultivation-based methods. Due to its extremely wide range of application areas, fast metagenome sequencing simulation systems with high fidelity are in great demand to facilitate the development and comparison of metagenomics analysis tools. Results: We present here a customizable metagenome simulation system: NeSSM (Next-generation Sequencing Simulator for Metagenomics). Combining complete genomes currently available, a community composition table, and sequencing parameters, it can simulate metagenome sequencing better than existing systems. Sequencing error models based on the explicit distribution of errors at each base and sequencing coverage bias are incorporated in the simulation. In order to improve the fidelity of simulation, tools are provided by NeSSM to estimate the sequencing error models, sequencing coverage bias and the community composition directly from existing metagenome sequencing data. Currently, NeSSM supports single-end and pair-end sequencing for both 454 and Illumina platforms. In addition, a GPU (graphics processing units) version of NeSSM is also developed to accelerate the simulation. By comparing the simulated sequencing data from NeSSM with experimental metagenome sequencing data, we have demonstrated that NeSSM performs better in many aspects than existing popular metagenome simulators, such as MetaSim, GemSIM and Grinder. The GPU version of NeSSM is more than one-order of magnitude faster than MetaSim. Conclusions: NeSSM is a fast simulation system for high-throughput metagenome sequencing. It can be helpful to develop tools and evaluate strategies for metagenomics analysis and it's freely available for academic users at http://cbb.sjtu.edu.cn/,ccwei/pub/software/NeSSM.php.

14:40-15:00 *Novel Bioinformatics Method for Identification of Genome-Wide Non-Canonical Spliced Regions Using RNA-Seq Data*

Yongsheng Bai
The Center for Genomic Advocacy (TCGA), Department of Biology, Indiana State University

Abstract: During endoplasmic reticulum (ER) stress, the endoribonuclease (RNase) Ire1α initiates removal of a 26 nt region from the mRNA encoding the transcription factor Xbp1 via an unconventional mechanism (atypically within the cytosol). This causes an open reading frame-shift that leads to altered transcriptional regulation of numerous downstream genes in response to ER stress as part of the unfolded protein response (UPR). Strikingly, other examples of targeted, unconventional splicing of short mRNA regions have yet to be reported.

Our goal was to develop an approach to identify non-canonical, possibly very short, splicing regions using RNA-Seq data and apply it to ER stress-induced Ire1α heterozygous and knockout mouse embryonic fibroblast (MEF) cell lines to identify additional Ire1α targets.We developed a bioinformatics approach called the Read-Split-Walk (RSW) pipeline, and evaluated it using two Ire1α heterozygous and two Ire1α-null samples. The 26 nt non-canonical splice site in Xbp1 was detected as the top hit by our RSW pipeline in heterozygous samples but not in the negative control Ire1α knockout samples. We compared the Xbp1 results from our approach with results using the alignment program BWA, Bowtie2, STAR, Exonerate and the Unix "grep" command. We then applied our RSW pipeline to RNA-Seq data from the SKBR3 human breast cancer cell line. RSW reported a large number of non-canonical spliced regions for 108 genes in chromosome 17, which were identified by an independent study. We conclude that our RSW pipeline is a practical approach for identifying non-canonical splice junction sites on a genome-wide level. We demonstrate that our pipeline can detect novel splice sites in RNA-Seq data generated under similar conditions for multiple species, in our case mouse and human.

15:00-15:20 *SMAL: A Resource of Spontaneous Mutation Accumulation Lines*

Wen Wei, Lu-Wen Ning, Yuan-Nong Ye, Shi-Jie Li, Hui-Qi Zhou, Jian Huang, and **Feng-Biao Guo***
Center of Bioinformatics and Key Laboratory for NeuroInformation of the Ministry of Education, School of Life

Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China.

Abstract: Mutation is the ultimate source of genetic variation and evolution. Mutation accumulation (MA) experiments are an alternative approach to study de novo mutation events directly. We have constructed a resource of the Spontaneous Mutation Accumulation Lines (SMAL, http://cefg.uestc.edu.cn/smal), which now contains all the current publicly available MA lines identified by high-throughput sequencing. We have relocated and mapped the mutations based on the most recent annotations. A total of 5608 single base mutations and 540 other mutations were obtained and are recorded in the current version of the SMAL database. The integrated data in SMAL provide further more detailed information that can be used in new theoretical analyses. We believe the SMAL resource will help researchers better understand the processes of genetic variation and the incidence of disease.

15:20-15:40 *Compendium of Protein Lysine Modifications: from acetylation, ubiquitination to new modifications*

Zexian Liu and Yu Xue

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

Abstract：Background: Recently, lysine was discovered as a hot spot amino acid for the protein post-translation modifications. Besides relatively well-studied protein lysine modifications (PLMs) such as methylation, acetylation and ubiquitination, a number of new PLMs were discovered to modify lysine residue, for example, butyrylation, crotonylation and succinylation. Although the detailed regulatory mechanisms are far from understanding, it is anticipated that these PLMs play critical roles in various biological processes. The curation of the experimental progresses of PLMs will be helpful for further studies.

Methods: Previously, we developed the CPLA database to maintain the lysine acetylation data from scientific literature (1). Here, we updated it as CPLM (Compendium of Protein Lysine Modification, http://cplm.biocuckoo.org) to reserve the data for PLMs (2). The experimentally identified substrates and sites for 12 types of PLMs from published literature in PubMed were manually collected, including acetylation, ubiquitination, methylation, sumoylation, propionylation, butyrylation, succinylation, crotonylation, glycation, malonylation, and pupylation.

Results: In total, 203,972 modification events on 189,919 modified lysines in 45,748 proteins for 122 species were collected and integrated into the CPLM databse. Most of the residues were modified by well-studied PTMs including ubiquitination (139,950 sites in 32,429 proteins) and acetylation (58,563 sites in 20,088 proteins). The third PLM with most substrates is succinylation, which was discovered as a novel PLM in 2011 and identified with 2,523 sites in 897 substrates. From the dataset, we totally identified 76 types of PLM cooccurrences at same lysine residues, including 40 types of pairwise crosstalks and 36 types of multiple (three or more) crosstalks. We observed that the pairwise crosstalks among acetylation, ubiquitination and succinylation are mostly abundant.

Conclusions: Since these PLMs attracted great attention recently, we anticipate that such a comprehensive resource will be useful for the research community.

Session H3：Highlight Track 16:20-18:00 October 25 (Saturday)

16:20-16:40 *Precision medicine: translating genomics to clinical applications using networks and ontologies*

Yves A. Lussier

The University of Arizona

Abstract: Translating biomolecular modules to clinical practice remains challenging, especially when their constituent molecular organization is discovered from analysis and integration of genomic, transcriptomic and proteomic assays. This presentation describes how the sciences of computation and genomics can further individualize therapy with a case report of biologically validated tumor genes and a clinical trial of a repositioned therapy. I will demonstrate new approaches to interpret an individual human transcriptome (mRNA), beyond classical genomic (DNA) level analyses. For example, network modeling identified synergy between mTOR and EGFR, with potential to overcome resistance and aberrations in the PI3K-MTOR pathway that occur in ⩾30% of head adenocarcinomas (subsequent NIH-funded clinical trial). “Ontology transforms,” which we define as the often-overlooked model transformations that can be performed over ontology-encoded datasets. The implication of moving to personal prediction integrating the dynamics of the genome (mRNA) with its blueprint (DNA) are numerous: from peripheral blood “omics provocation” essays to predicting individual response to therapy.

16:40-17:00 *MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes*

Xiaoqi Zheng, Qian Zhao, Hua-Jun Wu, et al.

Shanghai Normal Univerisity, China

Abstract: Tumor impurity has been a major technical issue in tumor profiling studies. We propose a statistical algorithm MethylPurify that uses regions with bisulfite reads showing discordant methylation levels to infer tumor purity from tumor samples alone. With the purity estimates, MethylPurify can identify differentially methylated regions (DMRs) from individual tumor samples. It is the first computational method to estimate tumor purity and make differential DNA methylation calls from tumor methylome data alone, without genomic variation information or prior knowledge from other datasets. In simulations with mixed bisulfite reads from cancer and normal cell lines, MethylPurify correctly inferred tumor purity and identified over 96% of the DMRs. On real patient data where tumor to normal comparison were used as golden standard, MethylPurify gave satisfactory DMR calls from tumor samples alone. Comparison with TCGA methylation results further suggests that DMRs called from tumor samples alone are equally accurate as the tumor to normal comparison, and included DMRs missed by the latter due to tumor heterogeneity.

17:00-17:20 *Prediction, Prevention and Treatment of CNS Metastases*

Xuefeng Bruce Ling
Stanford University, USA

Abstract: CNS metastases are a devastating cancer complication that afflicts hundreds of thousands of patients in the US every year. It represents a major cause of morbidity and mortality, not only due to the cancer itself but also to the treatments (i.e., irradiation) that must be used to control them. Better strategies are therefore urgently needed to either prevent this complication from developing or to effectively treat them once they develop. We hypothesize, based on several experimental observations made both here at Stanford and by others, a more unified sequence of events in which a premetastatic niche is formed by invasion of myeloid type cells that creates a favorable environment for cancer cell implantation. We therefore subdivide the CNS metastatic process into three intervention points: (i) prior to the formation of the niche, (ii) establishment of myeloid niches (premetastasis) and (iii) overt metastasis. An interdisciplinary group at Stanford has been formed that will dedicate itself to finding more effective strategies for managing CNS metastases. We have three independent projects that will assess whether tumors destined to produce CNS metastases have a characteristic genetic signature (Project 1), whether early metastases can be assessed at the niche stage using a novel MR imaging technique that can identify macrophages with excellent resolution (Project 2), and whether experimental CNS metastases can be more effectively and safely treated via chemokine manipulation (Project 3). Considering the importance of CNS involvement to essentially all cancer disciplines, we believe our work has the potential to positively impact cancer care for nearly all cancer patients.

