ICIBM

International Conference on Intelligent Biology and Medicine

2018

June 10-12, 2018
Los Angeles, CA, USA

Co-hosted by
The International Association for Intelligent Biology and Medicine (IAIBM) and Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston

2018 International Conference on Intelligent Biology and Medicine (ICIBM 2018)

June 10-12, 2018

Los Angeles, CA, USA

Hosted by:

The International Association for Intelligent Biology and Medicine (IAIBM)

and

Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston

TABLE OF CONTENTS

Welcome to ICIBM 2018!

On behalf of all our conference committees and organizers, we welcome you to the 2018 International Conference on Intelligent Biology and Medicine (ICIBM 2018), co-hosted by The International Association for Intelligent Biology and Medicine (IAIBM) and the Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston (UTHealth). Given the rapid innovations in the fields of bioinformatics, systems biology, and intelligent computing and their importance to scientific research and medical advancements, we are pleased to once again provide a forum that fosters interdisciplinary discussions, educational opportunities, and collaborative efforts among these ever growing and progressing fields.

We are proud to have built on the successes of previous years' conferences to take ICIBM 2018 to the next level. In this year, our keynote speakers include Drs. Josh Denny, Alexander Hoffmann, Jason Moore, and Paul Thomas. These individuals are world-renowned experts in their respective fields, and are privileged to host their talks at ICIBM 2018. Throughout the conference, we will also feature eminent scholar talks and speakers in four workshops and tutorials that will each provide an in-depth presentation or lesson on some of the most popular informatics topics in the biological and biomedical areas. We also have faculty members, postdoctoral fellows, PhD students and trainee level awardees selected from a substantial number of outstanding manuscripts and abstracts that span a diverse array of research subjects. These researchers, chosen through a rigorous review process, will showcase the innovative technologies and approaches that are the hallmark of our featured interdisciplinary fields and their related applications.

Overall, we anticipate this year's program will be incredibly valuable to research, education, and innovation, and we hope you are as excited as we are to experience ICIBM 2018's program. We'd like to extend our thanks to our sponsors for making this event possible, including National Science Foundation, The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Data Science and Informatics Core for Cancer Research, Illumina, Nextomics, Novogene, Macrogen, BGI Americas, and IBM. Furthermore, our sincerest thanks to the members of all our committees and our volunteers for their valuable efforts; we could not have accomplished so much without your dedication to making ICIBM 2018 a success.

On behalf of all of us, we hope that our hard work has provided a conference that is thought provoking, fosters collaboration and innovation, and is enjoyable for all of our attendees. Thank you for attending ICIBM 2018. We look forward to your participation in all our conference has to offer!

Sincerely,

Zhongming Zhao, PhD
ICIBM General Chair
Professor and Director,
Center for Precision Health
School of Biomedical Informatics
UTHealth, Houston

Kai Wang, PhD
ICIBM Program co-Chair
Associate Professor,
Raymond G. Perelman Center for Cellular and Molecular Therapeutics & Department of Pathology,
Children's Hospital of Philadelphia

Degui Zhi, PhD
ICIBM Program co-Chair
Associate Professor,
Center for Precision Health
School of Biomedical Informatics
UTHealth, Houston

ACKNOWLEDGEMENTS

General Chairs

Zhongming Zhao, The University of Texas Health Science Center at Houston, USA

Steering Committee

Yidong Chen, The University of Texas Health Science Center at San Antonio, USA
Kun Huang, Indiana University, USA
Tony Hu, Drexel University, USA, USA
Lang Li, Ohio State University, USA
Yi Xing, University of California Los Angeles, USA
Jiajie Zhang, The University of Texas Health Science Center at Houston
W. Jim Zheng, The University of Texas Health Science Center at Houston

Program Committee

Kai Wang, Co-Chair, Children's Hospital of Philadelphia, USA
Degui Zhi, Co-Chair, The University of Texas Health Science Center at Houston, USA
Genevera Allen, Rice University, USA
Jianlin Cheng, University of Missouri Columbia, USA
Luca Giancardo, The University of Texas Health Science Center at Houston, USA
Assaf Gottlieb, The University of Texas School of Biomedical Informatics at Houston, USA
Mike Guo, Bioinformatics Shared Resource, UNMCCC, University of New Mexico, USA
Zhi Han, Indiana University School of Medicine, USA
Matthew Hayes, Xavier University of Louisiana, USA
Kun Huang, Indiana University, USA
Weichun Huang, NIH/NIEHS, USA
Yang Huang, Kaiser Permanente, USA
Yufei Huang, University of Texas at San Antonio, USA
Zhi-Linag Ji, Xiamen University, China
Peilin Jia, University of Texas Health Sciences Center at Houston, USA
Victor Jin, University of Texas at San Antonio, USA
Yufang Jin, University of Texas at San Antonio, USA
Jun Kong, Emory University, USA
Dmitry Korkin, Worcester Polytechnic Institute, USA
Rui Kuang, University of Minnesota Twin Cities, USA
Kb Kulasekera, University of Louisville, USA
Qiwei Li, Rice University, USA
Fuhai Li, The Methodist Hospital, Weill Medical College of Cornell University, USA

Tao Li, Nankai University, China
Zhaohui Li, Nankai University, China
Aimin Li, Xi'an University of Technology, China
Ping Liang, Brock University, Canada
Li Liao, University of Delaware, USA
Honghuang Lin, Department of Medicine, Boston University School of Medicine, USA
Nan Liu, Chinese Academy of Sciences, China
Yin Liu, University of Texas Health Science Center at Houston, USA
Zhandong Liu, Baylor College of Medicine, USA
Xiaoming Liu, The University of Texas Health Sciences Center at Houston, USA
Zhiyong Lu, National Center for Biotechnology Information, USA
Mirjana Maletic, Baylor College of Medicine, USA
Patricio Manque, Universidad Mayor, Chile
Huaiyu Mi, University of Southern California, USA
Nitish Mishra, University of Nebraska Medical Center, USA
Qiangxing Mo, Baylor College of Medicine, USA
Tabrez Anwar Shamim Mohammad, Greehey Children's Cancer Research Institute (GCCRI), USA
Hatice Ozer, Ohio State University, USA
Ranadip Pal, Texas Tech University, USA
Jiang Qian, Johns Hopkins School of Medicine, USA
Guimin Qin, Xidian University, China
Thomas Rindflesch, National Library of Medicine, USA
Jianhua Ruan, The University of Texas at San Antonio, USA
Jianhua Ruan, The University of Texas at San Antonio, USA
Bairong Shen, Soochow Univerity, China
Xiaofeng Song, Nanjing University of Aeronautics and Astronautics, China
Fengzhu Sun, University of Southern California, USA
Wing-Kin Sung, National University of Singapore,
Manabu Torii, Kaiser Permanente, USA
Ying-Wooi Wan, Baylor College of Medicine, USA
Jun Wan, Indiana University, USA
Yufeng Wang, University of Texas at San Antonio, USA
Junbai Wang, Radium Hospital, USA
Qingguo Wang, Memorial Sloan Kettering Cancer Center, USA
Jiaying Wang, Xian Jiaotong University, China
Chaochun Wei, Shanghai Jiao Tong University, China
Xiwei Wu, City of Hope, USA
Yonghui Wu, University of Texas Health Science Center at Houston, USA
Zhijin Wu, Brown University, USA
Jungfeng Xia, Institute of Health Sciences, Anhui University, USA
Lu Xie, Shanghai Center for Bioinformation Technology, USA
Lei Xie, City University of New York, USA

Hua Xu, The University of Texas School of Biomedical Informatics at Houston, USA
Jianhua Xuan, Virginia Tech, USA
Yu Xue, Huazhong University of Science and Technology, China
Zhenqing Ye, University of Texas Health Science Center at San Antonio, USA
Sungroh Yoon, Seoul National University, Korea
Feng Yue, Penn State University, USA
Habil Zare, Texas State University, USA
Rui Zhang, University of Minnesota, USA
Shaojie Zhang, University of Central Florida, USA
Han Zhang, Nankai University, China
Bing Zhang, Baylor College of Medicine, USA
Zhongming Zhao, The University of Texas Health Science Center at Houston, USA
Min Zhao, Vanderbilt University, USA
Xianghong Zhou, University of California Los Angeles, USA
Yunyun Zhou, University of Mississippi, USA

Publication Committee
Zhijin Wu, Co-Chair, Brown University, USA
Jianhua Ruan, Co-Chair, The University of Texas at San Antonio, USA

Workshop/Tutorial Committee
Feng Yue, Co-Chair, Pennsylvania State University, USA
Yan Guo, Co-Chair, University of New Mexico, USA

Publicity Committee
Lana Garmire, Co-Chair, University of Hawaii, USA
Sun Kim, Co-Chair, Seoul National University, Korea
Yu Xue, Co-Chair, Huazhong University of Science and Technology, China

Award Committee
Fuhai Li, Co-Chair, Ohio State University, USA
Lei Xie, Co-Chair, City University of New York, USA

Trainee Committee
Abolfazl Doostparast, Co-Chair, Columbia University, USA
Qian Liu, Co-Chair, Children's Hospital of Philadelphia, USA

Local Organization Committee
Jessica Li, Co-Chair, University of California Los Angeles, USA
Matteo Pellegrini, Co-Chair, University of California Los Angeles, USA
Xiaoming Liu, Co-Chair, University of Texas Health Science Center at Houston, USA

International Conference on Intelligent Biology and Medicine
Program at-a-glance (June 10-12, 2018)

Sunday, June 10th

12:00	**Registration** opens		
CONCURRENT WORKSHOPS			
Room: Santa Monica		**Room: Grand Imperial North**	
1:00 - 2:30	Dr. Yan Guo University of New Mexico **Machine Learning**	1:00 - 2:30	Dr. Feng Yue Pennsylvania State University **3D Genome Organization**
2:30 - 2:40	***Break***		
2:40 - 4:10	Chi Zhang; Xiao Dong Indiana University; Albert Einstein College of Medicine **Single-cell Sequencing Analysis**	2:40 - 4:10	Dr. Ting Wang Washington University **WashU Epigenome Browser**
4:10 - 4:30	***Coffee/Tea Break***		
4:30 - 5:20	**Keynote Lecture (Room: Grand Imperial North)** **Josh Denny, MD, MS** **Professor of Biomedical Informatics and Medicine** **Vanderbilt University** **Member, National Academy of Medicine** Title: Huge cohorts, genomics, and clinical data to personalize medicine		
5:20 - 5:30	***Break***		
5:30 - 7:30	**Poster Session (Room: Grand Imperial Foyer)**		

Monday, June 11st

7:30 - 8:30	**Registration Open and Buffet Breakfast**
8:30 - 8:40	**Opening Remarks**
8:40 - 9:30	**Keynote Lecture (Room: Grand Imperial North)** **Paul Thomas, Ph.D.** **Associate Professor of Preventive Medicine** **Director of Division of Bioinformatics** **University of Southern California** **PI, Gene Ontology Consortium** Title: Reconstructing the large-scale evolution of genomes and gene functions
9:30 - 9:40	***Break***
9:40 - 10:00	**Eminent Scholar Talk (Room: Grand Imperial North)** **Xinghua Lu, MD, PhD** **Professor of Biomedical Informatics** **University of Pittsburgh** Title: From Big Data to Bedside (BD2B): Precision Oncology in an Era of Artificial Intelligence
10:00-10:05	***Break for parallel session***
CONCURRENT SESSIONS	

Room: Grand Imperial North **NGS & Tools** Session Chair: Youping Deng		Room: Santa Monica **Systems Biology** Session Chair: Yin Liu		Room: Hollywood **Bioinformatics** Session Chair: Li Liu	
10:05 -10:25	*Development of somatic mutation signatures for risk stratification and*	10:05 -10:25	*A Hidden Markov Model-based approach to reconstructing double*	10:05 - 10:25	*A Graph-based Algorithm for Estimating Clonal Haplotypes of Tumor*

	prognosis in lung and colorectal adenocarcinomas Mark Menor, Yong Zhu, Bin Jiang and Youping Deng		*minute chromosome amplicons* Ruslan Mardugalliamov, Kamal Al Nasr and Matthew Hayes		*Sample from Sequencing Data* Yixuan Wang, Rong Zhang, Xinyu Sun, Yu Geng, Jianye Liu, Zhongmeng Zhao, Xuanping Zhang, Yi Huang and Jiayin Wang
10:25 -10:45	*Genomic copy number variations in large screening for pediatrics sarcomas chemotherapy* Lijun Cheng, Pooja Chandra, Limei Wang, Karen Pollok, Pankita Pandya, Mary Murray, Jacquelyn Carter, Michael Ferguson, Mohammad Reza Saadatzadeh, Mashall Mark, Li Lang and Jamie Renbarger	10:25 -10:45	*Classifying Mild Traumatic Brain Injuries with Functional Network Analysis* Francis San Lucas, John Redell, Dash Pramod and Yin Liu	10:25 - 10:45	*A new insight into underlying disease mechanism through semi-parametric latent differential network model* Yong He, Jiadong Ji, Lei Xie, Xinsheng Zhang and Fuzhong Xue
10:45 -11:05	*Computational identification of deleterious synonymous variants in human genomes using a feature-based approach* Fang Shi, Yao Yao and Junfeng Xia	10:45 -11:05	*scDNA: a fast and comprehensive tool for single-cell differential network analysis* Yu-Chiao Chiu, Tzu-Hung Hsiao, Li-Ju Wang, Yidong Chen and Yu-hsuan Shao	10:45 - 11:05	*Genetic-Epigenetic Interactions in Asthma Revealed by a Genome-Wide Gene-Centric Search* Vladimir Kogan, Joshua Millstein, Stephanie J London, Carole Ober, Steven R White, Edward T Naureckas, W James Gauderman, Daniel J Jackson, Albino Barraza-Villarreal, Isabelle Romieu, Benjamin A Raby and

					Carrie V Breton
11:05 -11:15	***Coffee/Tea Break***				
11:15 -11:35	*High scoring segment selection for pairwise whole genome sequence alignment with the maximum scoring subsequence and GPUs* Abdulrhman Aljouie, Ling Zhong and Usman Roshan	11:15 -11:35	*Multiple transcription factors contribute to inter-chromosomal interaction in yeast* Yulin Dai, Chao Li, Guangsheng Pei, Dong Xiao, Guohui Ding, Zhongming Zhao, Yixue Li and Jia Peilin	11:15 - 11:35	*DLAD4U: deriving and prioritizing disease lists from PubMed literature* Junhui Shen, Suhas Vasaikar and Bing Zhang
11:35 -11:55	*Detecting virus integration sites based on multiple related sequencing data by VirTect* Yuchao Xia, Yun Liu, Minghua Deng and Ruibin Xi	11:35 -11:55	*Metabolomics of Mammalian Brain Reveals Regional Differences* William Choi, Mehmet Tosun, Cemal Karakas, Fatih Semerci, Zhandong Liu and Mirjana Maletic-Savatic	11:35 - 11:55	*A multitask bi-directional RNN model for named entity recognition on electronic medical records* Shanta Chowdhury, Xishuang Dong, Lijun Qian, Xiangfang Li, Yi Guan, Jinfeng Yang and Qiubin Yu
11:55 -12:15	*Comprehensive assessment of genotype imputation performance* Shuo Shi, Na Yuan, Ming Yang, Zhenglin Du, Jinyue Wang, Sheng Xin, Jiayan Wu and Jingfa Xiao	11:55 -12:15	*Boosting Gene Expression Clustering with System-Wide Biological Information: A Robust Autoencoder Approach* Hongzhu Cui, Chong Zhou, Xinyu Dai, Yuting Liang, Randy Paffenroth and	11:55 - 12:15	*iMEGES: integrated Mental-disorder GEnome score for prioritizing the susceptibility genes for mental disorders in personal genomes* Atlas Khan, Qian Liu and Kai Wang

			Dmitry Korkin		
12:15 - 1:35	**Lunch Break -** Buffet Lunches				
1:35 - 1:55	**Eminent Scholar Talk (Room: Grand Imperial North)** **Ting Wang, PhD** **Associate Professor of Genetics** **Washington University** Title: Exploring the dark matter in genomics data				
1:55 - 2:00	***Short Break***				
2:00 - 2:50	**Keynote Lecture (Room: Grand Imperial North)** **Alex Hoffmann, Ph.D.** **Thomas M. Asher Professor of Microbiology** **Director of the Institute for Quantitative and Computational Biosciences** **University of California Los Angeles** Title: Learning how to predict immune responses				
2:50 - 3:00	***Break for parallel sessions***				
CONCURRENT SESSIONS					
Room: Grand Imperial North **NGS & Tools** Session Chair: Lei Xie		Room: Santa Monica **Systems Biology** Session Chairs: Matthew Hayes		Room:Hollywood **Medical Informatics** Session Chairs: W Jim Zheng	
3:00 - 3:20	*Reconstructing the High-resolution Chromosomal 3D structure by Hi-C Complex Network* Tong Liu and Zheng Wang	3:00 - 3:20	*Prediction of Protein Self-Interactions using Stacked Long Short-Term Memory from Protein Sequences Information* Yanbin Wang, Zhuhong You, Xiao	3:00 - 3:20	*Building a high performance computing infrastructure for cancer research* W Jim Zheng

			Li, Tonghai Jiang, Li Cheng and Zhanheng Chen		
3:20 - 3:40	*CeL-ID: Cell Line Identification using RNA-seq data* Tabrez A Mohammad, Yun S Tsai, Safwa Ameer, Hung-I H Chen, Yu-Chiao Chiu and Yidong Chen	3:20 - 3:40	*Circular RNA Expression Profiles during the Differentiation of Mouse Neural Stem Cells* Qichang Yang, Jing Wu, Jian Zhao, Tianyi Xu, Zhongming Zhao, Xiaofeng Song and Ping Han	3:20 - 3:40	*Dynamic Prediction of Hospital Admission with Medical Claim Data* Tianzhong Yang, Yang Yang, Yugang Jia and Xiao Li
3:40 - 4:00	*Comparison of SureSelect and Nextera exome capture performance in single-cell sequencing* Wendy Huss, Qiang Hu, Sean Glenn, Kalyan Gangavarapu, Jianmin Wang, Jesse Luce, Paul Quinn, Elizabeth Brese, Fenglin Zhan, Jeffrey Conroy, Gyorgy Paragh, Barbara Foster, Carl Morrison, Song Liu and Lei Wei	3:40 - 4:00	*Identification of Gene Signatures from RNA-seq Data Using Pareto-optimal Cluster Algorithm* Saurav Mallik and Zhongming Zhao	3:40 - 4:00	*Early prediction of acute kidney injury following ICU admission* Lindsay P. Zimmerman, Paul A. Reyfman, Angela D. R. Smith, Zexian Zeng, Abel Kho, L. Nelson SanchezPinto and Yuan Luo
4:00 - 4:10	***Coffee/Tea Break***				

4:10 - 4:30	*A PheWAS Study of a Large Observational Epidemiological Cohort of African Americans: the REGARDS Study* Xueyan Zhao, Xin Geng, Vinodh Srinivasasainagendra, Suzanne Judd, Virginia Wadley, Orlando Gutierrez, Henry Wang, Ethan Lange, Leslie Lange, Daniel Woo, Fred Unverzagt, Monika Safford, Mary Cushman, Nita Limdi, Rakale Quarells, Donna Arnett, Marguerite Irvin and Degui Zhi	4:10 - 4:30	*An experimental design framework for Markovian gene regulatory networks under stationary control policy* Roozbeh Dehghannasiri, Mohammad Shahrokh Esfahani and Edward Dougherty	4:10 - 4:30	*Epileptic foci localization based on mapping the synchronization of dynamic brain network* Mei Tian, Wei Xiaoyan, Chen Ziyi, Tian Xianghua, Dong Nan, Li Dongmei and Zhou Yi
4:30 - 4:50	*NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data* Qian Liu, Daniela C. Georgieva, Dietrich M. Egli and Kai Wang	4:30 - 4:50	*GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization* Hung-I Chen, Yu-Chiao Chiu, Tinghe Zhang, Songyao Zhang, Yufei Huang and Yidong Chen	4:30 - 4:50	*Gene Fingerprint model for Literature based detection of the associations among complex diseases: A case study of COPD* Guocai Chen, Yuxi Jia, Lisha Zhu, Ping Li, Lin Zhang, Cui Tao and W. Jim Zheng
4:50 – 5:00	*Joint Principal Trend Analysis for Longitudinal High-Dimensional Genomic Data*	4:50 – 5:00	*Comparative gene co-expression network analysis of epithelial to mesenchymal*	4:50 – 5:10	*Integrating Sentence Sequence Representation and Shortest Dependency Path into a Deep*

	Yuping Zhang and Zhengqing Ouyang (**selected abstract talk**)		*transition reveals lung cancer progression stages* Daifeng Wang, John Haley and Patricia Thompson (**selected abstract talk**)		*Learning Framework for Relation Extraction in Clinical Text* Zhiheng Li, Zhihao Yang, Chen Shen, Jun Xu, Yaoyun Zhang and Hua Xu
5:00 – 5:10	*Inferring Drug-Target Associations based on perturbational profiles in L1000 Data* Pei-Han Liao, Tzu-Hung Hsiao, Liang-Chuan Lai, Mong-Hsun Tsai, Tzu-Pin Lu and Eric Y. Chuang (**selected abstract talk**)	5:00 – 5:10	*Epigenomic Patterns Are Associated with Gene Haploinsufficiency and Predict Risk Genes of Developmental Disorders* Siying Chen, Xinwei Han and Yufeng Shen (**selected abstract talk**)		
5:10 – 5:20	Predict effective drug combination by deep believe network and Ontology Fingerprints Guocai Chen, Lam Tsoi, Hua Xu and W. Jim Zheng (**selected abstract talk**)	5:10 – 5:20	*Genetic Association of Arterial Stiffness Index with Incident Coronary Artery Disease and Congestive Heart Failure* Seyedeh Zekavat, Mary Haas, Krishna Aragam, Connor Emdin, Amit Khera, Derek Klarin, Hongyu Zhao and Pradeep Natarajan (**selected abstract talk**)	5:10 – 5:30	

Tuesday, June 12nd

7:30 - 8:40	**Registration Open and Buffet Breakfast**
8:40 - 9:30	**Keynote Lecture (Room: Grand Imperial North)** **Jason Moore, Ph.D.** **Edward Rose, M.D. And Elizabeth Kirk Rose, M.D. Professor of Biostatistics, Epidemiology and Informatics** **Director, Institute for Biomedical Informatics** **University of Pennsylvania** Title: Accessible artificial intelligence for data science
9:30 - 9:40	***Break***
9:40 - 10:00	**Eminent Scholar Talk (Room: Grand Imperial North)** **Grace Xiao, Ph.D.** **Professor of Integrative Biology and Physiology** **UCLA** Title: Deciphering the function of single-nucleotide variants in the RNA
10:00 - 10:05	***Break for parallel sessions***

CONCURRENT SESSIONS

Room: Grand Imperial North **Cancer Genomics** Session Chair: Bin Chen		Room: Santa Monica **Systems Biology** Session Chair: Fan Zhang		Room: Hollywood **Computational drug discovery** Session Chair: Zhifu Sun	
10:05 -10:25	*Identification of exon skipping events associated with Alzheimer's disease in the human*	10:05 -10:25	*Investigation of multi-trait associations using pathway-based analysis of GWAS summary statistics*	10:05 -10:25	*Drug-Drug Interaction Prediction based on Co-Medication Patterns and Graph Matching*

	hippocampus Seonggyun Han, Jason Miller, Seyoun Byun, Dokyoon Kim, Shannon Leigh Risacher, Andrew Saykin, Younghee Lee and Kwangsik Nho		Guangsheng Pei, Hua Sun, Yulin Dai, Peilin Jia and Zhongming Zhao		Wen-Hao Chiang, Li Shen, Lang Li and Xia Ning
10:25 -10:45	*Brain-wide structural connectivity alterations under the control of Alzheimer risk genes* Jingwen Yan, Vinesh Raja V, Zhi Huang, Amico Enrico, Kwangsik Nho, Shaifen Fang, Olaf Sporns, Yu-Chien Wu, Andrew Saykin, Joaquin Goni and Li Shen	10:25 -10:45	*Lilikoi: an R package for personalized pathway-based classification modeling using metabolomics data* Sijia Huang, Fadhl Alakwaa and Lana Garmire	10:25 -10:45	*Predicting drug response of tumors from integrated genomic profiles by deep neural networks* Yu-Chiao Chiu, Hung-I Chen, Tinghe Zhang, Songyao Zhang, Li-Ju Wang, Yufei Huang and Yidong Chen
10:45 -11:05	*Context-sensitive Network Analysis Identifies Food Metabolites Associated with Alzheimer's Disease: An Exploratory Study* Yang Chen and Rong Xu	10:45 -11:05	*Gene2Vec: Distributed Representation of Genes Based on Co-Expression* Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao and Degui Zhi	10:45 -11:05	*Predicting Adverse Drug Reactions through Interpretable Deep Learning Framework* Sanjoy Dey, Heng Luo, Achille Fokoue-Nkoutche, Jianying Hu and Ping Zhang
11:05 -11:15	***Coffee/Tea Break***				
11:15 -11:35	*Using natural language processing and machine learning to identify breast*	11:15 -11:35	*A robust fuzzy rule based integrative feature selection strategy for gene*	11:15 -11:35	*Predicting drug sensitivity of cancer cells with pathway*

	cancer local recurrence Zexian Zeng, Sasa Espino, Ankita Roy, Xiaoyu Li, Seema Khan, Susan Clare, Xia Jiang, Richard Neapolitan and Yuan Luo		*expression data in TCGA* Shicai Fan		*activity inference* *Xuewei Wang, Zhifu Sun, Michael Zimmerman, Andrej Bugrim and Jean-Pierre Kocher*
11:35 -11:55	*Identification of long non-coding RNA-related and – coexpressed mRNA biomarkers for hepatocellular carcinoma* *Fan Zhang, Linda Ding, Li Cui, Robert Barber and Bin Deng*	11:35 -11:55	*Pessimistic optimization for modeling microbial communities with uncertainty* Meltem Apaydin, Liang Xu, Bo Zeng and Xiaoning Qian	11:35 -11:55	*Application of Transfer Learning for Cancer Drug Sensitivity Prediction* *Saugato Rahman Dhruba, Raziur Rahman, Kevin Matlock, Souparno Ghosh and Ranadip Pal*
11:55 -12:15	*Comparison of different functional prediction scores using a gene-based permutation model for identifying cancer driver genes* Alice Djotsa and Xiaoming Liu	11:55 -12:15	*Evaluation of Top-Down Mass Spectral Identification with Homologous Protein Sequences* Ziwei Li, Yunlong Liu, Xiaowen Liu and Weixing Feng	11:55 -12:15	*Large-scale mining disease comorbidity relationships from post-market drug adverse events surveillance data* Chunlei Zheng and Rong Xu
12:15 - 1:30	**Lunch Break -** Buffet lunch				
1:30 - 1:50	**Eminent Scholar Talk (Room: Grand Imperial North)** **Charles Wang, MD, PhD, MPH** **Professor & Director Center for Genomics** **Loma Linda University** Title: Vegetarian diet-modulated epigenetic reprogramming/epigenetic clocks and longevity				

1:50 - 2:10	***Award presentation***				
2:10 - 2:20	***Coffee/Tea Break***				
CONCURRENT SESSIONS					
Room: Grand Imperial North **International PI Talk** Session Chair: Yu Xue, Xiaofeng Song				Room: Santa Monica **Cancer Genomics** Session Chair: Jun Wan	
2:20 - 2:40	*Bioinformatics research on potential ability of circular RNA encoding protein* Xiaofeng Song			2:20 - 2:40	*Selecting precise reference normal tissue samples for cancer research using a deep learning approach* William Zeng, Benjamin Glicksberg, Yangyan Li and Bin Chen
2:40 - 3:00	*The impact of genetic admixture and natural selection on driving population differences in East Asia* Shuhua Xu			2:40 - 3:00	*Modeling of Hypoxia gene expression for three different cancer cell lines* Babak Soltanalizadeh, Erika Gonzalez Rodriguez, Vahed Maroufy and Hulin Wu
3:00 - 3:20	*Identification of functional PTM events in autophagy* Yu Xue			3:00 - 3:20	*Inferring gene-disease association by an integrative analysis of eQTL GWAS and Protein-Protein Interaction data* Jun Wang, Jiashun

					Zheng, Zengmiao Wang, Hao Li and Minghua Deng
3:20 - 3:40	*Tumor heterogeneity in hepatocellular carcinoma and intrahepatic cholangiocarcinoma* Ruibin Xi			3:20 - 3:40	*Network-based identification of critical regulators as putative drivers of human cleft lip* Aimin Li, Guimin Qin, Akiko Suzuki, Mona Gajera, Junichi Iwata, Peilin Jia and Zhongming Zhao
3:40	Adjourn				

Keynote Speaker
Josh Denny, MD, MS
Sunday, June 10, 2018
4:30 – 5:20 PM
Grand Imperial North

BIO

Dr. Joshua Denny is Professor of Biomedical Informatics and Medicine, Director of the Center for Precision Medicine and Vice President of Personalized Medicine at Vanderbilt University Medical Center. He is a Fellow of the American College of Medical Informatics and a member of the National Academy of Medicine.