17:20-17:40 *Development and validation of a novel computational approach to identify epigenetic biomarkers associate with cancer prognosis*

Li Xu, Xue Xiao and Shanguang Chen
Harbin Institute of Technolog

Abstract: Background: DNA methylation is one of the essential epigenetic mechanisms that are closely correlated with the prognosis progression in human cancers. Advances in computational approaches have generated many candidate biomarkers with acceptable accuracies which are being applied into cancer outcome prediction and personalization of therapy to improve patient care. However, the identification of DNA methylation signatures in cancer prognosis is still being explored. Besides, it is still not yet proved whether the prognosis related DNA methylation biomarkers could appropriate be proposed for cancer prognosis prediction, nor whether their aberrant methylation levels have the potential to affect coding/non-coding genes' expression leading to different prognosis outcome.

Methods: We developed a machine learning based feature selection approach to identify prognosis related DNA methylation signatures using approximately 4012 tumor samples crossing 10 type of human cancers (LUSC, UCEC, OV, BRCA, GBM, LUAD, KIRC, HNSC, READ, COAD) which is freely available at The Cancer Genome Atlas (TCGA) datasets. Six different classification/regression algorithms (svm, naiveBayes, knn1,3,5, lasso, randomForest, lda) were proposed to build a performance baseline for extracting the best predictive Prognosis-related-CpGs (PR-CpGs). The impact of PR-CpGs on Overall Survival (OS) was assessed using multi/uni-variate Cox regression and Kaplan-Meier analyses. The effect of prognosis related methylation on gene expression were further investigated using RNASeq profiles of the same tumor samples.

Results: We identified the best performance predictor sets of 290 PR-CpGs with optimal bootstrap accuracy in both training and testing set in 10 human cancers. Among the PR-CpGs, 58.6% were detected in promoter region, 27.58% in 5'UTR, 6.8% in 1st Exon and 6.8% in body region. Correlation analysis identified specific PR-CpGs which were highly associated (p adjusted under 0.05) with clinical trails in each cancer, respectively. A highly significant difference (logrank test, $p < 0.05$) in survival was observed between High Risk (HR) and Low Risk (LR) outcome groups in all PR-CpGs sets using univariate Cox analysis. Notably, patients with high level methylation had a significantly better OS (86% vs 34%, logrank $p < 0.001$) compared with less methylated cases. Of all 290 PR-CpGs, 56% were found with matching transcripts in RNASeq profiles, 88% in protein coding regions and 12% in non-coding regions. Importantly, 66% PR-CpGs in promoter region showed a repressive relationship with their protein coding genes of the same tumor samples, whereby hypo-methylated genes IL12RB and SEMA3B were identified in 87% and 79% of the tumors were associated with poor survival.

Conclusion: Our study proposed a comprehensive approach to identify a panel of prognostic DNA methylation biomarkers which can successfully predict cancer outcomes for the first time. We also discovered the association of the PR-CpGs with specific clinical features in each cancer, respectively. In addition, the validation of epigenetic changes on gene activity indicated that the expression of PR-CpGs related genes were negative associated with their promoter DNA methylation and could directly lead to differential prognosis outcome and validated the effectiveness of PR-CpGs in biological roles. The R package of our approach is freely available at http://www.escience.cn/people/lixu/index.html.

17:40-18:00 *Breast tumor subgroups reveal diverse clinical prognostic power*

Zhaoqi Liu, Xiang-Sun Zhang & Shihua Zhang
National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Predicting the outcome of cancer therapies using molecular features and clinical observations is a key goal of cancer biology, which has been addressed comprehensively using whole patient datasets without considering the effect of tumor heterogeneity. We hypothesized that molecular features and clinical observations have different prognostic abilities for different cancer subtypes, and made a systematic study using both clinical observations and gene expression data. This analysis revealed that (1) gene expression profiles and clinical features show different prognostic power for the five breast cancer subtypes; (2) gene expression data of the normal-like subgroup contains more valuable prognostic information and survival associated contexts than the other subtypes, and the patient survival time of the normal-like subtype is more predictable based on the gene expression profiles; and (3) the prognostic power of many previously reported breast cancer gene signatures increased in the normal-like subtype and reduced in the other subtypes compared with that in the whole sample set.

Session H4：Highlight Track 10:50-12:30   October 26 (Sunday)

10:50-11:10   Transforming Trillions of Points of Data into Diagnostics, Therapeutics, and New Insights into Disease

Atul Butte
Department of Pediatrics, Stanford University School of Medicine

Abstract: With the end of the United States NIH budget doubling and completion of the Human Genome Project, there is a need to translate genome-era discoveries into clinical utility. The difficulties in making bench-to-bedside translations have been well described. The nascent field of translational bioinformatics may help. Dr. Butte's lab at Stanford builds and applies tools that convert trillions of points of molecular, clinical, and epidemiological data -- measured and often released to the public by researchers and clinicians over the past decade and now colloquially termed “big data” -- into diagnostics, therapeutics, and new insights into disease. Dr. Butte, a bioinformatician and pediatric endocrinologist, will highlight his lab's work on using publicly-available molecular measurements to find new uses for drugs and medically evaluating large populations now with whole genomes sequenced.

11:10-11:30   *The systematice approach to cancer chemoresisance for better mechanistic understanding and DNA methylation diagonstics, a personal journey bigining with the discoveries made from the integrative multi-omic analysis to the robust assays fit to the clinical practice*

Jingde Zhu
Anhui Cancer Hospital Hefei and Shanghai Cancer Institute, Shanghai, China

Abstract: Cancer is a highly heterogeneous disease of the dysproliferating cells, suffering from extensive genomic and epigenomic defects. DNA methylation is the best characterized epigenetic mechanism, ensuring for the lasting transcription memory maintained across the cell divisions. The aberrant DNA methylation dictates the cancer-state specific changes of transcription and therefore phenotype and is a promising molecular target for cancer diagnosis. Chemo-resistance prevents curative chemotherapy of cancer. The focus of our research to develop and clinical use of the robust DNA methylation diagnostics to detect early and guide better of personalize chemotherapy of cancer. In this talk, I will use our studies in bladder cancer as a paradigm to demonstrate the robustness and promise of both approaches and discoveries for an effective cancer management.

11:30-11:50 *A Systems Kinetic Metbabolic Model for Xiamenmycin Biosynthetic Pathway*

Minjuan Xu, Yong-Cong Chen, Xiao-Mei Zhu, Jun Xu and Ping Ao
Shanghai Jiao Tong University

Abstract: Viability is a key constraint for metabolic engineering. Xiamenmycin is a leading drug candidate for anti-fibrosis. A kinetic metabolic model based on generic enzymatic rate equations was constructed and used to evaluate fluxes in engineered S. lividans with xiamenmycin-oriented genetic modification. Lyapunov function was used in the optimization of stability on a metabolic network. Biologically it represents a viability optimization. We simulated stoichimetrically and kinetically the production levels under two types of nutrient conditions for the potential of secondary metabolite production. The model reveals the links between primary and secondary metabolism in the biosynthesis of xiamenmycin and provides guides to enhance the production. Our approach is the first metabolic model with viability constraint.

11:50-12:10   *Stochastic modelling of biochemical systems of multi-step reactions using a simplified two-variable model*

Qianqian Wu, Kate Smith-Miles, Tianshou Zhou and Tianhai Tian
Monash University

Abstract: A fundamental issue in systems biology is how to design simplified mathematical models for describing the dynamics of complex biochemical reaction systems. In particular, a key question is how to use simplified reactions to describe the chemical events of multi-step reactions that are ubiquitous in biochemistry and biophysics. The widely used approach in literature is to use one-step reaction or reaction with time delay to represent the multi-step chemical events. However, the accuracy of these approaches is not satisfactory. This work designs a novel two-variable model to simplify chemical events of multi-step reactions. In addition to the total molecule number of a species, we introduce a new concept regarding the location of molecules in the multi-step reactions, which is the second variable to represent the system dynamics. We propose a simulation algorithm to compute the probability for the firing of the last step reaction in the multi-step events, which is evaluated using a deterministic model of ordinary differential equations and a stochastic model in the framework of the stochastic simulation algorithm. The efficiency of the proposed two-variable model is demonstrated by the realization of mRNA degradation process based on the experimentally measured data. Numerical results suggest that the proposed new two-variable model produces predictions that match the multi-step chemical reactions very well. The successful realization of the mRNA degradation dynamics indicates that the proposed method is a promising approach to reduce the complexity of biological systems.

12:10-12:30 *Characterizing and controlling the inflammatory network during influenza A virus infection*

Suoqin Jin, Yuanyuan Li, Ruangang Pan, Xiufen Zou,
School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China.

Abstract: To gain insights into the pathogenesis of influenza A virus (IAV) infections, this study focused on characterizing the inflammatory network and identifying key proteins by combining high-throughput data and computational techniques. We constructed the cell-specific normal and inflammatory networks for H5N1 and H1N1 infections through integrating high-throughput data. We demonstrated that better discrimination between normal and inflammatory networks by network entropy than by other topological metrics. Moreover, we identified different dynamical interactions among TLR2, IL-1β, IL10 and NFκB between normal and inflammatory networks using optimization algorithm. In particular, good robustness and multistability of inflammatory sub-networks were demonstrated. Furthermore, we identified a complex, TNFSF10/HDAC4 /HDAC5, which may play important roles in controlling inflammation, and demonstrated that changes in network entropy of this complex negatively correlated to those of three proteins: TNF □-2, NFκB and COX. These findings provide significant hypotheses for further exploring the molecular mechanisms of infectious diseases and developing control strategies.

Workshop abstracts

Workshop WA1 Translational biomedical and clinical informatics

14:00-14:20 *Evidence Based Disease Network Construction towards Drug Repositioning*

Liwei Wang, Jiabei Wang and Qian Zhu
Department of Medical Informatics, Jilin University, Changchun, China
Paper ID: 140

Abstract: I. BACKGROUND Drug repositioning (DR) is one of emerging approaches dedicated to find alternative usages of existing drugs efficiently and economically, especially with the advance in computational technology. The current progress made for computational DR is primarily focusing on informatics approach development/improvement or exploration on different type of data in order to identify possible drug candidates. Network analysis, as one of computational approaches, has gained much popularity for DR with promising findings. In addition, lots of biomedical knowledge is buried in a large amount of literatures from biomedical databases, such as Semantic MEDLINE (SM), while currently only little research attempted to apply or extend their findings identified from SM for DR. II. MATERIALS AND METHODS Comparing to the existing studies, we proposed a novel method for constructing the disease based network (DBN) by applying data extracted from SM. Phenotypical associations (disease-disease associations) can be identified from this network, which can drive drug repositioning study by targeting on specific domain. Phenotypical association extraction from SM We obtained existing phenotypical associations by using semantic types of concepts from SM, as well as inferred non-exisiting phenotypical associations through the two-step inference algorithm. Phenotypical association prioritization With a large number of phenotypical associations, especially for the inferred ones extracted from the SM, we performed prioritization steps to screen the most significant phenotypical associations via three steps, 1) based on the semantic types of Concept B defined in the inference rule, 2) based on the predicates associated with the inferred associations, and 3) based on the weights assigned to the phenotypical associations. Disease based network (DBN) construction. All resulted phenotypical associations have been used for constructing the DBN, where the nodes correspond to the disease terms extracted from SM, the edges correspond to the associations, and the weight of the edges. III. RESULTS. There were 160,090 existing phenotypical associations involving 11,896 concepts from SM. Meanwhile, 16,003 concepts were extracted through the inference rule, consisting of 80,713,255 phenotypical associations. After we applied prioritization steps to phenotypical associations extracted from SM, 2970 unique disease terms and 300,843 phenotypical associations were obtained for DBN construction. Among these inferred associations, 42,982 associations proved to have direct relationship in SM, and the remaining 257,168 pairs revealed the potential new associations between diseases. We will release the DBN when it is ready. Case studies using network analysis approach revealed potential associations in the DBN for further study, demonstrating the validity and feasibility of our method. IV. DISCUSSION. In this study, we introduced a disease network built by integrating phenotypical associations extracted/inferred from literature in SM. Disease orientated network was designed not only to provide comprehensive information specific to diseases that will strongly support epidemiology related study, e.g., finding treatment paradigm upon the new disease pattern identified from the DBN, but also to lead a novel direction towards drug repositioning. Two tags can be highlighted for the DBN, 1) evidence supported, all phenotypical associations were extracted from the SM and can be tracked back to the relevant published manuscripts accordingly. 2) concepts included in the DBN are labeled with standardized identifiers, UMLS CUIs. Data standardization will facilitate future data integration with more additional resources and support cross evaluation with other relevant resources. In the future, we will explore more systematical/intelligent algorithms, such as page rank, topic modeling to assist association selection to overcome the limitation of human interference in case studies. V. CONCLUSION. In this study we have built a DBN through inference in SM. The network offers information on phenotypical associations, supporting medical studies, especially providing a orientation and paradigm to gain insights for drug repurposing. The capability and the potential of the DBN have been successfully demonstrated by case studies.

14:20-14:40 *Network-based Analysis of Time Series RNA-Seq Gene Expression Data by Integrating the Interactome and Gene Ontology Information*

Yuji Zhang
University of Maryland School of Medicine, USA
Paper ID: 137

Abstract: **Background:** Monitoring the changes in gene expression patterns over time provides the distinct possibility of unraveling the mechanistic drivers characterizing cellular responses. Such time series gene expression data allows us to broadly “watch” the dynamics of the system. However, one challenge in the analysis of time series data is to establish and characterize the interplay between genes that are activated, deactivated or sustained in the context of a biological process or functional category. To address such challenges, novel algorithms are required to improve the interpretation of these data by integrating multi-source prior functional evidence. **Methods:** In this paper, we introduced a novel network-based approach to extract functional knowledge from time-dependent biological processes at a system level using time series mRNA deep sequencing data. First, a list of differentially expressed genes (DEGs) at each time point was identified. Second, GO terms that are enriched in each DEG list were identified. Third, the significance of interactions between DEGs in these GO terms at consecutive time points was measured. Finally, the significant interactions between DEGs in different GO terms were used to construct the interaction networks among GO terms between two consecutive time points, called GO networks. **Results:** The proposed method was applied to investigate 1α, 25(OH)2D3-altered mechanisms in zebrafish embryo development. GO networks were constructed over 4 consecutive time points. Results suggest that biological processes such as cartilage development and one-carbon compound metabolic process are temporally regulated by 1α,25(OH)2D3. Such discoveries could not have been identified with canonical gene set enrichment analyses. **Conclusion:** We have developed a network-based approach to analyzing the DEGs at different time points by integrating molecular interactions and gene ontology information. These results demonstrate that the proposed approach can provide insight on the molecular mechanisms taking place in vertebrate embryo development upon treatment with 1α,25(OH)2D3. Our approach enables the monitoring of biological processes that can serve as a basis for generating new testable hypotheses. Such network-based integration approach can be easily extended to any temporal- or condition-dependent genomic data analyses.

14:40-15:00 *Evidence based computational drug repositioning candidate screening pipeline design: Case Study*

Qian Zhu, Yuji Zhang, Hongfang Liu and Jiabei Wang
University of Maryland, Baltimore County, USA
Paper ID: 139

Abstract: Traditional drug development is time and cost consuming process, conversely, drug repositioning is an emerging approach to discover novel usages of existing drugs with a better risk-versus-reward trade-off. Computational technology is playing a key role in drug repositioning to screening the best drug repositioning candidates from a large candidate library. Recent efforts made for computer aided drug repositioning are mostly focusing on applying/developing data mining algorithms against wild type of large scale of biomedical data. In this paper, we introduce a novel computational pipeline designed for drug repositioning candidate screening based on existing phenotypical association (disease-disease association) discovery and pathway enrichment analysis by exploring systems biology data relevant to the interested phenotypical association specifically. To demonstrate usability and evaluate efficacy of this novel pipeline, we successfully conducted a case study by identifying potential drug repositioning candidates for Alzheimer's disease (AD) based on the studied phenotypical association between cancer and AD.

15:00-15:20 *PRECISE:PRivacy-prEserving Cloud-assisted quality Improvement Service in hEalthcare*

Feng Chen, Shuang Wang, Noman Mohammed, Samuel Cheng and Xiaoqian Jiang
School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, USA
Paper ID: 111

Abstract: Quality improvement (QI) requires systematic and continuous efforts to enhance healthcare services. A healthcare provider might wish to compare local statistics with those from other institutions in order to identify problems and develop intervention to improve the quality of care. However, the sharing of institution information may be deterred by institutional privacy as publicizing such statistics could lead to embarrassment and even financial damage. In this article, we propose a PRivacy-prEserving Cloud-assisted quality Improvement Service in hEalthcare (PRECISE), which aims at enabling cross-institution comparison of healthcare statistics while protecting privacy. The proposed framework relies on a set of state-of-the-art cryptographic protocols including homomorphic encryption and Yao's garbled circuit schemes. By securely pooling data from different institutions, PRECISE can rank the encrypted statistics to facilitate QI among participating institutes. We conducted experiments using MIMIC II database and demonstrated the feasibility of the proposed PRECISE framework.

15:20-15:40 *Network cluster analysis of protein–protein interaction network identified biomarker for type 2 diabetes*

Zhonghui Li, Zijun Qiao, Wenli Ma and Wenling Zheng
Southern Medical University, Institute of Genetic Engineering, Guangzhou, China
Paper ID: 92

Abstract: Type 2 diabetes mellitus (T2DM) is a complex disease caused by an impairment in β-cell insulin secretion and by peripheral insulin resistance. The prominent features presented by most patients with T2DM and obesity are the occurrence of insulin resistance in the muscles, liver and fat, resulting in a reduced response of these tissues to insulin. Under normal conditions, the increasing blood glucose levels is regulated by insulin secretion from the pancreatic islet β-cells. Once β-cells fail to function, people will develop T2DM. Despite the progress achieved in recent years in this field, the genetic causes for insulin resistance and T2DM have not yet been fully discovered. The present study aims to characterize type 2 diabetes by analyzing its gene expression compared with normal controls and to identify biomarkers of early type 2 diabetes. The gene expression profiles were downloaded from Gene Expression Omnibus and the differentially expressed genes (DEGs) in type 2 diabetes were identified. Furthermore, the functional analysis of gene ontology and the pathway enrichment analysis were conducted. Total 781 DEGs were identified from the type 2 diabetes samples as compared to healthy controls. These genes were found to be involved in several biological processes, including cell communication, cell proliferation, cell shape and apoptosis. We constructed a protein-protein interaction network and the clusters in PPI were analyzed by ClusterONE. We find six functional modules that may play important roles in the initiation of type 2 diabetes were identified in the network. Then we analyzed the methylation of a microarray of GSE21232 to find weather the six genes have something to do with the DNA methylation.