Dr. Denny has substantial experience in the design, development, and implementation of electronic health record (EHR) data mining algorithms and was the primary author of several natural language processing systems to support phenotype extraction algorithms for genomic research projects, including development of the phenome-wide association study (PheWAS) method. He is principal investigator (PI) of nodes in the Electronic Medical Records and Genomics (eMERGE) Network, Pharmacogenomics Research Network (PGRN), and the Implementing Genomics into Practice (IGNITE) Network. He is also PI of the Data and Research Center of the Precision Medicine Initiative All of Us Research Program (previously called the Precision Medicine Initiative Cohort Program), which will eventually enroll at least 1 million Americans in an effort to understand the genetic, environmental, and behavioral factors that influence human health and disease.

To date, Dr. Denny has led >100 genome-wide and candidate gene association studies using EHR data linked to genetic data. He serves on a number of mentoring committees and has trained >30 postdoctoral and predoctoral trainees.

Keynote Speaker
Josh Denny, MD, MS
Sunday, June 10, 2018
4:30 – 5:20 PM
Grand Imperial North

Title: Huge cohorts, genomics, and clinical data to personalize medicine

Precision medicine offers the promise of improved diagnosis and for more effective, patient-specific therapies. Typically, such studies have been pursued using research cohorts. At Vanderbilt, we have linked de-identified electronic health records (EHRs), to a DNA repository, called BioVU, which has nearly 250,000 samples. Through BioVU and a NHGRI-funded network using EHRs for discovery, the Electronic Medical Records and Genomics (eMERGE) network, we have used clinical data of genomic basis of disease and drug response using real-world clinical data. The EHR also enables the inverse experiment – starting with a genotype and discovering all the phenotypes with which it is associated – a phenome-wide association study. By looking for clusters of diseases and symptoms through phenotype risk scores, we find unrecognized genetic variants associated with common disease. The era of huge international cohorts such as the UK Biobank, Million Veteran Program, and the newly started *All of Us* Research Program will make millions of individuals available with dense molecular and phenotypic data. *All of Us* launched May 6, 2018 and will engage one million diverse individuals across the US who will contribute data and also receive results back.

Keynote Speaker
Paul Thomas, Ph.D
Monday, June 11, 2018
8:40 – 9:30 AM
Grand Imperial North

BIO

Paul D. Thomas, Ph.D. is an Associate Professor in the Preventive Medicine Department, and heads the Bioinformatics Division, at the University of Southern California Keck School of Medicine. Dr. Thomas's research lab focuses on the development and application of computational methods for reconstructing gene evolution, and using these techniques to understand the function of human genes, and how genetic factors may impact disease risk. Dr. Thomas is a Principal Investigator for the Gene Ontology project, which is among the world's largest bioinformatics projects.

Keynote Speaker
Paul Thomas, Ph.D
Monday, June 11, 2018
8:40 – 9:30 AM
Grand Imperial North

Title: Reconstructing the large-scale evolution of genomes and gene functions

Over evolutionary time periods, the gene content of genomes evolves by processes of gene duplication, horizontal transfer, de novo gene origination, and gene loss. My group has reconstructed the evolutionary history of over 1 million genes in 15,000 gene families, covering all domains of life. From these gene family trees, we have inferred the gene content of common ancestral genomes, and the history of duplication, transfer, origination and loss along each branch of the species tree. I will give an overview of our reconstruction methods, and findings from this reconstruction, such as the deep history of the human genome. I will also discuss a major application of our work: inferring human gene function from experiments in "model organisms" such as the fruit fly and yeast.

Keynote Speaker
Alex Hoffmann, Ph.D
Monday, June 11, 2018
2:00 – 2:50 PM
Grand Imperial North

BIO

Dr. Alexander Hoffmann is Professor of Microbiology, Immunology, and Molecular Genetics, and Director of the Institute for Quantitative and Computational Biosciences (QCB) at UCLA. Alex's interests began to focus on biology when undergraduate research on topoisomerases provided the thrill of discovery while studying for a Physics BA at Cambridge. During his graduate research with Dr. Bob Roeder at Rockefeller University, he cloned genes for TBP and some components of the TFIID complex, and developed the now popular His-tag to purify and characterize recombinant proteins. During his postdoctoral training with Dr. David Baltimore at MIT and Caltech, he first focused on HIV and then aimed to understand the dynamic control of the NFκB signaling network and its ability to produce distinct gene expression programs. Reactivating undergraduate math and physics skills, and with the help of really smart students and postdocs, Alex has pursued a Systems Biology approach (iterating between quantitative experimentation and computational modeling) in order to understand how molecular networks generate precise immune responses to pathogens and control development of the immune system. A recurring theme of our research is that it is the kinetic properties of these regulatory networks that provide the explanations for understanding specificity, robustness, diversification, fine-tuning, and other characteristics of biological processes.

Alex is PI of the Signaling Systems Laboratory, first at UCSD (2003-2013) and then at UCLA (since 2014). At UCSD he was Professor of Chemistry and Biochemistry, and Director of the Graduate Program in Bioinformatics and Systems Biology, was PI of the P50 Center of Excellence for Systems Biology (SDCSB) and co-founded the BioCircuits Institute (BCI). At UCLA he is the Asher Professor of Microbiology in the Department of Microbiology, Immunology and Molecular Genetics (MIMG), and the director of the Institute for Quantitative and Computational Biosciences (QCBio).

Keynote Speaker
Alex Hoffmann, Ph.D
Monday, June 11, 2018
2:00 – 2:50 PM
Grand Imperial North

Learning how to predict immune responses

Precision Medicine initiatives aim to leverage the availability of clinical Big Data and the power of statistical machine learning approaches to produce predictive models for clinical decision making. In our studies of Immune Responses, we leverage the availability of Big Knowledge in Pubmed, to construct mechanistic models that we parameterize with experimental data. I will discuss some of our recent progress in understanding how molecular network dynamics and molecular noise affect immune cell function, and some of the modeling strategies that allow for prediction and insight.

Keynote Speaker
Jason Moore, Ph.D
Tuesday, June 12, 2018
8:40-9:30 AM
Grand Imperial North

BIO

Jason Moore is the Edward Rose Professor of Informatics and Director of the Penn Institute for Biomedical Informatics. He also serves as Senior Associate Dean for Informatics and Chief of the Division of Informatics in the Department of Biostatistics, Epidemiology, and Informatics. He came to Penn in 2015 from Dartmouth where he was Director of the Institute for Quantitative Biomedical Sciences. Prior to Dartmouth he served as Director of the Advanced Computing Center for Research and Education at Vanderbilt University where he launched their first high-performance computer. He has a Ph.D. in Human Genetics and an M.S. in Applied Statistics from the University of Michigan. He leads an active NIH-funded research program focused on the development of artificial intelligence and machine learning algorithms for the analysis of complex biomedical data. He is an elected fellow of the American Association for the Advancement of Science (AAAS), an elected fellow of the American College of Medical Informatics (ACMI), an elected fellow of the American Statistical Association (ASA), and was selected as a Kavli fellow of the National Academy of Sciences. He is currently serves as Editor-in-Chief of the journal *BioData Mining*.

Keynote Speaker
Jason Moore, Ph.D
Tuesday, June 12, 2018
8:40-9:30 AM
Grand Imperial North

Accessible artificial intelligence for data science

Artificial intelligence (AI) has finally emerged as a useful tool for big data analytics. This is due to decades of research that have provided powerful algorithms and the availability of plentiful computing resources. Despite these advances, AI is not widely accessible due to a steep learning curve and the expense and black box nature of software provided by commercial vendors such as IBM. In response to this need, we have developed an accessible, open-source, and user-friendly AI system at the University of Pennsylvania (PennAI) to bring AI and automated machine learning technology to everyone who wants to incorporate this technology into their big data analytics agenda. We will provide an overview of the system and an update on its features.

Eminent Scholar Talk
Xinghua Lu, MD, PhD
Monday, June 11, 2018
9:40-10:00 AM
Grand Imperial North

BIO

Dr. Xinghua Lu has broad research experiences in clinical and basic biomedical sciences. His current research interest concentrates on developing artificial intelligence (AI) methodologies and their application in translational medicine, particularly precision oncology. The ongoing research in his group spans a broad spectrum of translational informatics: using causal inference methods to understand cancer disease mechanisms, using probabilistic graphic models and deep learning models to studying cancer signaling pathways, using NLP technology to mine and annotate biomedical texts (publications and medical records), and finally using AI techniques to facilitate clinical decision making.

Eminent Scholar Talk
Xinghua Lu, MD, PhD
Monday, June 11, 2018
9:40-10:00 AM
Grand Imperial North

From Big Data to Bedside (BD2B): Precision Oncology in an Era of Artificial Intelligence

Cancer is mainly caused by somatic genome alterations. Currently, genome-scale data from individual patients are readily available, and it is anticipated that precisely targeting patient-specific genomic alterations of individual tumors will lead to more effective therapies. However, due to the sheer volume and complexity of contemporary cancer genomics data, it remains a significant challenge to efficiently utilize genomic information to guide personalized therapy. In this presentation, I will discuss different artificial intelligence technologies that can advance precision oncology, including causal inference methods for revealing the disease mechanisms of each individual tumor, causal network methods for discovering cancer pathways, and deep learning methods to infer the state of signaling machinery of tumor cells. I will further discuss how the information derived from such analyses can be used to predict efficacy of anti-cancer drugs in pre-clinical settings and their potential clinical applications.

Eminent Scholar Talk
Ting Wang, PhD
Monday, June 11, 2018
1:35-1:55 PM
Grand Imperial North

BIO

Ting Wang, Ph.D., is an Associate Professor of Genetics, Computer Science and Engineering, and Biostatistics, and the Director of Computational and Systems Biology Program at Washington University School of Medicine. Dr. Wang received undergraduate degree from Peking University in Beijing, China, and PhD in Computational Biology in 2006 from Washington University. He was a Helen Hay Whitney Fellow at University of California Santa Cruz before returning to Washington University in 2009 to start his own lab in the Department of Genetics and the Center for Genome Sciences and Systems Biology. Dr. Wang's research focuses on understanding genetic and epigenetic impact of transposable element on regulatory networks. He investigates DNA methylation dynamics during normal development, cancer development, and in evolution. He develops algorithms for identifying regulatory motifs and modules, and analytical and visualization technologies to integrate large genomic and epigenomic data. He invented the WashU Epigenome Browser as part of the NIH Roadmap Epigenome Project. Dr. He is the Principle Investigator of six NIH funded projects, including an R01 to develop tools for integrative epigenomic data analysis and visualization, two R01s from NIH to investigate transposable elements' role in gene regulation, one U01 from NIH to investigate 3D genome structure and to establish a data visualization center for the 4D Nucleome Network, one U24 from NIH to establish a data coordination center for environmental epigenomics, and one U01 from NIH to participate the ENCODE Consortium. He is also funded by a research scholar grant from American Cancer Society to study DNA methylation of enhancers in cancer development.

Eminent Scholar Talk
Ting Wang, PhD
Monday, June 11, 2018
1:35-1:55 PM
Grand Imperial North

Title: Exploring the dark matter in genomics data

Advances in next-generation sequencing platforms have reshaped the landscape of genomic and epigenomic research towards the understanding of roles of human transposable elements. It is now possible to map the epigenetic landscape of transposable element (TE) across many tissue and cell types as well as in diseases. In the presentation I will discuss tools developed for this purpose and present some results from investigating regulatory roles of transposable elements using public genomic and epigenomic resources.