Workshop WA2 Big Data Study for Bioinformatics

16:00-16:20 *Dissecting the obesity disease landscape: identifying gene-gene interactions that are highly associated with Body Mass Index*

Rishika De, Shefali S. Verma, Michael V. Holmes, Folkert Asselbergs, Jason H. Moore, Brendan Keating, Marylyn D. Ritchie and Diane Gilbert-Diamond
Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH, USA
Paper ID: 76

Abstract: Despite heritability estimates of 40-70% for obesity, less than 2% of its variation is explained by Body Mass Index (BMI) associated loci that have been identified so far. Epistasis, or gene-gene interactions are a plausible source to explain portions of the missing heritability of BMI. Using genotypic data from 18,686 individuals across five study cohorts – ARIC, CARDIA, FHS, CHS, MESA – we filtered SNPs (Single Nucleotide Polymorphisms) using two parallel approaches. SNPs were filtered either on the strength of their main effects of association with BMI, or on the number of knowledge sources supporting a specific SNP-SNP interaction in the context of obesity. Filtered SNPs were specifically analyzed for interactions that are highly associated with BMI using QMDR (Quantitative Multifactor Dimensionality Reduction). QMDR is a nonparametric, genetic model-free method that detects non-linear interactions associated with a quantitative trait. We identified seven novel, epistatic models with a Bonferroni corrected p-value of association < 0.06. Prior experimental evidence helps explain the plausible biological interactions highlighted within our results and their relationship with obesity. We identified interactions between genes involved in mitochondrial dysfunction (POLG2), cholesterol metabolism (SOAT2), lipid metabolism (CYP11B2), cell adhesion (EZR), cell proliferation (MAP2K5), and insulin resistance (IGF1R). This study highlights a novel approach for discovering gene-gene interactions by combining methods such as QMDR with traditional statistics.

16:20-16:40 *Next-Generation Sequencing Data Analysis on Cloud Computing*

Taesoo Kwon, Won Gi Yoo, Won-Ja Lee, Won Kim and Dae-Won Kim
Korea Center for Disease Control and Prevention, Korea
Paper ID: 71

Abstract：With the advent of next-generation sequencing (NGS), including whole genome sequencing (WGS), RNA sequencing (RNA-seq), and chromatin immunoprecipitation followed by sequencing (ChIP-seq), many biologists and computer scientists are highlighting the urgent need for computing power, storage, and various bioinformatics software for analysing large quantities of sequence data. Currently, building the computational infrastructure required for massive data processing and providing maintenance services are among the most important tasks. However, technology platforms for handling large amounts of information pose multiple challenges for data access and processing. To overcome these challenges, cloud computing technologies are emerging as a possible infrastructure for tackling the intensive use of computing power and communication resources in NGS data analysis. Thus, in this review, we explain the concepts and key technologies of cloud computing, such as Map and Reduce, and discuss the problem of data transfer. To reveal the performance and usefulness of these technologies, we analysed NGS data using cloud platforms and compared them with a local cluster. From the benchmark results, we concluded that cloud computing is still more expensive than local cluster, but provides reasonable performance for NGS data analysis with acceptable prices and could be a good alternative to local cluster systems.

16:40-17:00 *Robust high-order gene-gene interaction analysis in Genome Wide Association Studies*

Yongkang Kim and Taesung Park

Seoul national university, Korea
Paper ID: 78

Abstract：Genome-wide association studies (GWAS) have successfully found hundreds of associations between genetic variants and complex traits. Most GWAS have focused on identification of single variants. It was shown that most variants found by GWAS could explain a small part of diseases heritability. Many researchers have pointed out that this

missing heritability might be explained by gene-gene (GG) or gene-environment (GE) interactions and other structural variants. Ritchie et al. (2001) proposed multifactor dimensionality reduction (MDR) method for detection of GG and GE interactions for qualitative traits such as case-control studies. This MDR has been extended to Generalized MDR by Lou et al. (2007) to handle both quantitative and qualitative traits in the presence of covariates. While GMDR has been quite powerful in detecting GG and GE interactions, it may suffer from a few outlying quantitative traits. In this paper, we first demonstrate the effects of outlying traits in GMDR analysis. Next, we propose robust MDR for reducing the effects caused by outlying traits. Robust MDR uses robust estimation based on L-estimator and M-estimator. Our robust GMDR is compared to the original MDR and is shown to perform better through simulation studies. Robust MDR is illustrated through a real GWA example of 8577 samples.

17:00-17:20 *Improving Mental Health using Sentiment Analysis on a Social Network*

Giryong Choi, Hyo Jin Do and Ho-Jin Choi
Korea Advanced Institute of Science and Technology, Korea
Paper ID: 70

Abstract：Healthcare social networks allow doctors, patients, and families to collaborate and share resources online, through which patients can get useful information and receive healthcare guidelines. In this paper, we propose a novel approach for improving mental health of general public using a healthcare social network. Our method consists of three steps. In data analysis step, system analyzes data to provide personalized service. In contents recommendation step, system selects good contents and recommend them to the user. In friends recommendation step, system finds new friend candidates who can give positive influence to the user. We define an index called emotional well-being index (EWI) using the technique of sentiment analysis. We conducted a preliminary experiment as a proof of concept to EWI.

17:20-17:40 *The Correlation and Regression Analysis on Aerosol Optical Depth, Ice Cover and Cloud Cover in Greenland Sea*

Bo Qu, Albert Gabric, Peijuan Gu and Meifang Zeng
Nantong University, China
Paper ID: 119

Abstract：Researches on Arctic aerosol, ice cover and cloud cover have received great attention and it related to the regional even global climate changing. We here study the distributions and the coupling relationships of AOD, cloud cover (CLD) and ice cover (ICE) in the Greenland Sea (20°W-10°E, 70°N-80°N) during 2003-2012. Enhanced statistics methods, such as lag regression method and co-integration analysis method are used for correlation and regression analysis. ARMA model was used to predict AOD time series in the future 3 years. According to the 10 years satellite data, AOD was high in spring, and low in summer. Generally, AOD was higher down south and lower up north. CLD and AOD mainly had negative correlations and ICE and AOD had positive correlations. According to the lag regression analysis by statistical software EViews, both the peaks of CLD and peaks of ICE were all 1 month earlier than the peak of AOD. The co-integration test suggested that both ICE(-1) and CLD(-1) and AOD were all zero-order integration, and there was no unit root in the residual, so there all had long-run equilibrium relationships. ICE and AOD were stationary series, and the residual had no unit root, they were good coupling. The melting of sea ice and decreasing of cloud cover would all result in the increasing of the AOD content. However, the relationship between AOD and CLD was weaker than the relationship between AOD and ICE, indicating that the aerosol in Arctic mostly came from the sea rather than from the air.

Workshop WB1：Network Biology

14:00-14:20 A Novel Markov Chain Modeling Method for Identifying Differential Pathways

Zhirui Zhang and Hong-Qiang Wang
Institute of Intelligent Machines, Chinese Academy of Science, China
Paper ID:59

Abstract: pathway analysis plays an important role in exploring underlying connections between genomic data and complex diseases. In this paper, we propose a gene link‑based method for identification of differentially expressed gene pathways. By viewing gene links in a pathway as a Markov chain, the proposed method first develops a gene link Markov chain model (MCM) and devises a Markov chain model-based classification rule to measure the biological importance of a gene link. Then, the expression difference of a pathway is estimated based on all the gene links in the pathway using the gene link MCM. The use of gene links, instead of individual genes, allows for exploring pathway topology that is crucial to pathway activity in cells. Results on two real-world gene expression data sets demonstrate that the effectiveness and efficiency of the proposed method in identifying differential gene pathways.

14:20-14:40 *Network-based detection of Disease Modules and Potential Drug Targets in Intractable Epilepsy*

Hongwei Chu, Changkai Sun, Xuezhong Zhou, Guangming Liu, Lin Liu, Minghui Lv, Xiaofeng Zhou, Yiwei Wang, Xing Li, Pin Sun and Yizhun Zhu

Liaoning Provincial Key Laboratory of Cerebral Diseases, Institute for Brain Disorders Dalian Medical University

Paper ID: 80

Abstract: Epilepsy is one of the common nervous system diseases and a complex brain disease that severely damages the life and health of humans. One-third of all epilepsy patients have medically intractable epilepsy (IE), for which anti-epileptic drugs are not effective. Therefore, discovery of potential drug targets is urgent and meaningful for better clinical solutions. Using the IE terms from Medical Subject Headings (MeSH) terminology, we integrated literature-based disease-gene relationships, which were extracted from the CoreMine PubMed search engine system, protein-protein interactions (PPI) and drug-target relationships from heterogeneous data sources, and used the network medicine approach to identify disease modules and detect enriched pathways. The potential drug targets and the underlying mechanisms were confirmed by chemical-protein interaction network and published literatures. Using 23 IE MeSH terms, we manually searched the CoreMine system to obtain 1,400 disease- gene associations, which had 871 distinct genes from the PubMed database. With the help of the PPI database (i.e. String 9), we mapped the genes to the PPI network and utilized the BGL community detection method to find 33 disease-related topological PPI modules that contain 640 proteins and 2,483 links. After that, we used the enrichment analysis method to obtain the PPI modules with pathway and gene ontology enrichment. Finally, we confirmed nine significant PPI modules that are considered as epilepsy disease modules with significant functional signatures. We combined genes with drugs in the DrugBank database to confirm the four proteins, MT-CYB, UQCRB, UQCRC1 and UQCRH, which would be potential drug targets for IE. The results of this study demonstrated that integrated network data sources and network-based approach are useful to understand the molecular mechanism and extract potential drug targets for IE.

14:40-15:00 *Detection of SNP-SNP Interaction based on the Generalized Particle Swarm Optimization Algorithm*

Ma Changyi, Shang Junliang, Li Shengjun and Sun Yan

School of Information Science and Engineering, Qufu Normal University, China

Paper ID: 94

Abstract: Most of complex diseases, such as cancer, heart disease, and diabetes, are believed to be mainly caused by epistatic interactions of two single nucleotide polymorphisms (SNPs), namely, SNP-SNP interactions. Though many works have been done for the detection of SNP-SNP interactions, the algorithmic development is still ongoing due to their mathematical and computational complexities. In this study, we proposed a method, PSOMiner, based on the generalized particle swarm optimization algorithm for the detection of SNP-SNP interaction that has the highest pathogenic effect in a data set. Experiments of PSOMiner are performed on lots of simulation data sets under the criteria of detection power. Results demonstrate that PSOMiner is promising for the detection of SNP-SNP interaction. PSOMiner might be an alternative to existing methods for detecting SNP-SNP interactions.

15:00-15:20 *cLP: Linear Programming with Biological Constraints and its Application in Classification Problems*

Manli Zhou, Youxi Luo, Guoqin Mai and Fengfeng Zhou

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, P.R. China

Paper ID:91

Abstract: Feature selection represents a major challenge in the biomedical data mining problem, and numerous algorithms have been proposed to select an optimal subset of features with the best classification performance. However, the existing algorithms do not take into account the vast amount of biomedical knowledge from the literature and experienced researchers. This work proposes a novel feature selection algorithm, cLP, with the optimized binary classification accuracy. The proposed algorithm incorporates the biomedical knowledge as constraints in the linear programming based optimization model. The experimental data shows that cLP outperforms the other feature selection algorithms, and its constrained version performs similarly well with the unconstrained version. Although theoretically constraints will reduce the classification model performance, our data shows that the constrained cLP sometimes even outperforms the unconstrained version. This suggests that besides the benefit of including biomedical knowledge in the model, the constrained cLP may also achieve better classification performance.

15:20-15:40 *Comparative genomics reveals a global map of selenium utilization and evolution in prokaryotes*

Ting Peng, Jie Lin and Yan Zhang

Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

Paper ID: 88

Abstract: Selenium (Se) is an important micronutrient that is used in a wide variety of biological processes in many organisms. The known biological forms of Se in prokaryotes include selenocysteine, selenouridine and the recently identified Se- containing cofactor. In the recent decade, several key genes involved in different Se utilization traits have been characterized; however, systematic studies on the occurrence and the evolution of these Se utilization traits are very limited. In this work, we analyzed more than 5200 bacterial and archaeal genomes to examine the occurrence of different Se traits in prokaryotes. First, a global species map of Se utilization has been provided based on the occurrence pattern of key genes involved in each Se utilization trait, which demonstrates the most detailed understanding of Se utilization in different phyla/clades of prokaryotes so far. A highly mosaic distribution of species that use Se (in different forms) was observed in spite that most organisms do not use this element. The three known Se traits showed different relationship with each other in bacteria and archaea, implying somewhat different evolutionary trends of Se utilization between the two kingdoms. In addition, the selenophosphate synthetase gene, which is used to define the overall Se utilization, is also detected in some organisms that do not use any of the known Se forms, indicating that a novel and unknown Se utilization trait may be present in prokaryotes. Further phylogenetic analysis revealed that frequent horizontal gene transfer events could be observed for each Se utilization trait, but co-transfer of two or three Se traits could not be detected yet. Finally, we characterized the selenoproteomes of completely sequenced Sec-utilizing organisms in selenoprotein- rich phyla. In conclusion, our data provide more information about Se utilization in prokaryotes and should be useful for a further understanding of the evolutionary dynamics of Se in nature.

Workshop WB2：Computational Systems Biology

16:00-16:20 *Incorporating feature reliability in false discovery rate estimation improves statistical power to detect differentially expressed features*

Elizabeth Chong, Yijian Huang, Hao Wu, Tianwei Yu, Dean Jones, Arshed Quyyumi, Karan Uppal and Nima Ghasemzadeh

Department of Biostatistics and Bioinformatics, Emory University, USA

Paper ID: 108

Abstract: Feature selection is a critical step in translational omics research. False discovery rate (FDR) is an integral tool of statistical inference in feature selection from high-throughput data. It is commonly used to screen features (SNPs, genes, proteins, or metabolites) for their relevance to the specific clinical outcome under study. Traditionally, all features are treated equally in the calculation of false discovery rate. In many applications, different features are measured with different levels of reliability. In such situations, treating all features equally will cause substantial loss of statistical power to detect significant features. Feature reliability can often be quantified in the measurements. Here we present a new method to estimate the local false discovery rate that incorporates feature reliability. We also propose a composite reliability index for metabolomics data. Combined with the new local false discovery rate method, it helps to detect more differentially expressed metabolites that are biologically meaningful in a real metabolomics dataset.

16:20-16:40 *Bioinformatic Inference of Changes in Levels of Reactive Oxygen Species and Their Carcinogenic Effects in Papillary Thyroid Carcinoma with Hashimoto Thyroiditis*

Jin Wook Yi, Sang Huyk Kwak, Jo-Heon Kim, Eun Kyung Paik, Ji-Youn Sung, Jihan Yu, Ji Hyun Chang, Sang Yun Ha, Kyu Eun Lee, Yeo-Kyu Youn and Ju Han Kim

Department of Surgery, Seoul National University Hospital, Seoul, Korea

Paper ID: 20

Abstract: Elevation of reactive oxygen species level has been proposed as a risk factor for the development of thyroid cancer in patients with Hashimoto thyroiditis. However, no experimental or biological evidence has yet clearly demonstrated that the levels of reactive oxygen species increase to sufficient levels to contribute to carcinogenesis in this context. Because experimental measurement of reactive oxygen species is complicated and technically challenging, we investigated this issue using a bioinformatics approach. Specifically, we analyzed data from The Cancer Genomic Atlas, an open-access repository of RNA- sequencing results: 33 datasets from normal thyroid tissue, 232 from papillary thyroid carcinoma without Hashimoto thyroiditis, and 60 from papillary thyroid carcinoma accompanied by Hashimoto thyroiditis. Differential expression analysis revealed that 28 genes related to reactive oxygen species, including several well-known antioxidant-related genes, were highly expressed in the Hashimoto thyroiditis group. Gene Set Enrichment Analysis revealed that the reactive oxygen species gene set was significantly enriched only in the Hashimoto thyroiditis group. These results suggest that the levels of reactive oxygen species increase and contribute to thyroid carcinogenesis in patients with Hashimoto thyroiditis.

16:40-17:00 *In silico analysis of mutations in PITX3 gene*

Abida Arshad, Rashda Abbasi, Christian Sieber, Muhammad Arshad and Nafees Ahmad
Department of Zoology, PMAS Arid Agriculture University, Rawalpindi, Pakistan

Paper ID: 29

Abstract: PITX3 belongs to a class of heomeodomain transcription factors involved in the development of dopaminergic neurons and ocular lens. Despite a great degree of homology, the mutation in human and mouse Pitx3 gene exhibit differences in the range and extent of phenotypic effects. The current study was designed to predict the effect of mutations in the mouse and human PITX3 gene using in silico tools. We used publically available bioinformatics tools to identify the secondary structure, functional domains, three-dimensional structure and DNA binding residues. Analysis of functional domains in the PITX3 revealed a lack of OAR domain in the G219fs mutation and in the mouse eyeless mutation. There was no difference in the functional motifs of the S13N and K111E mutation compared to the wild- type PITX3. However, an additional helix-turn-helix (HTH) domain is predicted in K111E mutation. Comparison of three-dimensional structures of the wild- type and mutant proteins did not show significant differences except 220delG. The eyeless mouse mutant protein exhibited a very different structure compared to the wild-type mouse Pitx3. Our results indicate that three- dimensional structure of the protein is a good predictor of the in vitro and in vivo behavior of the PITX3 protein and provides guidelines for performing the functional assays of the mutant proteins.

17:00-17:20 *A New Staging Framework by Fusion Molecular and Clinical Variables through CART Model*

Hongmin Cai and Ying Jin
School of Computer Science & Engineering, South China University of Technology Guangzhou, China

Paper ID: 39

Abstract: Gastric cancer is one of the most lethal cancers in China and its classification is currently based on standard TNM system. A few molecular proteins have been reported to possess prognostic and are promising to subside for clinical variables in identifying subgroup of patients suffering some common cancers. However, few work has been done in gastric cancer, possibly due to the great verity of such patients. To find new staging scheme for gastric patients, this study proposed a new framework to the requirements. The first step in the proposed framework is to select variables by method of Local Hyperplane based RELIEF[1, 2] to have a prognostic panel. It is then fed to the revised classification and regression tree (CART) to search for staging patterns, directed by a new “impurity” criterion. Finally, an original staging scheme is obtained by fusing the predication after CART and clinical variables using expert- marking strategy. Experimental results on a dataset of empirical gastric cancer demonstrated the high prognostics power and its potential sub-staging capabilities.

17:20-17:40 *Evolution Analysis for HA Gene of Human Influenza A H3N2 Virus (1990 - 2013)*

Su-Li Li, Meng-Zhe Jin and Zhao-Hui Qi
College of Information Science and Technology, Shijiazhuang Tiedao University Shijiazhuang, China

Paper ID: 55

Abstract: Human influenza A virus is an important pathogen which threatens the health of human in a long time. The mutation study of HA gene is the most important. Here we investigate the evolution characteristics of HA gene of H3N2 influenza virus from 1990 to 2013. Numerical mapping and PCA clustering analysis are applied to the gene evolution analysis. The clustering diagram by MATLAB represents the mapping of HA gene in 2D space. The first two principal components account for 78.48% by PCA analysis. And the points are clustered into three parts, 1990~1999, 2000~2005 and 2006~2013. However, there is no obvious interval among them. Then we show the graphical representation of HA gene sequences according to the emerging time of isolates and different continents. Results show that during 1990 to 2013 human influenza A H3N2 virus has been evoluting gradually. There was not large genetic recombination. Even so, it is necessary to continuously monitor the human influenza A (H3N2) viruses.

17:40-18:00 *Selecting Representative Topics in Biomedical Research Articles using MeSH Descriptors*

Chae-Gyun Lim, Byeong-Soo Jeong and Ho-Jin Choi
Kyung-Hee University, Korea

Paper ID: 73

Abstract: Due to the huge number of research articles in the biomedical domain, it becomes more and more important to develop methods to find relevant articles of our specific research interests. Latent Dirichlet allocation (LDA), a probabilistic topic modeling technique for unsupervised analysis of text, can be used to find related articles based on the representation of topics latent in the articles. Unfortunately, however, it is very difficult to understand the proper meanings of the topics in the output of topic modeling because topics are represented

simply as words distribution. Human experts may help interpret the topics, but in general too many topics and vocabularies are generated while learning a topic model. In this paper, we propose a method for selecting representative topics for a given set of biomedical research articles from PubMed. Our method extracts significant MeSH vectors as the descriptors for each topic in the learned LDA model. The vectors represent the frequencies of MeSH terms indexed in those articles that are high-ranked on particular topics, and hence are used to determine the topics representative to a specific subfield. In the experiments, we evaluated the performance of the method by measuring the perplexities of the representative topics found and the accuracy of recommendation for particular subfields given unseen documents.