Eminent Scholar Talk
Grace Xiao, Ph.D
Tuesday, June 12, 2018
9:40-10:00 AM
Grand Imperial North

BIO

Dr. Xinshu (Grace) Xiao is a Professor and Vice Chair of the Department of Integrative Biology and Physiology at the University of California, Los Angeles (UCLA). She is also a member of the Institute for Quantitative and Computational Biology, the Molecular Biology Institute and the Jonsson Comprehensive Cancer Center at UCLA. Dr. Xiao received her Bachelor's degree from Tsinghua University in China and her PhD degree from the Massachusetts Institute of Technology (MIT). Dr. Xiao's research focuses on the bioinformatics and genomics of RNA biology. She has won a number of prestigious honors including an Alfred P. Sloan Foundation Research Fellowship. Work in the Xiao Lab is highly interdisciplinary, bridging bioinformatics, genomics, systems biology and basic molecular biology. Current research topics of the Xiao Lab include the computational and experimental studies of alternative splicing and its regulation, RNA editing and small RNA regulation of gene expression. The Xiao lab is supported by various funding sources, including an ENCODE Computational Analysis Award.

Eminent Scholar Talk
Grace Xiao, PhD
Tuesday, June 12, 2018
9:40-10:00 AM
Grand Imperial North

Deciphering the function of single-nucleotide variants in the RNA

Single-nucleotide variants (SNVs) are abundant in both the genome and the transcriptome. High-throughput sequencing technologies have greatly facilitated the identification and functional characterization of SNVs, thus enabling substantial progress towards precision medicine. Compared to studies of SNVs in the DNA, approaches to identify and characterize expressed SNVs in the RNA are powerful in revealing multiple types of SNVs and their potential function in gene regulation. To this end, we have developed a battery of bioinformatic methods to study SNVs in RNA-seq data. With sequencing errors excluded, SNVs in the RNA-seq reflect existence of genetic variants or RNA editing sites, both of which could be essential players in disease diagnostics, basic mechanisms and biology. As part of the ENCODE consortium, we have developed and applied new methods to identify functional SNVs in mediating alternative splicing. Application of these methods to a large amount of RNA-seq data revealed novel insights into the functional roles of many SNVs.

Eminent Scholar Talk
Charles Wang, MD, PhD, MPH
Tuesday, June 12, 2018
1:30-1:50 PM
Grand Imperial North

BIO

Dr. Wang, MD, PhD, MPH, is the founding Director of the Center for Genomics and a tenured full-professor in the School of Medicine, Loma Linda University (LLU). He received his MD (certified by the US ECFMG) and MPH from the Tongji Medical University, Wuhan, China, and his PhD from University of Washington, Seattle, WA. Prior to joining LLU, Dr. Wang has held various positions both in the US government agencies such as the US DOE's Argonne National Lab, the US FDA's National Center for Toxicological Research, the US Department of Defense's Air Force Research Laboratory, and in the academia such as Assistant & Associate Professor of Medicine at the David Geffen School of Medicine at UCLA, Director of the Clinical Transcriptional Genomics Core at Cedars-Sinai Medical Center and Scientific Director of the Functional Genomics Core at City of Hope National Medical Center. Dr. Wang is a recipient of several awards, including the Qianjiang Distinguished Expert (Hangzhou, China), the Distinguished Professor (Guangzhou Medical University, China), the AACR-Bristol-Myers Squibb Young Investigator Award of the American Association for Cancer Research (AACR), and the first place award for Research Excellence of Society of Toxicology (SOT). Dr. Wang serves as Editorial Board Members for more than 10 different journals including the *Scientific Data (Nature Publishing Group)*, the *Frontiers*, the *Single Cell Genomics & Proteomic*s, and the *Clinical Precision Medicine (as an Associate Editor)*. Dr. Wang is a well-recognized expert on genomics, epigenomics, transcriptomics and single-cell sequencing. Dr. Wang was one of the key project leaders for the FDA Sequencing Quality Control (SEQC) consortium-based international collaborations, and he is currently leading the SEQC-2 consortium single-cell sequencing collaboration project. Dr. Wang has published many high impact peer-reviewed papers in several high profile journals such as *Nature Biotechnology*, *Nature Communications* and *PNAS*. He has been an invited speaker and/or a session chair for many prestigious genomics conferences such as the Global Technology Community (GTC) Single Cell Analysis Conference, Cambridge, MA, the Wellcome Trust Scientific Conference, Cambridge, UK and the Festival of Genomics, San Francisco, CA.

Eminent Scholar Talk
Charles Wang, MD, PhD, MPH
Tuesday, June 12, 2018
1:30-1:50 PM
Grand Imperial North

Vegetarian diet-modulated epigenetic reprogramming and longevity

Diet importantly impacts health, disease resistance, and longevity, but the molecular mechanisms mediating the effects of diet are poorly understood. The major reason for this profound knowledge gap is that we don't yet have a comprehensive understanding of how diet influences specific epigenomes. To address this knowledge gap, we exploit a powerful, well-characterized, long-standing cohort that has already made many significant advances: the Adventist Health Study 2 (AHS2), which constitutes a multi-ethnic group of Seventh Day Adventists whose detailed diets are known. Our **CENTRAL MODEL** is that long-term dietary patterns cause epigenomic reprogramming in ways that alter disease susceptibility, maintenance of health, and longevity. We performed DNA methylome studies using Illumina 450K BeadChips, TruSeq Epic next-gen sequencing (NGS) technologies and the frozen human white blood cells derived DNA from AHS2 participants consisting of three different groups: vegan, pescetarian and non-vegetarians. Age, gender, race, BMI, exercise, drinking and smoking status were balanced in each group. We determine the impact of dietary factors on both Horvath and Hannum epigenetic clocks, which are algorithms that use DNA methylation patterns to estimate biological age of specific cells and tissues. We also developed a new epigenetic clock, the "LLU clock", which, unlike Horvath and Hannum epigenetic clocks that only work for the array-based data, can measure the DNA methylation age using both array and NGS data. Our LLU epigenetic clock shows consistent better prediction of biological age and performance metrics than Horvath and Hannum clocks using many benchmark samples. Overall, our results show that a lifetime vegetarian diet can reprogram the epigenomes, which contributes approximately 5 extra years of lifespan compared to non-vegetarians based on our AHS2 cohort. Our pilot DNA methylome study support our **HYPOTHESIS** that maintaining a specific diet causes epigenomic reprogramming leading to programming of health and longevity in humans.

Deep architectures are not necessary for anomaly classification of matrix-formed omics data

Yan Guo PhD,
Associate & Endowed Professor in Cancer Bioinformatics and Computational Biology.
Director of Bioinformatics Shared Resources, Comprehensive Cancer Center
University of New Mexico

Abstract

Motivation

While deep learning has made breathtaking successes in tackling sequence-based problems, its effectiveness in phenotype classification using numerical matrix-formed omics data remains under studied. It is informative to compare deep learning neural networks with classical machine learning methods in the setting of high throughput omics data for anomaly classification purpose. Using 37 high throughput datasets, covering transcriptomes and metabolomes, we evaluated the classification power of deep learning and traditional machine learning methods. Representative deep learning methods, Multi-Layer Perceptrons (MLP) and Convolutional Neural Networks (CNN), were deployed and explored in seeking optimal architectures for the best classification performance. Together with five classical supervised classification methods (Linear Discriminant Analysis, Multinomial Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine), MLP and CNN were comparatively tested on the 37 datasets to predict disease stages or discriminate diseased samples from normal samples.

Results

Single layer MLPs with ample hidden units outperformed deeper MLPs, and 64 to 128 hidden neurons seemed sufficient to yield highest prediction accuracy. MLPs of two hidden layers returned nearly as good performance as single layered MLPs and they were more robust than single layer MLPs in case of extremely imbalanced class composition. CNN was not conspicuous in either accuracy or robustness, even non-comparable to most traditional methods. Summarizing the results across all 37 datasets, MLP achieved the highest accuracy among all methods tested, and it was one of the most robust methods against imbalanced class composition and inaccurate class labels. Secondary findings include that a drop-out layer consistently promoted the classification accuracy for MLP but not for CNN, and that MLP and CNN cost tens of thousands times more computation time than Linear Discriminant Analysis.

Conclusions

The comparative results proved that well-configured single-layer or two-layer MLPs are a good choice for anomaly classification of matrix-formed omics data. Shallow MLPs with ample hidden neurons are sufficient to achieve superior classification performance in handling numerical matrix-formed omics data. Conclusions and guidance generated from this study are helpful for improving future neural network applications on omics data matrices.

Concurrent-Workshops
Sunday, June 10, 2018
2:30-4:20 PM
Santa Monica

Single-cell sequencing analysis

Dr. Chi Zhang, Indiana University School of Medicine
Dr. Xiao Dong, Albert Einstein College of Medicine

Single cell sequencing technologies provide quantification of genomics/transcriptomics variations on single cell level. The workshop will include two parts, Dr. Chi Zhang will introduce single cell RNA-sequencing technology and its data modeling, and Dr. Xiao Dong will focus on single cell genomics data.

Single cell RNA-seq (scRNA-Seq) technology emerged to unmask individual cellular transcriptomic profiles in a tissue or bulk cell sample, which are ideal objects to build reference maps of single cell behaviors in a real disease environment. New computational challenges arises in analysis of scRNA-seq data due to heterogeneity among cells and limitations of the technology. The scRNA-seq introduction will include: (1) a general overview of the experimental workflow of C1 Fluidigm and 10x genomics techniques, (2) scRNA-seq data processing, quality control, analysis and visualization, (3) statistical models for scRNA-seq data and differential expression test methods, and (4) identification of cell types in scRNA-seq data, with a focus on intra-tumor heterogeneity.

Workshop on 3D genome organization

Feng Yue PhD,
Director, Bioinformatics Division
Institute for Personalized Medicine
Assistant Professor of Biochemistry and Molecular Biology
Penn State College of Medicine

Abstract

Recent advent of 3C-based technologies such as Hi-C and ChIA-PET provides an opportunity to explore chromatin interactions and 3D genome organization in an unprecedented scale and resolution. However, it remains challenging to visualize and explore chromatin interaction data due to its size and complexity. In this workshop, we will introduce the basic concept and methods to study 3D genomics. We will illustrate how to use some popular tools for users to explore both their own chromatin interaction data and publicly available data, including those generated by the ENCODE and 4D Nucleome projects. We will also show users how integrate other “omics” data sets, such as GWAS, ChIP-Seq and RNA-Seq data, and therefore gain a complete view of the regulatory landscape for any given gene. Our workshop will also provides multiple methods to link distal *cis*-regulatory elements with their potential target genes, and therefore represents a valuable resource for the study of gene regulation in mammalian genomes.

WashU Epigenome Browser

Ting Wang
Associate Professor of Genetics and Computer Science and Engineering,
Washington University at St. Louis.

Abstract

This interactive workshop will introduce The WashU Epigenome Browser and associated tools (http://epigenomegateway.wustl.edu/). The Epigenome Browser hosts data from ENCODE Roadmap Epigenomics and 4DN projects, and support navigation of the data and its interactive visualization, integration, comparison, and analysis. Annotating the non-coding, regulatory genome with genomic and epigenomic data has provided new venues to interpret the functional consequences of genetic variants associated with human complex traits. Attendees will gain hands-on experience with exploring the most current epigenomic resources, and with advanced visual-bioinformatics tools including gene set view, genome juxtaposition, and chromatin-interaction display, inventions unique to the WashU Epigenome Browser. Through specific examples the workshop will demonstrate the power of annotating genetic variants with epigenomic data, and how it can be easily achieved with the Epigenome Browser.

Development of somatic mutation signatures for risk stratification and prognosis in lung and colorectal adenocarcinomas

Mark Menor[1], Yong Zhu[2], Bin Jiang[2], Youping Deng[1,3*]
* Corresponding author
[1]Department of Complementary & Integrative Medicine, University of Hawaii John A. Burns School of Medicine, Honolulu, HI, USA
[2]National Medical Centre of Colorectal Disease, The Third Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, People's Republic of China
[3]University of Hawaii Cancer Center, Honolulu, HI, USA

Background:
Prognostic signatures are vital to precision medicine. However, development of somatic mutation prognostic signatures for cancers remains a challenge. In this study we developed a novel method for discovering somatic mutation based prognostic signatures.

Results:
Somatic mutation and clinical data for lung adenocarcinoma (LUAD) and colorectal adenocarcinoma (COAD) from The Cancer Genome Atlas (TCGA) were randomly divided into training (n=328 for LUAD and 286 for COAD) and validation (n=167 for LUAD and 141 for COAD) datasets. A novel method of using the log2 ratio of the tumor mutation frequency to the paired normal mutation frequency is computed for each patient and missense mutation. The missense mutation ratios were mean aggregated into gene-level somatic mutation profiles. The somatic mutations were assessed using univariate Cox analysis on the LUAD and COAD training sets separately. Stepwise multivariate Cox analysis resulted in a final gene prognostic signature for LUAD and COAD. Performance was compared to gene prognostic signatures generated using the same pipeline but with different somatic mutation profile representations based on tumor mutation frequency, binary calls, and gene-gene network normalization. Signature high-risk LUAD and COAD cases had worse overall survival compared to the signature low-risk cases in the validation set (log-rank test p-value=0.0101 for LUAD and 0.0314 for COAD) using mutation tumor frequency ratio (MFR) profiles, while all other methods, including gene-gene network normalization, have statistically insignificant stratification (log-rank test p-value ≥ 0.05). Most of the genes in the final gene signatures using MFR profiles are cancer-related based on network and literature analysis.

Conclusions:
We demonstrated the robustness of MFR profiles and its potential to be a powerful prognostic tool in cancer. The results are robust according to validation testing and the selected genes are biologically relevant.