Poster abstracts

Poster session

Poster01: *Detecting Causal Gene Regulations from Short Time-series Data based on Prediction of Topologically Equivalent Attractors*

Ben-gong Zhang, Xiaoping Liu, Huanfei Ma, Kazuyuki Aihara, Luonan Chen
School of Mathematics & Computer Science, Wuhan Textile University, Wuhan 430073, China

Paper ID: 38

Abstract: Detecting causal relation from multivariate time-series is an important but difficult task. It can be widely used in ecological systems, economics, complex networks, biological systems and many other fields. Many effective methods have been proposed to study this problem. However, to detect causal relation among the observed variables, all the existing methods generally require long time-series data or many samples. In real-world applications, particularly in biology, the observed time-series are typically short or small samples (e.g.,≈10 time points) due to either experiment or resource restrictions. Based on the theme of attractor embedding theory for a nonlinear dynamical system, here we develop a new prediction-based causality detection method called topologically equivalent position method which can detect the causal relation even from short observed time-series or a small number of samples. We show that this method significantly extends the classic Granger causality paradigm to a general case. Both theoretical models and gene expression data are used to verify the effectiveness of our approach.

Poster02: *Deciphering Disease Progression across Multiple Clinical Stages by Stepwise Module Network*

Tao Zeng, Luonan Chen
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Paper ID: 103

Abstract: Motivation: The presence and absence of dysfunction is not only related to single genes but also cooperative genes. The study of cooperative genes' dysfunctional presence and absence always requires the temporal data corresponding to disease progression. But, there are difficulties in clinical applications to obtain enough time-course data due to long term of complex disease and huge cost of postoperative follow-up. Meanwhile, more and more cancer stage information gives us another opportunity to model pathogen progression as a dynamical transition of a biological system among distinct patient groups. It is necessary to decipher disease progression across multiple clinical stages by modeling the presence and absence of biological dysfunctions in a systematical and united way.

Method: Our proposed stage-specific dysfunctional module were used to both represent the functional appearance when genes in a module have coherent expression behaviors under some condition/stage, and the functional disappearance when the same genes are out of co-expression under another condition/stage. (i) Multiple stage-specific dysfunctional modules are organized in the form of a network of modules, which provide a united approach to model the presence and absence of relevant functions in particular disease stage. (ii) The presence and absence of functions (dysfunctional modules) among adjacent clinical stages can further be natural biological indicators (candidate causes) for disease progression. Finally, the stepwise module network (SMN) can be reconstructed by combining the stage-specific dysfunctional modules as nodes (e.g. pathogen functions), and the causal relations among those modules as edges (e.g. functional presence and absence), which provide a systematic view of the functional cascade of a biological system during disease progression.

Results: In the present work, we implemented a computational framework to construct the stepwise module network; built a network instance based on the gene expression data of individuals in clinical stages as normal, cirrhosis, dysplastic, early HCC and advanced HCC; and analyzed the presence and absence of dysfunctional modules related to hepatitis C and cancer. Experimental findings revealed several key dynamical characteristics of biological network components as causes of HCV-HCC progression, such as the causes of virus (e.g. the genes of hepatitis C pathway), the causes of oncogene (e.g. the genes in pathway of cancer) and the causes of dysregulated genes (e.g. a few novel or less well-studied HCV-HCC associated genes).

Poster03: *Genome-wide Identification of Implicit Phenotypic Function-networks by Network Alteration Analysis*

Chuan-chao Zhang, Tao Zeng, Juan Liu, Luonan Chen
School of Computer, Wuhan University, Wuhan 430072, China

Paper ID: 105

Abstract: To study the complex diseases, a major challenge is how to identify disease-relevant networks/functions, e.g. phenotypic functions of particular disease. From the viewpoints of systems biology, it is to capture significantly differentially expressed molecules in active signaling pathways, modules, or functions in phenotype-specific molecular networks. But, many genes, gene networks, or gene modules actually have slight but important changes on their expressions or functions among different phenotypes, e.g. the phenotypic change of expression of transcriptional factor. They are known as implicit factors to determine particular phenotypes; however, there are few systematical studies on them. In this article, we purpose a computational framework as network alteration analysis (NAA) to identify the classification-defined explicit and implicit function-networks in a genome-wide way, i.e. all known biological functions in gene ontology database or the corresponding to sub-networks. By NAA, we can recognize two different patterns of biological functions and its corresponding sub-networks associated with particular disease simultaneously, which can be divided into explicit and implicit functions/sub-networks related to disease development and progression. We used NAA to investigate the diabetes preliminarily, and mainly found: (1) the quantified scores of sub-networks corresponding to explicit or implicit functions/sub-networks can be used as markers to distinguish the non-diabetic samples and diabetic samples, although the similarity distance of scores of implicit functions/sub-networks were too close; (2) disease-associated genes have different locations on the sub-networks corresponding to explicit and implicit functions, so that, they (especially implicit functions) are key phenotypic functions and would play different roles (e.g. causes of diseases) in the disease development and progression.

Poster04: *Predicting Disease Genes based on Consistently Differential Interactions for Complex Diseases*

Qianqian Shi, Xiaoping Liu, Luonan Chen
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
Paper ID: 106

Abstract: A complex disease is generally caused by genetic alterations or disorder of biological processes. Systematic identification of causal disease genes can shed light on the mechanisms underlying complex diseases, and provide new ideas to develop efficient diagnosed biomarkers or therapies. In this paper, we proposed a novel approach to predict potential disease genes for complex diseases, based on a consistency-detection scheme for molecular interactions from normal and disease samples using heterogeneous datasets, rather than single dataset. In particular, we can determine reliable differential interactions between normal and disease states by identifying the consistent interactions on a protein-protein interaction network, from which the disease genes are further decided based on those consistent interactions and also their topological structure on the network. For validating the method, the breast cancer data is used to identify the consistently differential interactions from normal to breast cancer onset, and the results well agree with the known information, thereby implying predictive power of our method. Our method also provides superior and meaningful results by compared with some typical methods. In addition, we demonstrated that the differential interactions are informative in complex disease study, in particular for detecting novel disease genes, and actually those interactions can be used as new edgetic targets from the network viewpoint.

Poster05: *Detecting Biomarkers and Disease Pathways for Nonalcoholic Fatty Liver Disease in Mouse*

Xiaoping Liu, Luonan Chen
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
Paper ID: 117

Abstract: Nonalcoholic fatty liver disease (NAFLD) is a kind of common disease in the world, and is the most common liver disease. NAFLD is also regarded as causing by fat accumulation in the liver and associating with obesity and insulin resistance. Although, there are a few researches about disease mechanism of fatty liver due to non-lethal disease, it harms human health without treatment. So it would benefit to treat the NAFLD in human if we can understand the mechanism of NAFLD onset in mouse model. In this paper, we identified some biomarkers which can be used to test the disease development of NAFLD by differential expression of genes, and some potential disease pathways to attempt to depict the onset progress of NAFLD by detecting the differential interactions in different time points. By the KEGG and GO enrichment analysis, we can see these potential disease pathways of NAFLD can significantly enriched to some liver disease progress and fatty metabolism.

Poster06: *Phytoplankton response to an intense dust storm in the Tasman Sea in September-October, 2009.*

Albert Gabric, Roger Cropp, Grant McTainsh, Barbara Johnston Johnston, Harry Butler
Griffith University, Nathan Campus, Brisbane, Queensland, Australia

Paper ID: 118

Abstract: Here we present a detailed analysis of the marine biological response in the Tasman Sea (25-40ºS, 150-170°E) after the "Red Dawn" dust storm, which was one the largest recorded in SE Australia in the last 70 years. We examine the impact of dust-derived nutrients deposited to the ocean surface on satellite-derived estimates of phytoplankton biomass as indicated by surface chlorophyll-a. We have simulated contemporaneous atmospheric dust load and deposition over the adjacent ocean using a regional dust transport model that provides daily data from September to December 2009. The phytoplankton response was confined to the region south of 30ºS, with the greatest positive anomalies (>0.6 mgm-3) occurring south of 35ºS, even though deposition was recorded further north. Contrary to previous reports of little biological impacts from dust storms in the Tasman Sea, our results suggest the regional phytoplankton can respond strongly to inputs of aeolian nutrients during the austral spring if deposition is strong and ocean conditions are favourable.

Poster07: *Detecting Causality from Nonlinear Dynamics with Short-term Time Series*

Huanfei Ma, Luonan Chen
School of Mathematics, Soochow University, Suzhou 215006, China

Paper ID: 126

Abstract: Quantifying causality between variables from observed time series data is of great importance on various fields but also a challenging task, especially when the observed data are short. Unlike the conventional methods, we find it possible to detect causality only with very short time series data, based on embedding theory of an attractor for nonlinear dynamics. Specifically, we first show that measuring the smoothness of a cross map between two observed variables can be used to detect a causal relation. Then, we provide a very effective algorithm to computationally evaluate the smoothness of the cross map, or “Cross Map Smoothness” (CMS), and thus to infer the causality, which can achieve high accuracy even with very short time series data. Analysis of both mathematical models from various benchmarks and real data from biological systems validates our method.