Genomic copy number variations in large screening for pediatrics sarcomas chemotherapy

Lijun Cheng[1], Pooja Chandra[2], Limei Wang[1], Pankita H Pandya[3], Karen E. Pollok[3], Mary E. Murray[3], Jacquelyn Carter[3], Michael Ferguson[3], Mohammad Reza Saadatzadeh[3], Mashall Mark[3], Li Lang[1,2*] and Jamie L Renbarger[3,4*]
* Corresponding author
[1] Department of Biomedical informatics, College of medicine, Ohio State University, Columbus, OH 43210
[2] Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, IN 46202
[3] Department of Pediatrics, hematology/oncology, School of Medicine, Indiana University, Indianapolis, IN, 46202
[4] Indiana Institute of Personalized Medicine, Indiana University, Indianapolis, IN 46202

Research into sarcoma has historically been challenging as rare and heterologous. The activity of first-line systemic chemotherapy in advanced pediatric sarcoma remains suboptimal, with only 20-40% patients demonstrating objective radiographic responses (ORRs). Complex chromosomal aberrations such as amplification and deletion of DNA copy number are frequently observed to pediatric sarcoma. This systematic copy number variation (CNV) comparison makes it possible to catch DNA copy number changes and identify chromosomal regions containing “target genes” responsible for sarcoma recurrent development and/or progression.

The paper aims to detect the prognosis biomarkers for Osteosarcoma (OS), Rhabdomyosarcoma (RMS) and Ewing’s Sarcoma (ES) based on copy number aberration for clinical chemotherapy systematically. The 206 CNV profiles from pediatric sarcoma patients across three types of sarcoma are collected and analyzed. Systematically significant CNVs of chromosome are detected. By comparing with healthy population, frequently amplification and deletion of sixty three genes recurrent-related are observed and validated in OS, RMS and EWS.

By bridging molecular biomarkers between sarcomas tumors and cell lines, integrating large scale of drugs screening with CNV analysis on 38 sarcoma cancer cells, 19 chemotherapies associated with 32 potential prognosis biomarkers of DNA copy variation are recommended to guide pediatrics OS, ES, and RMS treatment respectively, including oncogenes MYC, RAD21, MAPK1, ATF1 and MDM2 etc amplification. The research not only detects CNVs both of sarcoma cell lines and tumors, but also provides novel insights into chemotherapy biomarkers based on copy number amplification or deletion in pediatric bone and soft tissue sarcoma irrespectively.

Computational identification of deleterious synonymous variants in human genomes using a feature-based approach

Fang Shi[1,#], Yao Yao[2,#], Yannan Bin[2], Chun-Hou Zheng[1], and Junfeng Xia[2,*]

[#]These authors contributed equally to this work.
[*]Corresponding author

[1]College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230601, China
[2]Institute of Physical Science and Information Technology, School of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China

Background: Although synonymous single nucleotide variants (sSNVs) do not alter the protein sequences, they have been shown to play an important role in human disease. Distinguishing pathogenic sSNVs from neutral ones is challenging because pathogenic sSNVs tend to have low prevalence. Although many methods have been developed for predicting the functional impact of single nucleotide variants, only a few have been specifically designed for identifying pathogenic sSNVs.
Results: In this work, we describe a computational model, IDSV (Identification of Deleterious Synonymous Variants), which uses random forest (RF) to detect deleterious sSNVs in human genomes. We systematically investigate a total of 74 multifaceted features across seven categories: splicing, conservation, codon usage, sequence, pre-mRNA folding energy, translation efficiency, and function regions annotation features. Then, to remove redundant and irrelevant features and improve the prediction performance, feature selection is employed using the sequential backward selection method. Based on the optimized 10 features, a RF classifier is developed to identify deleterious sSNVs. The results on benchmark datasets show that IDSV outperforms other state-of-the-art methods in identifying sSNVs that are pathogenic.
Conclusions: We have developed an efficient feature-based prediction approach (IDSV) for deleterious sSNVs by using a wide variety of features. Among all the features, a compact and useful feature subset that has an important implication for identifying deleterious sSNVs is identified. Our results indicate that besides splicing and conservation features, a new translation efficiency feature is also an informative feature for identifying deleterious sSNVs. While the function regions annotation and sequence features are weakly informative, they may have the ability to discriminate deleterious sSNVs from benign ones when combined with other features. The data and source code are available on website http://bioinfo.ahu.edu.cn:8080/IDSV.

High scoring segment selection for pairwise whole genome sequence alignment with the maximum scoring subsequence and GPUs

Abdulrhman Aljouie, Ling Zhong, and Usman Roshan*
Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102

* Corresponding author

Whole genome alignment programs use exact string matching with hash tables to quickly identify high scoring fragments between a query and target sequence around which a full alignment is then built. In a recent large-scale comparison of alignment programs called Alignathon it was discovered that while evolutionary similar genomes were easy to align, divergent genomes still posed a challenge to existing methods. As a first step to fill this gap we explore the use of more exact methods to identify high scoring fragments which we then pass on to a standard pipeline. We identify such segments between two whole genome sequences with the maximum scoring subsequence instead of hash tables. This is computationally extremely expensive and so we employ the parallelism of a Graphics Processing Unit to speed it up. We split the query genome into several fragments and determine its best match to the target with a previously published GPU algorithm for aligning short reads to a genome sequence. We then pass such high scoring fragments on to the LASTZ program which extends the fragment to obtain a more complete alignment. Upon evaluation on simulated data, where the true alignment is known, we see that this method gives an average of at least 20% higher accuracy than the alignment given by LASTZ at the expense of a few hours of additional runtime. We make our source code freely available at web.njit.edu/\˜usman/maxsubgenomealign.

Detecting virus integration sites based on mul- tiple related sequencing data by VirTect

Yuchao Xia[1#], Yun Liu[1#], Minghua Deng[1,2] and Ruibin Xi[1,3,*]
[1] School of Mathematical Sciences, Peking University, Beijing, 100871, China.
[2] Center for Quantitative Biology, Peking University, Beijing, 100871, China
[3] Center for Statistical Science, Peking University, Beijing, 100871, China

[#] These authors contributed equally to this work.
[*] Corresponding author

Background: Since tumor often has a high level of intra-tumor heterogeneity, multiple tumor samples from the same patient at different locations or different time points are often sequenced to study tu- mor intra-heterogeneity or tumor evolution. In virus-related tumors such as human papillomavirus- and Hepatitis B Virus-related tumors, virus genome integrations can be critical driving events. It is thus important to investigate the integration sites of the virus genomes. Currently, a few algorithms for detecting virus integration sites based on high-throughput sequencing have been developed, but their insufficient performance in their sensitivity, specificity and computational complexity hinders their ap- plications in multiple related tumor sequencing.

Results: We develop VirTect for detecting virus integration sites simultaneously from multiple relat- ed-sample data. This algorithm is mainly based on the joint analysis of short reads spanning break- points of integration sites from multiple samples. To achieve high specificity and breakpoint accuracy, a local precise sandwich alignment algorithm is used. Simulation and real data analyses show that, compared with other algorithms, VirTect is significantly more sensitive and has a similar or lower false discovery rate.

Conclusions: VirTect can provide more accurate breakpoint position and is computationally much more efficient in terms both memory requirement and computational time.

Comprehensive assessment of genotype imputation performance

Shuo Shi[1,2,3], Na Yuan[2], Ming Yang[4], Zhenglin Du[2], Jinyue Wang[1,2,3], Xin Sheng[1,2,3], Jiayan Wu[1]*, Jingfa Xiao[1,2,3]*

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, People's Republic of China.
[2]Big Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, People's Republic of China.
[3]University of Chinese Academy of Sciences, 100049 Beijing, People's Republic of China.
[4]Office of General Affairs, Chinese Academy of Sciences, Beijing, 100864, China.

*Corresponding author

Background: Genotype imputation is a process of estimating missing genotypes from the haplotype or genotype reference panel. It can effectively boost the power of detecting SNPs in genome-wide association studies (GWASs), integrate multi-studies for meta-analysis, and be applied in fine-mapping studies. The performance of genotype imputation is affected by many factors, including software, reference selection, sample size, and single-nucleotide polymorphism (SNP) density/sequencing coverage. Systematically evaluation of the imputation performance of current popular software will benefit future studies.

Results: The result indicated the accuracy of IMPUTE2 (99.18%) is slightly higher than the others (Beagle4.1: 98.94%, MACH+Minimac3: 98.51%, and SHAPEIT2+IMPUTE2: 99.08%). To achieve good and stable imputation quality, the minimum requirement of SNP density needed to be greater than 200/Mb. The imputation accuracies of IMPUTE2 and Beagle4.1 were under the minor influence of the study sample size. The contribution extent of reference to genotype imputation performance relied on software selection. We assessed the imputation performance on SNPs generated by next-generation whole genome sequencing, found that SNPs set detected by sequencing with 15x depth could be mostly got by imputing from haplotype reference panel of 1000 Genomes Project based on SNP data detected by sequencing with 4x depth. All of the imputation software had a weaker performance in low minor allele frequency SNP regions because of the bias of reference or software. In the future, more comprehensive reference panel or new algorithm development may rise up to this challenge.

Conclusions: Systematically evaluation of the influences of software, reference selection, SNP density, and sample size on imputation performance can give a general and practical guidance for future imputation study.

Reconstructing the high-resolution chromosome three-dimensional structures by Hi-C complex networks

Tong Liu[1] and Zheng Wang[1,*]

[1]Department of Computer Science, University of Miami, 1365 Memorial Drive, P.O. Box 248154, Coral Gables, FL, 33124, USA

*Corresponding author

Background: Hi-C data have been widely used to reconstruct chromosomal 3D structures. One of the key limitations of Hi-C is the unclear of the relationship between spatial distance and the number of Hi-C contacts. Many methods used a fixed parameter when converting number of Hi-C contacts to wish distances. However, a single parameter cannot properly explain the relationships between wish distances and genomic distances or locations of topologically associating domains (TADs).

Results: We tried to address one of the key issues of using Hi-C data, that is, the unclear relationship between spatial distance and the number of Hi-C contacts, which is crucial to understand significant biological functions, such as the enhancer-promoter interactions. Specifically, we developed a new method to infer this converting parameter and pairwise Euclidean distances based on the topology of Hi-C complex network (HiCNet). The inferred distances were modeled by clustering coefficient and multiple other types of constraints. We found that our inferred distances between beads within the same TAD are apparently smaller than those distances between beads from different TADs. Our inferred distances had a higher correlation with fluorescence in situ hybridization (FISH) data,

fitted the localization patterns of Xist transcripts on DNA, and better matched 156 pairs of protein-enabled long-range chromatin interactions detected by ChIA-PET. Using the inferred distances and another round of optimization, we further reconstructed 40 kb high-resolution 3D chromosomal structures of mouse male ES cells. The high-resolution structures illustrate TADs and DNA loops (peaks in Hi-C contact heatmaps) that usually indicate enhancer-promoter interactions.

CeL-ID: Cell line identification using RNA-seq data

Tabrez A Mohammad[1], Yun S Tsai[1], Safwa Ameer[1], Hung-I Harry Chen[1], Yu-Chiao Chiu[1], Yidong Chen[1, 2*]

[1]Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA
[2]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

*Corresponding author

Background: Cell lines form the cornerstone of cell-based experimentation studies into understanding the underlying mechanisms of normal and disease biology including cancer. However, it is commonly acknowledged that contamination of cell lines is a prevalent problem affecting biomedical science and available methods for cell line authentication suffer from limited access as well as being too daunting and time-consuming for many researchers. Therefore, a new and cost effective approach for authentication and quality control of cell lines is needed.

Results: We have developed a new RNA-seq based approach named CeL-ID for cell line authentication. CeL-ID uses RNA-seq data to identify variants and compare with variant profiles of other cell lines. RNA-seq data for 934 CCLE cell lines downloaded from NCI GDC were used to generate cell line specific variant profiles and pair-wise correlations were calculated using frequencies and depth of coverage values of all the variants. Comparative analysis of variant profiles revealed that variant profiles differ significantly from cell line to cell line whereas identical, synonymous and derivative cell lines share high variant identity and are highly correlated ($\rho > 0.9$). Our benchmarking studies revealed that CeL-ID method can identify a cell line with high accuracy and can be a valuable tool of cell line authentication in biomedical science. Finally, CeL-ID estimates about the possible cross contamination using linear admixture model if no perfect match was detected.

Conclusions: In this study, we show utility of an RNA-seq based approach for cell line authentication. Our comparative analysis of variant profiles derived from RNA-seq data revealed that variant profiles of each cell line are distinct and overall share low variant identity with other cell lines whereas identical or synonymous cell lines show significantly high variant identity and hence variant profiles can be used as a discriminatory/identifying feature in cell authentication model.

Comparison of SureSelect and Nextera exome capture performance in single-cell sequencing

Wendy J. Huss[1,#], Qiang Hu[2,#], Sean T. Glenn[3,4,#], Kalyan J. Gangavarapu[1], Jianmin Wang[2], Jesse D. Luce[3], Paul K. Quinn[3], Elizabeth A. Brese[4], Fenglin Zhan[5], Jeffrey M. Conroy[3], Gyorgy Paragh[6], Barbara A. Foster[1], Carl D. Morrison[3], Song Liu[2], Lei Wei[2,*]

[#] These authors contributed equally
*Corresponding author

[1]Department of Pharmacology and Therapeutics,
[2]Departments of Biostatistics and Bioinformatics,
[3]Center of Personalized Medicine,
[4]Department of Molecular and Cellular biology,
[3]Department of Cancer Genetics,
[4]Department of Pathology,
[6]Department of Dermatology
[5]Roswell Park Comprehensive Cancer Center, Buffalo, NY, 1314263 PET/CT center,
The First Affiliated Hospital of University of Science and Technology of China, Hefei, China, 230001

Background: Advances in single-cell sequencing (SCS) provide unprecedented opportunities for clinical examination of circulating tumor cells, cancer stem cells and other rare cells responsible for disease progression and drug resistance. On the genomic level, single-cell whole exome sequencing (scWES) started to gain popularity with its unique potentials in characterizing mutational landscapes at a single cell level. Currently, thereis little known about the performance of different exome capture kits in scWES. Nextera rapid capture (NXT, Illumina Inc.) has been the only exome capture kit recommended for scWES by Fluidigm C1, a widely accessed system in single-cell preparation.

Results: In this study, we compared the performance of NXT with Agilent SureSelect XT Target Enrichment System (AGL), another exome capture kit widely used for bulk sequencing. We created DNA libraries of 192 single cells isolated from spheres grown from a melanoma specimen using Fluidigm C1. Twelve high-yield cells were selected to perform dual-exome capture and sequencing using AGL and NXT in parallel. After mapping and coverage analysis, AGL outperformed NXT in coverage uniformity, mapping rates of reads, exome capture rates and low PCR duplicate rates. For germline variant calling, AGL achieved better performance in overlap with known variants in dbSNP and transition-transversion ratios. Using calls from high coverage bulk sequencing from blood DNA as the golden standard, AGL-based scWES demonstrated high positive predictive values, and medium to high sensitivity. Lastly, we evaluated somatic mutation calling by comparing single cell data with the matched blood sequence as control. On average three hundred mutations were identified in each cell. In 10 of 12 cells, higher numbers of mutations were identified by using AGL than NXT, probably caused by coverage depth. When mutations are adequately covered in both AGL and NXT data, the two methods showed very high concordance (93-100% per cell).

Conclusions: Our results suggest AGL can also be used for scWES when there is sufficient DNA, and yields better data quality than NXT.

A PheWAS Study of a Large Observational Epidemiological Cohort of African Americans: the REGARDS Study

Xueyan Zhao[1,#], Xin Geng[2,#], Vinodh Srinivasasainagendra[1], Suzanne Judd[1], Virginia Wadley[3], Orlando Gutierrez[3], Henry Wang[4], Ethan Lange[5], Leslie Lange[5], Daniel Woo[6], Fred Unverzagt[7], Monika Safford[8], Mary Cushman[9], Nita Limdi[10], Rakale Quarells[11], Donna K. Arnett[12], Marguerite R. Irvin[13*], Degui Zhi[2,14,*]

[1] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, 35233, USA
[2] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, 77030, USA
[3] Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, 35233, USA
[4] Department of Emergency Medicine, University of Alabama at Birmingham, Birmingham, AL, 35233, USA
[5] Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, 80045 USA
[6] Department of Neurology, University of Cincinnati Medical Center, OH, 45219, USA
[7] School of Medicine, Indiana University, Indianapolis, IN, 46202, USA
[8] Division of General Internal Medicine, Weill Cornell Medical College, Cornell University, New York, NY, 10065, USA
[9] Department of Pathology, Larner College of Medicine at the University of Vermont, Burlington, VT, 05405, USA
[10] Department of Neurology, University of Alabama at Birmingham, Birmingham, AL, 35294, USA

[11] Community Health & Preventive Medicine, Morehouse School of Medicine, Atlanta, GA, 30310
[12] College of Public Health, University of Kentucky, Lexington, KY, 40506, USA
[13] Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, 35233,USA
[14] School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, 77030, USA

[#] These authors contributed equally
*These authors jointly supervised this work.

Cardiovascular disease, diabetes, and renal disease are the leading causes of death and disability worldwide. However, knowledge of genetic determinants of those diseases in African Americans remains limited. In our study, associations between 4,956 GWAS catalog reported SNPs and 67 traits were examined using 7,726 African Americans from REasons for Geographic and Racial Differences in Stroke (REGARDS) study, which is focused on identifying factors that increase stroke risk. The prevalent and incident phenotypes studied included inflammation, kidney traits, cardiovascular traits and cognition. Our results validated 29 known associations, of which eight associations in African Americans were reported for the first time. Our cross-ethnic validation of GWAS findings provide additional evidence for the important roles of these loci in the disease process and may help identify genes especially important for future functional validation.

NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data

Qian Liu[1#], Daniela C. Georgieva[2], Dieter M. Egli[3], Kai Wang[1,4*#]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[2] Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY 10032, USA
[3] Department of Pediatrics, Columbia University, New York, NY 10032, USA
[4] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[#]Previous address: Department of Biomedical Informatics and Institute for Genomic Medicine, Columbia University, New York, NY 10032, USA.
[*]Corresponding author

Background: Recent advances in single-molecule sequencing techniques, such as Nanopore sequencing, improved read length, increased sequencing throughput, and enabled direct detection of DNA modifications through the analysis of raw signals. These

DNA modifications include naturally occurring modifications such as DNA methylations, as well as modifications that are introduced by DNA damage or through synthetic modifications to one of the four standard nucleotides.

Methods: To improve the performance of detecting DNA modifications, especially synthetically introduced modifications, we developed a novel computational tool called NanoMod. NanoMod takes raw signal data on a pair of DNA samples with and without modified bases, extracts signal intensities, performs base error correction based on a reference sequence, and then identifies bases with modifications by comparing the distribution of raw signals between two samples, while taking into account of the effects of neighboring bases on modified bases ("neighborhood effects").

Results: We evaluated NanoMod on simulation data sets, based on different types of modifications and different magnitudes of neighborhood effects, and found that NanoMod outperformed other methods in identifying known modified bases. Additionally, we demonstrated superior performance of NanoMod on an E. coli data set with 5mC (5-methylcytosine) modifications.

Conclusions: In summary, NanoMod is a flexible tool to detect DNA modifications with single-base resolution from raw signals in Nanopore sequencing, and will greatly facilitate large-scale functional genomics experiments in the future that use modified nucleotides.

Joint Principal Trend Analysis for Longitudinal High-Dimensional Data

Yuping Zhang[1,2,3], Zhengqing Ouyang[4,5,6]
[1]Department of Statistics, University of Connecticut, Storrs, Connecticut, U.S.A.
[2]Center for Quantitative Medicine, University of Connecticut Health Center, Farmington, Connecticut, U.S.A.
[3]Institute for Systems Genomics, Institute for Collaboration on Health, Intervention, and Policy, CT Institute of the Brain and Cognitive Sciences, University of Connecticut, Storrs, Connecticut, U.S.A.
[4]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, U.S.A.
[5]Department of Biomedical Engineering, Institute for Systems Genomics, University of Connecticut, Storrs, Connecticut, U.S.A.
[6]Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, Connecticut, U.S.A

We consider a research scenario motivated by integrating multiple sources of information for better knowledge discovery in diverse dynamic biological processes. Given two longitudinal high-dimensional datasets for a group of subjects, we want to extract shared latent trends and identify relevant features. To solve this problem, we present a new

statistical method named as joint principal trend analysis (JPTA). We demonstrate the utility of JPTA through simulations and applications to gene expression data of the mammalian cell cycle and longitudinal transcriptional profiling data in response to influenza viral infections.

Inferring Drug-Target Associations based on perturbational profiles in L1000 Data

Pei-Han Liao, Tzu-Hung Hsiao, Liang-Chuan Lai, Mong-Hsun Tsai, Tzu-Pin Lu and Eric Y. Chuang

Background

The cost of new drug development has been increasing whereas the productivity has not been growing correspondingly. Drug repurposing is a new approach to shorten the procedure of new drug discovery. The important key of drug repurposing is to identify the drug target genes and affected functions. In this study, we developed a novel algorithm to identify drug target genes through gene expression profiles. Utilizing the LINCS L1000 dataset, which contains expression profiles of near 20,000 compound treatments, not only the putative compound affected genes could be identified, also the affected functions can be explored.

Methods

We hypothesized that the affected gene of a compound treatment and the targeted gene of a knockdown experiment would be the same. Under the assumption, we proposed an approach to elucidate the association between drugs and their targets based on the similarity of differentially expressed genes. The genes with significant changes of each expression profiles were identified. Next, we used hypergeometric test to assess the similarity between two profiles based on their differentially expressed gene lists. We also applied gene set enrichment analysis to further identify the enriched functions between the interactions of compounds and identified target genes based on the overlap of their differentially expressed genes. Through our method, not only compound-target pairs, but also their enriched functions could be identified.

Results

More than 0.6 million transcriptional profiles perturbed by chemical compounds and genetic treatments were collected from LINCS L1000 dataset, including 4,343 distinct shRNAs and 19,795 distinct compounds. After applying our novel algorithm to the dataset, we comprehensively identified the putative drug targets, and constructed a drug-target network. For a drug-target pair which affects the functions of cell cycle, the drug is considered to have potential to exert anti-proliferative properties and the target gene is regarded as a potential therapeutic target of the drug. We also validated our result though the reported drug-target information from Drug Repurposing Hub.

Conclusions

Here we developed an approach to elucidate the association between compounds and their potential targets systematically. The method based on transcriptional data from cell lines can be robustly applied to either new compounds or marketed drugs and has a great potential for drug discovery and drug repurposing.

Predict effective drug combination by deep believe network and Ontology Fingerprints

Guocai Chen[1], Alex Tsoi[2], Hua Xu[1], W. Jim Zheng[1]

[1]School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA
[2]Department of Dermatology and Biostatistics, University of Michigan, Ann Arbor, MI, USA

The synergistic effect of drug combination is one of the most desirable properties for treating cancer. However, systematically predicting effective drug combination is a significant challenge. We report here a novel method based on deep belief network to predict drug synergy from gene expression, pathway and the Ontology Fingerprints—a literature derived ontological profile of genes. Using data sets provided by 2016 DREAM competition, our analysis shows that this integrative method outperforms published results from the DREAM website, demonstrating the feasibility of predicting drug synergy from literature and the –omics data using advanced artificial intelligence approach.

Concurrent Session - Systems Biology
Monday, June 11, 2018
10:05 AM - 5:20 PM
Santa Monica

A Hidden Markov Model-based approach to reconstructing double minute chromosome amplicons

Ruslan T. Mardugalliamov[1], Kamal Al Nasr[1], Matthew Hayes[2]

[1]Department of Computer Science, Tennessee State University, Nashville, TN, USA
[2]Department of Physics and Computer Science, Xavier University of Louisiana,
New Orleans, Louisiana, USA E-mail: mhayes5@xula.edu

*Corresponding author

Double minute chromosomes (DMs) are circular fragments of extrachromosomal DNA. They are a mechanism for extreme gene amplification in the cells of some malignant tumors. Their existence strongly correlates with malignant tumor cell behavior and drug resistance. Locating DMs is important for informing precision therapy to cancer treatment. Furthermore, accurate detection of double minutes requires precise reconstruction of their amplicons, which are the highly-amplified gene-carrying contiguous segments that adjoin to form DMs. This work presents AmpliconFinder – a Hidden-Markov Model-based approach to detect DM amplicons. To assess its efficacy, AmpliconFinder was used to augment an earlier framework for DM detection (DMFinder), thus improving its robustness to noisy sequence data, and thus improving its sensitivity to detect DMs. Experiments on simulated genomic data have shown that augmenting DMFinder with AmpliconFinder significantly increased the sensitivity of DMFinder on these data. Moreover, DMFinder with AmpliconFinder found all previously reported DMs in three pediatric medulloblastoma datasets, whereas the original DMFinder framework found none.

Classifying Mild Traumatic Brain Injuries with Functional Network Analysis

F. Anthony San Lucas[1], John Redell[2], Dash Pramod[2,3], and Yin Liu[2,3,4*]

[1] Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, 1155 Pressler Street, Houston, Texas, USA

[2] Department of Neurobiology and Anatomy, University of Texas Health Science Center at Houston, 6431 Fannin Street, Houston, Texas, USA

[3] University of Texas Graduate School of Biomedical Science, 6767 Bertner Avenue, Houston, Texas, USA

[4]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Houston, Texas, USA

*Corresponding author

Background: Traumatic brain injury (TBI) represents a critical health problem of which timely diagnosis and treatment remain challenging. TBI is a result of an external force damaging brain tissue, accompanied by delayed pathogenic events which aggravate the injury. Molecular responses to different mild TBI subtypes have not been well characterized. TBI subtype classification is an important step towards the development

and application of novel treatments. The computational systems biology approach is proved to be a promising tool in biomarker discovery for central nervous system injury.

Results: In this study, we have performed a network-based analysis on gene expression profiles to identify functional gene subnetworks. The gene expression profiles were obtained from two experimental models of injury in rats: the controlled cortical impact and the fluid percussion injury. Our method integrates protein interaction information with gene expression profiles to identify biomarkers as subnetworks of genes. We have demonstrated that the selected gene subnetworks are more accurate to classify the heterogeneous responses to different injury models, compared to conventional analysis using individual marker genes selected without network information.

Conclusions: The systems approach can lead to a better understanding of the underlying complexities of the molecular responses after TBI and the identified subnetworks will have important prognostic functions for patients who sustain mild TBIs.

---

scDNA: a fast and comprehensive tool for single-cell differential network analysis

Yu-Chiao Chiu, Tzu-Hung Hsiao, Li-Ju Wang, Yidong Chen and Yu-hsuan Shao

Background

Single-cell RNA sequencing (scRNA-Seq) is an emerging technology that has revolutionized the research of the tumor heterogeneity. However, the highly sparse data matrices generated by the technology has posed an obstacle to the analysis of differential gene regulatory networks.

Results

Addressing the challenges, this study presents, as far as we know, the first bioinformatics tool for scRNA-Seq-based differential network analysis (scdNet). The tool features a gene set-based inference of per-cell states regarding a cellular function, sample size adjustment of gene-gene correlation, comparison of inter-state correlations, and construction of differential networks associated with the function. A simulation analysis demonstrated the power of scdNet in the analyses of sparse scRNA-Seq data matrices, with low requirement on the sample size, high computation efficiency, and tolerance of sequencing noises. Applying the tool to analyze a dataset of single circulating tumor cells (CTCs) of prostate cancer, we identified a close association between cell cycle phases to migration and other crucial functions, partially through the modulation of the IL17RC oncogene and other genes.