Poster08: *Kinase-inhibitor family map for kinase inhibitor selectivity*

Jinn-Moon Yang, Yi-Yuan Chiu, Chih-Tan Lin
National Chiao Tung University, Hsinchu 30013, Taiwan

Paper ID: 127

Abstract: Kinases play central roles in signaling pathways and are promising therapeutic targets for many diseases. Designing selective kinase inhibitors is an emergent and challenging task, because kinases share an evolutionary conserved ATP binding site. To understand kinase-inhibitor binding mechanisms, kinase inhibitor selectivity, and kinase-inhibitor-disease relationships are helpful to design selective kinase inhibitors for many diseases, such as cancers, neurological and metabolic diseases. Here, we propose kinase-inhibitor family (KIF) to address these issues. A KIF can be defined as follows: (i) the kinases in the KIF with significant sequence similarity, (ii) the inhibitors in the KIF with significant topology similarity and (iii) the kinase-inhibitor interactions (KIIs) in the KIF with significant interaction similarity. The KIIs within a KIF are often conserved on some consensus KIFMap anchors, which represent conserved interactions between the kinase subsites and consensus moieties of their inhibitors. Our experimental results reveal that the members of a KIF often possess similar inhibition profiles. The KIFMap anchors can reflect kinase conformations types, kinase functions and kinase inhibitor selectivity. Moreover, we construct KIFMap database, including 1208 KIFs, 962 KIDs, 55603 KIIs, 35788 kinase inhibitors, 399 human protein kinases, 339 diseases and 638 disease allelic variants, to explore kinase-inhibitor-disease relationships. We believe that KIFMap provides biological insights into kinase inhibitor selectivity, binding mechanisms and cancer network.

Poster09: *Molecular Features of Canine MAOA Gene: VNTR, Expression, and Methylation in Different Dog Breeds*

Jungwoo Eo, Yun-Jeong Kwon, Hee-Eun Lee, Heui-Soo Kim
Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Republic of Korea

Paper ID: 131

Abstract: Monoamine Oxidase A (MAOA), an enzyme that metabolizes serotonin and norepinephrine, has been known to be associated with some behavioral changes such as depression and antisocial/aggressive behavior. It has been reported that MAOA is expressed mostly in various parts of the dog brain. We previously analyzed canine MAOA gene such as genomic location, in silico expression, and interactions with other personality-related genes. However, genomic structures, expression patterns, and regulation of the MAOA gene in dogs are still not fully understood. To evaluate molecular features of the canine MAOA, we analysed genomic sequences including dog-specific VNTR, and the association between transcriptional levels and methylation status in promoter region of canine MAOA in brains of three dog breeds (Beagle, Sapsaree and Shepherd). We found the 7 dog breeds have the conserved dog-specific VNTR containing two blocks of 90bp and a truncated block in MAOA promoter region.

Additionally, our data showed the differential transcription levels in Beagle, Sapsaree, and Shepherd, showing higher expression levels in Beagle and Sapsaree than in Shepherd. To examine the regulatory mechanism, the methylation of the CpG island in the MAOA gene promoter was investigated by Bisulfite Seqeuencing PCR (BSP). Hypomethylation in the brain of Beagle and Sapsaree which showed high expression was observed, while hypermethylation patterns in the Shepherd brains. These results indicate that the high or low expression may be regulated by epigenetic changes in different dog breeds. Furthermore, the differential epigenetic controls and their effects on MAOA gene expression in the brain could contribute behavioral differences including aggressiveness in dog breeds.

Poster10: *HEpD: a database finding epigenetic differences between Thoroughbred and Jeju horses in four tissues*

Yong-Seok Choi, Jeong-An Gim, Sugi Lee, Dae-Soo Kim, Kwang-Seuk Jeong, Chang Pyo Hong, Jin-Han Bae, Jae-Woo Moon, Byung-Wook Cho, Hwan-Gue Cho, Jong Bhak, Heui-Soo Kim

Department of Statistics, College of Natural Sciences, Pusan National University, Busan 609-735, Republic of Korea

Paper ID: 132

Abstract: DNA methylation is very important as it regulates gene expression in organisms. With the advent of next-generation sequencing technology, genome-wide maps of DNA methylation are now available. The Thoroughbred horse is bred for racing, while the Jeju horse is a traditional Korean horse bred for racing or food. The methylation profiles of equine organs may provide genomic clues underlying their athletic traits. We have developed a publicly accessible Horse Epigenetic Database (HEpD, http://www.primate.or.kr/hepd) to elucidate genome-wide DNA methylation patterns of the cerebrum, lung, heart, and skeletal muscle from Thoroughbred and Jeju horses. Using methylated DNA immunoprecipitation coupled with next-generation sequencing, our database provides information regarding significantly enriched methylated regions beyond a threshold, methylation density of a specific region, and differentially methylated regions (DMRs) for tissues from two equine subspecies. It provided methylation patterns at 784 gene regions in the equine genome. This database can potentially help researchers identify DMRs in the tissues of these horse species and investigate the differences between the Thoroughbred and Jeju horse subspecies.

Poster11: *Clustering Analysis and Association Analysis of Subject in Literature relating to Hepatitis B*

Junli Liu, Xiumei Zhang

Healthcare Department, Beijing Wanfang Data Co., Ltd., Beijing 100038, China

Paper ID: 134

Abstract: The present study collected the studies related to hepatitis B published between 2010 and 2012 in the core Chinese medical journals, including Wanfang Med Online. Then, the present study conducted pre-treatment for the abstracts using a stop list to rule out obvious interference characters which cannot be components of medical terms. The repeated character strings in the texts were then extracted using a method based on the string frequency statistics and taken as a pre-filtering candidate terminology database. Finally, the candidate terminology database was filtered based on three aspects, including inverse document frequency, common substring, and term length by setting a C-value for the threshold. Finally, Basic data were obtained from the identified papers, and investigated the distribution of research subjects in the literature related to hepatitis B using co-word and cluster analyses. Moreover, our study determined the correlation between three subjects (Chinese medicine, urinary system disease, and metabolic syndromes) using pair wise association analysis. The combination of clustering and association analyses was superior at identifying the relationship between different studies in the literature.

Acknowledgement: This work was supported by National Natural Science Foundation of China under Grant No. 71133006.

Poster12: *Development of a Clinical Guideline-based Smartphone Application for Obesity Management*

Eunjoo Jeon, Hyeoun-Ae Park

College of Nursing, Seoul National University, Seoul, Korea

Paper ID: 135

Abstract: Objectives: The purpose of the study was to develop a smartphone application (“app”) based on clinical guideline and entity-attribute-value (“EAV”) model for obesity management. Methods: Obesity-related knowledge and functional requirements were extracted from clinical practice guidelines, a literature review, and consultations with experts. Extracted knowledge was used to design obesity management algorithms, and functions of the app are presented in a use case diagram and activity diagrams. Data models were developed and database was created using EAV model. User interface was designed based on user requirements. Finally an app was developed. The proficiency and efficiency of the algorithm were evaluated using scenarios, while user interface was assessed using a mobile heuristic evaluation tool and the usability of the app was assessed using the Post-Study System Usability Questionnaire. Results: In total, 131 obesity management-related rules and 11 functions for the app were extracted. Five algorithms—comprising 1 main algorithm and 4 subalgorithms—were developed. Total 806 data models were developed and categorised into 5 groups such as ‘Input’, ‘Diary’, ‘Output’, ‘Food’, and ‘Exercise’. The database comprised of 11 tables with 4 view tables, and 41 screens. An app was developed using Android SDK platform 4.0.3, JDK 1.7.0. The overall proficiency and efficiency scores of the algorithm were 88.0 and 69.1, respectively. In heuristic tests, 57 comments were made and mean usability score was 3.47 out of 5. Thirteen usability problems were identified by the heuristics and usability evaluations. Conclusion: The app developed in this study might be helpful for obesity management, since it can provide high-quality information and interventions without spatial or temporal constraints. However, the clinical effectiveness of this app requires further investigations. Keywords: Weight Loss, Telemedicine, Mobile Health Units. Acknowledgement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2010-0028631).

Poster13: *Development of Automatic Nursing Narratives Generation System from Entity-attribute-value Triplets*

Yul Ha Min, Eunjoo Jeon, Hyeoun-Ae Park
Research Institute of Nursing Science, Seoul National University, Seoul, Korea
Paper ID: 136

Abstract: Introduction: The aim of this study is to develop and evaluate a natural language generation (“NLG”) system to populate nursing narratives using detailed clinical models. The NLG of nursing narratives consists of three phases: identifying type of statement and plausible set of attributes to determine content of the narratives; determining the sequence of the terms and case markers to refine the narratives; and generating grammatically correct narratives with different levels of granularity by assigning values to attributes and removing the ambiguity of the narratives. Methods: Semantic, contextual and syntactic knowledge were extracted. Detailed clinical models (“DCM”) as semantic knowledge developed for signs and symptoms, nursing problems, and patient’s responses to nursing care in a previous study were used. Contextual knowledge such as the source of information and tense of verb was collected by consulting to the experts. Syntactical knowledge was extracted by analysing Korean grammar. Then NLG system was developed and implemented for pilot testing. We developed and implemented the NLG system for our task. Results: With 82 detailed clinical models, total 66,888 nursing narratives in four different types of statement were generated. The quality of generated nursing narratives was evaluated by the three nurse experts using a five-point rating scale. The mean scores for overall quality was 4.66, for content 4.60, for grammaticality 4.40, for writing style 4.13, and for correctness 4.60. Discussion: The system developed in this study was able to generate nursing narratives with different levels of granularity successfully. The generated nursing narratives can improve semantic interoperability of nursing data documented in nursing progress notes. Acknowledgement: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2012-012257).

Poster14: *Development of an Obesity Ontology for Collection and Analysis of Big Data*

Ae Ran Kim, Tae-Min Song, Hyeoun-Ae Park
College of Nursing, Seoul National University, Seoul, Korea
Paper ID: 141

Abstract: Background: We have recently seen a big surge in the level of interest in big data as tools to store and analyse large amount of unstructured data on social media networks become available. Obesity is one of major health concerns for people because it is a known risk factor for numerous chronic diseases. It is important to understand how social media interactions reflect and shape the public discourse on obesity. Purpose: The purpose of this study was to develop an obesity ontology with a list of terms for acquisition and analysis of unstructured social networking data on obesity. Methods: An obesity ontology was developed based on ‘Ontology development 101’: 1) determining the domain and scope of the ontology, 2) enumerating important terms in the ontology from obesity management guidelines of NICE(2006) and Korean Society for the Study of Obesity(2012), 3) extracting additional terms by analysing postings on social media, 4) defining the classes and the class hierarchy by a top-down development process 5) defining the properties of classes and the data types of properties, 6) describing the relationship between classes, and 7) developing terminology by presenting synonyms of each term. Results: Eight domains of obesity ontology were extracted from the guidelines. They are ‘types’, ‘risk factors’, ‘signs and symptoms’, ‘complications’, ‘diagnosis’, ‘treatment’, ‘prevention’ and ‘settings’. In total 598 terms were extracted from the guidelines and additional 288 terms from postings on social networking services. In total 55 classes and 101 properties were derived from the enumerated terms. Subclasses from the extracted terms of social media were ‘fad diet’, ‘oriental medicine’, and ‘other procedure or treatment’. Terminology was developed

with illustrating 740 terms and 606 synonyms. Conclusion: Obesity ontology and terminology developed in this study can be used as a framework to understand obesity using unstructured big data from social media. Key words: obesity, social media, ontology, terminology. Acknowledgements: This work was supported by the research grant (Efficient management of Big Data on Health & Welfare) funded by the Korea Institute for Health and Social Affairs (No.20140225). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2010-0028631).

Poster15: *Outcome-guided mutual information networks for investigating gene-gene interaction effects on clinical outcomes*

Hyun-hwan Jeong, So Yeon Kim, Kyubum Wee, Kyung-Ah Sohn
Department of Information and Computer Engineering, Ajou University, Suwon 443-749, S. Korea
Paper ID: 142

Abstract: Network-based analysis frameworks have gained huge popularity recently as network information can provide a more systematic and global view of the underlying biological system. For example, network-based linear regression for eQTL analysis or network-based Cox regression for survival analysis has successfully improved their performances by incorporating biological network structure in regression frameworks. However, most network-based studies rely on feature-wise networks such as a correlation network or a static network obtained from an existing knowledgebase. Such networks can reveal the relation between a pair of features, but do not consider the effect of pair-wise feature interactions on the outcome. To detect significant feature pairs associated with the outcome, we employ the Mutual Information measure, which is a non-parametric, information-theoretic measure and has been successfully used to detect linear or non-linear association between the features. Based on the extension of Mutual Information measure, we propose a simple but powerful scheme to construct an outcome-guided network with appropriate edge significance filtering. We demonstrate the utility of the proposed network construction approach in two main applications: the integrative network analysis of multiple genomic profiles, and the network-based survival analysis. In both applications, datasets from ovarian serous cystadenocarcinoma patients in The Cancer Genome Atlas (TCGA) are used. The results highlight the usefulness of the outcome-guided mutual information networks in both applications for investigating gene-gene interaction effects associated with clinical outcomes.

Poster16: *Predictive Model of the Survival Rates of the Major Cancers Using Semi-Supervised Learning*

Jung Min Yun, Peter Kang, Kang Hee Park, Sung-Hye Park, Jung Wook Seo, Peom Park
Department of Industrial Engineering, Ajou University, Suwon 443749, Korea
Paper ID: 145

Abstract: Objective: This study was to propose a model to predict the survival rates and times of the major cancers by analysis of clinical pathology data using semi-supervised learning techniques. Background: A study on the prediction of cancer survival rates and times based on the patient physical and diagnosis information is not enough in Korea even though the National Cancer Information Center provides various information including the major cancer incidence rates, five year cancer survival rates and cancer death rates. Methods: In this study, the major cancer survival rates by various time periods were predicted by an application of semi-supervised learning techniques using physical and medical information of 500 cancer patients. Results: Predictive algorithms in the cancer survival rates of stomach cancer, colon cancer, lung cancer and liver cancer by month and year were proposed, and the reliability of algorithms was verified using previous data from 1993 to 2011. Comparing to the five year cancer survival rates reported by the National Cancer Information Center, similar cancer survival rates were observed (Stomach cancer: 69%, colon cancer: 74%, lung cancer: 20%, and liver cancer: 28.6%) in this study. Conclusion: This study proposed the algorithms of the cancer survival rates by the length of major cancers, and predictive results showed that the survival rates of the major cancers increase as time passes. However, extra factors such as the development of medical technology and various cultural issues were not considered because that the proposed algorithms merely depended on statistic values. Therefore, it is suggested that the proposed algorithms should be verified and compared by considering various factors which may influence the cancer survival rates, and using other data mining methods.

Poster17: *NCC-AUC: an AUC-based method to identify multi-biomarker panel for prognosis and survival analysis*

Meng Zou, Zhaoqi Liu, Yong Wang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
Paper ID: 146

Abstract: In prognosis and survival studies, an important and challenging goal is to identify multiple-biomarker panels with predictive power. The challenge lies in dealing with the noisy, censored, and small-sample-size survival data and dig out crucial biomarkers from high-dimensional clinical or genomic data. Thus sophisticated models and algorithms are in pressing need. In this study, we propose a novel Area Under Curve (AUC)-based multi-biomarker panel identification method called NCC-AUC (Nearest Centroid Classifier for AUC evaluation) and apply it to lung cancer prognosis and survival analysis. Our method is inspired by the connection between AUC in classification accuracy evaluation and Harrell's concordance index in survival analysis. This connection allows us to convert the survival time regression problem to a binary classification problem. Then an optimization model is formulated to directly maximize AUC and meanwhile select a small group of biomarkers to construct a predictor. Specifically, NCC-AUC utilizes the nearest centroid classifier framework and optimizes the classification accuracy AUC. Simultaneously, NCC-AUC minimizes the number of features to regularize the classifier. We apply NCC-AUC to a lung cancer survival dataset and identify multi-biomarker panels to predict disease prognosis. A panel of 12 biomarkers was revealed for stage IB non-small-cell lung cancer with AUC 0.8284 in training cohort, and AUC 0.7589 in the independent validation cohort. The panel showed better performance by combining several biomarkers together to significantly improve univariate analysis with maximum AUC 0.6401 for single biomarker p21ras. In addition, NCC-AUC outperformed widely used feature selection method SVM-RFE, which obtained training AUC 0.8204 and validation AUC 0.5036. Furthermore, we compared NCC-AUC with classical survival analysis model such as Cox model or L1-Cox model. NCC-AUC obtained a p-value 0.0158 by log rank test while Cox model and L1-Cox model only got 0.5939 and 0.7166 in the validation cohort. In summary, NCC-AUC can serve as a useful tool to identify meaningful genomic or clinical biomarkers for prognosis and survival prediction.

Poster18: *Lung Caner Related EdgeMarkers*

Wanwei Zhang, Luonan Chen
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Abstract: Lung cancer is notorious as one of the leading cause of cancer-related deaths worldwide. Although it's widely studied, the early diagnosis and treatment of lung cancer are still big challenges, especially for non-small-cell lung cancer, of which the mechanism is still not clear. To get a better understanding about how this cancer develops, we applied our previously developed method, EdgeMarker, to the RNA-sequencing datasets of two major non-small-cell lung cancer types, adenocarcinoma and squamous-cell carcinoma, from TCGA Database, and identified edgetic biomarkers (markers that are the interactions or correlations of the molecular pairs) specific to both cancer types. Unlike the biomarkers that are discovered by traditional methods, the expression of the molecules that are involved in the edge biomarkers are not necessarily to be significantly different. Moreover, edge biomarkers directly use the correlation information of molecular pairs to characterize the biological state in contrast to traditional biomarkers which use the molecular expressions. The function analysis of these edge biomarkers implies the relevance between them and the diseases. These may give us a new inspiration to treat the disease.

Poster19: *Infer metagenomic abundance and reveal homologous genomes based on the structure of the taxonomy tree*

Yu-Qing Qiu, Xue Tian and Shihua Zhang
Chinese Academy of Sciences
Paper ID: 124

Abstract: **Background:** Metagenomic research uses sequencing technologies to investigate the genetic biodiversity of microbiomes present in various ecosystems or animal tissues. The composition of a microbial community is highly associated to the environment in which the organisms exist. As large amount of sequencing short reads of microorganism genomes obtained, accurately estimating the abundance of microorganisms within a metagenomic sample is becoming an increasing challenge in bioinformatics. **Results:** In this paper, we describe a hierarchical taxonomy tree based mixture model (HTTMM) for estimating the abundance of taxon within a microbial community by incorporating the structure of the taxonomy tree. In this model, genome specific short reads and homologous short reads among genomes can be distinguished and represented by leaf and intermediate nodes in the taxonomy tree respectively. An expectation-maximization algorithm has been adopted to solve this model. **Conclusions:** Using simulated and real-world applications, we demonstrate that the proposed method is superior to both flat mixture model and lowest common ancestry based methods. Moreover, this model revealed previously unaddressed homologous genomes.

Poster20: *SeedsGraph: an efficient assembler for next generation sequencing data*

Chunyu Wang, Maozu Guo, Xiaoyan Liu, Yang Liu and Quan Zou
Harbin Institute of Technology

Paper ID: 22

Abstract: DNA sequencing technologies has been rapidly evolving, and produces large number of short-reads in a fast rising tendency. This led to resurgence of research in whole-genome shotgun assembly algorithms. We start the assembly algorithm by clustering the short reads in cloud computing framework, and the clustering process groups fragments according their original consensus long sequence similarity. We condense each group of reads to a chain of seeds, which is a kind of substring with reads aligned, and then build a graph accordingly. Finally we analysis the graph to find Euler paths, and assembly the reads related in the paths for contigs, then layout contigs with mate-pair information for scaffolds. The result shows our algorithm is efficient and feasible for large set of reads like in Next Generation Sequencing technology.

Open Access
Open Access
GPB
Genomics, Proteomics & Bioinformatics
Genomics in Neurological Disorders