Conclusions

Overall, the tool is widely applicable to datasets generated by the emerging technology to bring biological insights into tumor heterogeneity. MATLAB implementation of scdNet is available at https://github.com/chiuyc/scdNet.

Multiple transcription factors contribute to inter-chromosomal interaction in yeast

Yulin Dai[1,2,3,#], Chao Li[2,3,#], Guangsheng Pei[1], Xiao Dong[2,3], Guohui Ding[2,4], Zhongming Zhao[1,5,6], Yixue Li[2,4,*], Peilin Jia[1,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2]Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Rd. Shanghai 200031, PR China

[3]Graduate School of Chinese Academy of Sciences, 19 Yuquan Rd. Beijing 100049, PR China

[4]Shanghai Center for Bioinformation Technology, 1278 Keyuan Rd. Shanghai 201203, PR China

[5]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[6]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

[#] These authors contribute equally to this work.

Chromatin interactions medicated by genomic elements located throughout the genome play important roles in gene regulation and can be identified with the technologies such as High-throughput Chromosome Conformation Capture (Hi-C), followed by next-generation sequencing. These techniques were wildly used to reveal the relative spatial disposition of chromatins in human, mouse and yeast. Unlike metazoan where CTCF plays major roles in mediating chromatin interactions, in yeast, the transcription factors (TFs) involved in this biological process are poorly known. Here, we presented two computational approaches to estimate the TFs enriched in the chromatin physical inter-chromosomal interactions in yeast. Through the Chi-square method, we found TFs whose binding data are differentially distributed in different interaction groups, including Cin5, Stp1 and Sut1, whose binding data are negatively correlated with the chromosome spatial distance. A multivariate linear regression model was employed to estimate the potential contribution of different transcription factors against the physical distance of chromosomes. Rlr1, Set12 and Dig1 were found to be top positively participated in these chromosomal interactions. Overall, we found 10 TFs enriched from both computational approaches. No TF was found to have a dominant impact on the inter- chromosomal interaction as CTCF did in human or other metazoan. In summary,

we presented a systematic examination of TFs involved in chromatin interaction in yeast and the results provide candidate TFs for future studies.

Metabolomics of Mammalian Brain Reveals Regional Differences

William T. Choi[#1-4], Mehmet Tosun[#4,5], Hyun-Hwan Jeong[#4,6], Cemal Karakas[4,5], Fatih Semerci[1,4], Zhandong Liu[*4,5,7], Mirjana Maletić-Savatić[*1,4,5,7,8]

[1]Program in Developmental Biology, Baylor College of Medicine, Houston, TX
[2]The National Library of Medicine Training Program in Biomedical Informatics, Houston, TX
[3]Medical Scientist Training Program, Baylor College of Medicine, Houston, TX
[4]Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX
[5]Department of Pediatrics-Neurology, Baylor College of Medicine, Houston, TX
[6]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX
[7]Quantitative Computational Biology Program, Baylor College of Medicine, Houston, TX
[8]Department of Neuroscience, Baylor College of Medicine, Houston, TX

# These authors contribute equally to this work.

*Corresponding author

The mammalian brain is organized into regions with specific biological functions and properties. These regions have distinct transcriptomes, but little is known whether they may also differ in their metabolome. The metabolome, a collection of small molecules or metabolites, is at the intersection of the genetic background of a given cell or tissue and the environmental influences that affect it. Thus, the metabolome directly reflects information about the physiologic state of a biological system under a particular condition. The objective of this study was to investigate whether various brain regions have diverse metabolome profiles, similarly to their genetic diversity. The answer to this question would suggest that not only the genome but also the metabolome may contribute to the functional diversity of brain regions. We investigated the metabolome of four regions of the mouse brain that have very distinct functions: frontal parenchyma, hippocampus, cerebellum, and olfactory bulb. We utilized gas- and liquid-chromatography mass spectrometry platforms and identified 215 metabolites. Principal component analysis, an unsupervised multivariate analysis, clustered each brain region based on its metabolome content, thus providing the unique metabolic profile of each region. A pathway-centric analysis indicated that olfactory bulb and cerebellum had most distinct metabolic profiles, while the cortical parenchyma and hippocampus were more similar in their metabolome content. Among the notable differences were distinct oxidative-anti-oxidative status and region-specific lipid profiles. Finally, a global metabolic connectivity analysis using the weighted correlation network analysis identified five hub metabolites that organized a unique metabolic network architecture within each examined brain region. These data indicate the diversity of global metabolome corresponding to specialized regional brain function and provide a new perspective on the underlying properties of brain regions.

Boosting Gene Expression Clustering with System-Wide Biological Information: A Robust Autoencoder Approach

Hongzhu Cui[1], Chong Zhou[2], Xinyu Dai[2], Yuting Liang[2], Randy Paffenroth[2,3,4], Dmitry Korkin[1,2,4*]

[1]Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA, USA 010609

[2]Data Science Program, Worcester Polytechnic Institute, Worcester, MA, USA 010609

[3]Mathematics Department, Worcester Polytechnic Institute, Worcester, MA, USA 010609

[4]Computer Science Department, Worcester Polytechnic Institute, Worcester, MA, USA 010609

*Corresponding author

Gene expression analysis provides genome-wide insights into the transcriptional activity of a cell. One of the first computational steps in exploration and analysis of the gene expression data is clustering. With a number of standard clustering methods routinely used, most of the methods do not take prior biological information into account. Here, we propose a new approach for gene expression clustering analysis. The approach benefits from a new deep learning architecture, Robust Autoencoder, which provides a more accurate high-level representation of the feature sets, and from incorporating prior system-wide biological information into the clustering process. We tested our approach on two gene expression datasets and compared the performance with two widely used clustering methods, hierarchical clustering and k-means, and with a recent deep learning clustering approach. Our approach outperformed all other clustering methods on the labeled yeast gene expression dataset. Furthermore, we showed that it is better in identifying the functionally common clusters than k-means on the unlabeled human gene expression dataset. The results demonstrate that our new deep learning architecture can generalize well the specific properties of gene expression profiles. Furthermore, the

results confirm our hypothesis that the prior biological network knowledge is helpful in the gene expression clustering.

Prediction of Protein Self-Interactions using Stacked Long Short-Term Memory from Protein Sequences Information

Yan-Bin Wang[1,#], Zhu-Hong You[1,#,*], Xiao Li[1,*], Tong-Hai Jiang[1], Li Cheng[1], Zhan-Heng Chen[1]

[1]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China. University of Chinese Academy of Sciences.

# These authors contributed equally to this work.

*Corresponding author

Background: Self-interacting Proteins (SIPs) plays a critical role in a series of life function in most living cells. Researches on SIPs are important part of molecular biology. Although numerous SIPs data be provided, traditional experimental methods are labor-intensive, time-consuming and costly and can only yield limited results in real-world needs. Hence，it's urgent to develop an efficient computational SIPs prediction method to fill the gap. Deep learning technologies have proven to produce subversive performance improvements in many areas, but the effectiveness of deep learning methods for SIPs prediction has not been verified.

Results: We developed a deep learning model for predicting SIPs by constructing a Stacked Long Short- Term Memory (SLSTM) neural network that contains “dropout”. We extracted features from protein sequences using a novel feature extraction scheme that combined Zernike Moments (ZMs) with Position Specific Weight Matrix (PSWM). The capability of the proposed approach was assessed on *S.erevisiae* and *Human* SIPs datasets. The result indicates that the approach based on deep learning can effectively resist data skew and achieve good accuracies of 95.69% and 97.88%, respectively. To demonstrate the progressiveness of deep learning, we compared the results of the SLSTM-based method and the celebrated Support Vector Machine (SVM) method and several other well-known methods on the same datasets.

Conclusion: The results show that our method is overall superior to any of the other existing state-of- the-art techniques. As far as we know, this study first applies deep learning method to predict SIPs, and practical experimental results reveal its potential in SIPs identification.

Circular RNA Expression Profiles during the Differentiation of Mouse Neural Stem Cells

Qichang Yang[1], Jing Wu[1], Jian Zhao[1], Tianyi Xu[1], Zhongming Zhao[2,3,*], Xiaofeng Song[1,*], Ping Han[4,*]
[1]Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, 211106, China
[2]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA
[4]The First Affiliated Hospital with Nanjing Medical University, Nanjing, Jiangsu, 210019, China

[*]Co-corresponding authors:

Background: Circular RNAs (circRNAs) have recently been found to be expressed in human brain tissue, and many lines of evidence indicate that circRNAs play regulatory roles in neurodevelopment. Proliferation and differentiation of neural stem cells (NSCs) are critical parts during development of central nervous system (CNS). To date, there have been no reports of circRNA expression profiles during the differentiation of mouse NSCs. We hypothesize that circRNAs may regulate gene expression in the proliferation and differentiation of NSCs.

Results: In this study, we obtained NSCs from the wild-type C57BL/6J mouse fetal cerebral cortex. We extracted total RNA from NSCs in different differentiation stages and then performed RNA-seq. By analyzing the RNA-Seq data, we found 37 circRNAs and 4182 mRNAs differentially expressed during the NSC differentiation. Gene Ontology (GO) enrichment analysis of the cognate linear genes of these circRNAs revealed that some enriched GO terms were related to neural activity. Furthermore, we performed a co-expression network analysis of these differentially expressed circRNAs and mRNAs. The result suggested a stronger GO enrichment in neural features for both the cognate linear genes of circRNAs and differentially expressed mRNAs.

Conclusion: We performed the first circRNA investigation during the differentiation of mouse NSCs. We found that 12 circRNAs might have regulatory roles during the NSC differentiation, indicating that circRNAs might be modulated during NSC differentiation. Our network analysis suggested the possible complex circRNA-mRNA mechanisms during differentiation, and future experimental work is need to validate these possible mechanisms.

Identification of Gene Signatures from RNA-seq Data Using Pareto-optimal Cluster Algorithm

Saurav Mallik[1], Zhongming Zhao[1,2*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

*Corresponding author

Background: Gene signatures are important to represent the molecular changes in the disease genomes or the cells in specific conditions, and have been often used to separate samples into different groups for better research or clinical treatment. While many methods and applications have been available in literature, there still lack powerful ones that can take account of the complex data and detect the most informative signatures.

Methods: In this article, we present a new framework for identifying gene signatures using Pareto-optimal cluster size identification for RNA-seq data. We first performed pre-filtering steps and normalization, then utilized the empirical Bayes test in Limma package to identify the differentially expressed genes (*DEGs*). Next, we used a multi-objective optimization technique, "Multi-objective optimization for collecting cluster alternatives" (MOCCA in R package) on these *DEGs* to find Pareto-optimal cluster size, and then applied k-means clustering to the RNA-seq data based on the optimal cluster size. The best cluster was obtained through computing the average Spearman's Correlation Score among all the genes in pair-wise manner belonging to the module. The best cluster is treated as the signature for the respective disease or cellular condition.

Results: We applied our framework to a cervical cancer RNA-seq dataset, which included 253 squamous cell carcinoma (SCC) samples and 22 adenocarcinoma (ADENO) samples. We identified a total of 582 *DEGs* by Limma analysis of SCC versus ADENO samples. Among them, 260 are up-regulated genes and 322 are down-regulated genes. Using MOCCA, we obtained seven Pareto-optimal clusters. The best cluster has a total of 35 *DEGs* consisting of all-upregulated genes. For validation, we ran PAMR (prediction analysis for microarrays) classifier on the selected best cluster, and assessed the classification performance. Our evaluation, measured by sensitivity, specificity, precision, and accuracy, showed high confidence.

Conclusions: Our framework identified a multi-objective based cluster that is treated as a signature that can classify the disease and control group of samples with higher classification performance (accuracy 0.935) for the corresponding disease. Our method is useful to find signature for any RNA-seq or microarray data.

An experimental design framework for Markovian gene regulatory networks under stationary control policy

Roozbeh Dehghannasiri[1*], Mohammad Shahrokh Esfahani[2], Edward R Dougherty[3,4]

[1]Department of Biochemistry, Stanford University, Stanford, CA 94305, USA.
[2]Division of Oncology, Stanford School of Medicine, Stanford, CA 94305, USA.
[3]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA.
[4]Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77845, USA.

*Corresponding author

Background: A fundamental problem for translational genomics is to find optimal therapies based on gene regulatory intervention. Dynamic intervention involves a control policy that optimally reduces a cost function based on phenotype by externally altering the state of the network over time. When a gene regulatory network (GRN) model is fully known, the problem is addressed using classical dynamic programming based on the Markov chain associated with the network. When the network is uncertain, a Bayesian framework can be applied, where policy optimality is with respect to both the dynamical objective and the uncertainty, as characterized by a prior distribution. In the presence of uncertainty, it is of great practical interest to develop an experimental design strategy and thereby select experiments that optimally reduce a measure of uncertainty.

Results: In this paper, we employ mean objective cost of uncertainty (MOCU), which quantifies uncertainty based on the degree to which uncertainty degrades the operational objective, that being the cost owing to undesirable phenotypes. We assume that a number of conditional probabilities characterizing regulatory relationships among genes are unknown in the Markovian GRN. In sum, there is a prior distribution which can be updated to a posterior distribution by observing a regulatory trajectory, and an optimal control policy, known as an "intrinsically Bayesian robust" (IBR) policy. To obtain a better IBR policy, we select an experiment that minimizes the MOCU remaining after applying its output to the network. At this point, we can either stop and find the resulting IBR policy or proceed to determine more unknown conditional probabilities via regulatory observation and find the IBR policy from the resulting posterior distribution. For sequential experimental design this entire process is iterated. Owing to the computational complexity of experimental design, which requires computation of many potential IBR policies, we implement an approximate method utilizing mean first passage times (MFPTs) – but only in experimental design, the final policy being an IBR policy.

Conclusions: Comprehensive performance analysis based on extensive simulations on synthetic and real GRNs demonstrate the efficacy of the proposed method, including the accuracy and computational advantage of the approximate MFPT-based design.

GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization

Hung-I Harry Chen[1,2], Yu-Chiao Chiu[2], Tinghe Zhang[1], Songyao Zhang[1,4], Yufei Huang[1,*], Yidong Chen[2,3,*]

[1]Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX 78249, USA
[2]Greehey Children's Cancer Research Institute, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA
[3]Department of Epidemiology & Biostatistics, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA
[4]Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, China

*Corresponding authors

Background: Bioinformatics tools have been developed to interpret gene expression data at the gene set level, and these gene set based analyses improve the biologists' capability to discover functional relevance of their experiment design. While elucidating gene set individually, inter gene sets association is rarely taken into consideration. Deep learning, an emerging machine learning technique in computational biology, can be used to generate an unbiased combination of gene set, and to determine the biological relevance and analysis consistency of these combining gene sets by leveraging large genomic data sets.

Results: In this study, we proposed a gene superset autoencoder (GSAE), a multi-layer autoencoder model with the incorporation of a priori defined gene sets that retain the crucial biological features in the latent layer. We introduced the concept of the gene superset, an unbiased combination of gene sets with weights trained by the autoencoder, where each node in the latent layer is a superset. Trained with genomic data from TCGA and evaluated with their accompanying clinical parameters, we showed gene supersets' ability of discriminating tumor subtypes and their prognostic capability. We further demonstrated the biological relevance of the top component gene sets in the significant supersets.

Conclusions: Using autoencoder model and gene superset at its latent layer, we demonstrated that gene supersets retain sufficient biological information with respect to

tumor subtypes and clinical prognostic significance. Superset also provides high reproducibility on survival analysis and accurate prediction for cancer subtypes.

Comparative gene co-expression network analysis of epithelial to mesenchymal transition reveals lung cancer progression stages

Daifeng Wang[1,2], John D. Haley[2,3], Patricia Thompson[2,3]

[1]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA
[2]Stony Brook Cancer Center, Stony Brook Medicine, Stony Brook, NY, USA
[3]Department of Pathology, Stony Brook Medicine, Stony Brook, NY, USA

The epithelial to mesenchymal transition (EMT) plays a key role in lung cancer progression and drug resistance. However, the dynamics and stability of gene expression patterns as cancer cells transition from E to M at a systems level and relevance to patient outcomes are unknown. To address this, using comparative network and clustering analysis, we systematically analyzed time-series gene expression data from lung cancer cell lines H358 and A549 that were induced to undergo EMT [1]. In particular, we predicted the putative regulatory networks controlling EMT expression dynamics, especially for the EMT-dynamic genes and related these patterns to patient outcomes using data from TCGA. Also, we validated the EMT hub regulatory genes using RNAi. From the network, we identified several novel genes distinct from the static states of E or M that exhibited temporal expression patterns or ‘periods’ during the EMT process that were shared in different lung cancer cell lines. For example, cell cycle and metabolic genes were found to be similarly down-regulated where immune-associated genes were up-regulated after middle EMT stages. The presence of EMT-dynamic gene expression patterns supports the presence of differential activation and repression timings at the transcriptional level for various pathways and functions during EMT that are not detected in pure E or M cells. Importantly, the cell line identified EMT-dynamic genes were found to be present in lung cancer patient tissues and associated with patient outcomes. In summary, our study suggests that in vitro identified EMT-dynamic genes capture elements of gene EMT expression dynamics at the patient level. Measurement of EMT dynamic genes, as opposed to E or M only, is potentially useful in future efforts aimed at classifying patient’s responses to treatments based on the EMT dynamics in the tissue.

Epigenomic Patterns Are Associated with Gene Haploinsufficiency and Predict Risk Genes of Developmental Disorders

Siying Chen[1], Xinwei Han[1], Yufeng Shen[1,*]

[1]Department of Systems Biology, Columbia University Medical Center, New York, NY

*Corresponding author

Haploinsufficiency is a major mechanism of genetic risk of human disease. Accurate prediction of haploinsufficient genes is essential for prioritizing and interpreting deleterious variants in genetic studies. Current methods based on mutation intolerance in population data suffer from inadequate power for genes with short transcripts or under modest selection. In this study we showed haploinsufficiency is strongly associated with regulatory complexity of gene expression measured by epigenomic profiles, and then developed a computational method (Episcore) to predict haploinsufficiency from epigenomic and gene expression data from a broad range of tissue and cell types using machine learning methods. Using data from recent exome sequencing studies of developmental disorders, Episcore achieved better performance in prioritizing loss of function de novo variants than current methods. We further showed that Episcore was complementary to mutation intolerance metrics for prioritizing loss of function variants. Our approach enables new applications of epigenomic and gene expression data and facilitates discovery of novel risk variants in studies of developmental disorders.

Genetic Association of Arterial Stiffness Index with Incident Coronary Artery Disease and Congestive Heart Failure

Seyedeh Zekavat, Mary Haas, Krishna Aragam, Connor Emdin, Amit Khera, Derek Klarin, Hongyu Zhao and Pradeep Natarajan

Despite current medical therapy, mortality from coronary artery disease (CAD) and congestive heart failure (CHF) remains high, making the discovery of orthogonal causal pathways particularly necessary. Arterial stiffness index (ASI) is a noninvasive, rapid measurement extracted from pulse oximeter waveforms derived from finger infrared analysis. ASI has been independently associated with cardiovascular disease risk in multiple epidemiological studies; however, it is unknown whether these associations represent causal relationships. While finger infrared analysis is used widespread clinically to measure oxygen saturation levels, the pulsatile waveforms derived from this technique are currently not clinically leveraged towards calculating arterial stiffness index. Determining whether arterial stiffness causally mediates cardiovascular diseases may help identify relevant biology and motivate novel prevention, monitoring, and therapeutic approaches that utilize this phenotype clinically.

Mendelian randomization uses human genetics to facilitate causal inference by leveraging the random assortment of genetic variants during meiosis at conception. This approach is therefore less susceptible to confounding or reverse causality which may occur with observational epidemiological analyses. Here, we used Mendelian randomization to

determine whether a genetic predisposition to increased arterial stiffness is associated with increased risk for incident CHF and CAD.

We performed the first genome-wide association analysis of arterial stiffness index in 131,686 participants from the UK Biobank. Genome-wide association analysis of ASI yielded two significant loci (P<5x10-8) at TEX41-ZEB2 and FOXO1, and three suggestive loci (P<5x10-7) at COL4A2-COL4A1, RNF126, and TCF20.

Using these variant-level results, we developed a 744-variant polygenic risk score that robustly associated with ASI (P<1x10-300, F-statistic=10,020), and tested its association with incident CHF and CAD through Mendelian randomization. Across nearly 400K participants of the UK Biobank and 200K participants of the CARDIOGRAMplusC4D consortium, we do not find evidence supporting a causal association with incident CAD (P>0.05). However, we do observe evidence supporting a causal association of ASI with incident CHF (HR=1.23 per SD increase in genetically-mediated ASI, P=5.7x10-3), independent of other cardiometabolic risk factors, including blood pressure. Furthermore, we find that the association with incident CHF is present only among individuals without hypertension (HR=1.52 per SD increase in genetically-mediated ASI, P=1.3x10-4) and not among those with hypertension (P>0.05) (Pinteraction=9.6x10-3).

These results are consistent with a causal association between ASI and CHF but not between ASI and CAD, and support ASI as an orthogonal clinical tool particularly for individuals without hypertension. Furthermore, these data suggest that reducing ASI, particularly among those without hypertension, may reduce risk for CHF.

Concurrent Sessions-Bioinformatics
Monday, June 11, 2018
10:05AM - 12:15 PM
Hollywood

A Graph-based Algorithm for Estimating Clonal Haplotypes of Tumor Sample from Sequencing Data

Yixuan Wang[1,3], Rong Zhang[2,3] , Xinyu Sun[2,3] , Yu Geng[1,3,4] , Jianye Liu[1,3] , Zhongmeng Zhao[1,3], Xuanping Zhang[1,3] , Yi Huang[1,3] , Jiayin Wang[*1,3]

[1]School of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an 710049, China
[2]School of Management, Xi’an Jiaotong University, Xi’an, Shaanxi 710049, China
[3]Shaanxi Engineering Research Center of Medical and Health Big Data, Xi’an Jiaotong University, Xi’an 710049, China
[4]Jinzhou Medical University, Jinzhou 121001, China

*Corresponding author

Tumor heterogeneity is an important characteristic of tumor tissue and is generally considered to be the result of evolution. It embodied with selective advantage of sub-clone presenting sub-clonal architecture inheritance. Identifying tumor heterogeneity is the basis for comprehensive understandings of the process of tumor evolution and has research and clinical implications. Most of the existing methods can only reconstruct genotype heterogeneity but cannot for haplotype heterogeneity. Even haplotype heterogeneity has greater value compared to genotype heterogeneity. Based on cancer sequencing data, this paper proposes an algorithm *MixSubHap* to reconstruct haplotype heterogeneity of tumor samples, estimates the number of sub-clones and provide the haplotypes of each clone into the following steps. Firstly, the number of sub-clones and the prior distribution of each sub-clone proportion are roughly obtained through clustering of variant allelic frequency (VAF). Secondly, an improved maximum spanning tree algorithm is used to extract at least two paired-end reads. Under the evolution rules for sub-clone, iterative stripping is conducted to reflect the ratio of each sub-clone's read depth till all variants contain. Thirdly, all the depth peeling reads are assembled to reconstruct the tumor clone haplotype. For a variation site with sufficient sequence depth, *MixSubHap* which based on convolution inverse partition stitching algorithm maximizes the depth stripping ratio of the inverse convolution and then tests probability. Through dividing and assembling the variation sites, the conflicts in the assembly drawing are resolved. The experimental results show that *MixSubHap* is suitable for the second and third generation sequencing data. When the sequencing depth is insufficient, this algorithm achieves about 95% accuracy; when the sequencing depth is more than 50×, the corrected recognition accuracy is generally close to 99%. Compared with HapCompass and other similar algorithms, *MixSubHap* has greatly improved in accuracy, computing speed and has greatly reduced requirement of the sequencing depth.

A new insight into underlying disease mechanism through semi-parametric latent differential network model

Yong He[1], Jiadong Ji[1*], Lei Xie[2,3], Xinsheng Zhang[4], Fuzhong Xue[5]

[1]School of Statistics, Shandong University of Finance and Economics, 250014 Jinan, China.
[2]Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, 10016 New York, USA.
[3]Department of Computer Science, Hunter College, The City University of New York, 10065 New York, USA.
[4]School of Management, Fudan University, 200433 Shanghai, China.
[5]School of Public Health, Shandong University, 250012 Jinan, China

*Corresponding author

Background: In genomic studies, to investigate how the structure of a genetic network differs between two experiment conditions is a very interesting but challenging problem, especially in high-dimensional setting. Existing literatures mostly focus on differential network modelling for continuous data. However, in real application, we may encounter discrete data or mixed data, which urges us to propose a unified differential network modelling for various data types.

Results: We propose a unified latent Gaussian copula differential network model which provides deeper understanding of the unknown mechanism than that among the observed variables. Adaptive rank-based estimation approaches are proposed with the assumption that the true differential network is sparse. The adaptive estimation approaches do not require precision matrices to be sparse, and thus can allow the individual networks to contain hub nodes. Theoretical analysis shows that the proposed methods achieve the same parametric convergence rate for both the difference of the precision matrices estimation and differential structure recovery, which means that the extra modeling flexibility comes at almost no cost of statistical efficiency. Besides theoretical analysis, thorough numerical simulations are conducted to compare the empirical performance of the proposed methods with some other state-of-the-art methods. The result shows that the proposed methods work quite well for various data types . The proposed method is then applied on gene expression data associated with lung cancer to illustrate its empirical usefulness.

Conclusions: The proposed latent variable differential network models allows for various data-types and thus are more flexible, which also provide deeper understanding of the unknown mechanism than that among the observed variables. Theoretical analysis, numerical simulation and real application all demonstrate the great advantages of the latent differential network modelling and thus are highly recommended.

Genetic-Epigenetic Interactions in Asthma Revealed by a Genome-Wide Gene-Centric Search

Vladimir Kogan[1#], Joshua Millstein[1#*], Stephanie J. London[2], Carole Ober[3], Steven R. White[4], Edward T. Naureckas[4], W. James Gauderman[1], Daniel J. Jackson[5], Albino Barraza-Villarreal[6], Isabelle Romieu[7], Benjamin A. Raby[8], Carrie V. Breton[1]

[1]Department of Preventive Medicine, Keck School of Medicine of the University of Southern California.

[2]Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Dept. of Health and Human Services, RTP, NC 27709.

[3]Department of Human Genetics, University of Chicago.

[4]Department of Medicine, University of Chicago.

[5]University of Wisconsin School of Medicine and Public Health.

[6]National Institute of Public Health of Mexico.

[7]International Agency for Research on Cancer, Section of Nutrition and Metabolism, Lyon, France.
[8]Brigham and Women's Hospital, Department of Medicine, Channing Division of Network Medicine.

#These two authors contributed equally to this work

*Corresponding author

Objectives: There is evidence to suggest that asthma pathogenesis is affected by both genetic and epigenetic variation independently, and there is some evidence to suggest genetic-epigenetic interactions affect risk of asthma. However, little research has been done to identify such interactions on a genome-wide scale. The aim of this studies was to identify genes with genetic-epigenetic interactions associated with asthma.
Methods: Using asthma case-control data, we applied a novel nonparametric gene-centric approach to test for interactions between multiple SNPs and CpG sites simultaneously in the vicinities of 18,178 genes across the genome.
Results: Twelve genes, PF4, ATF3, TPRA1, HOPX, SCARNA18, STC1, OR10K1, UPK1B, LOC101928523, LHX6, CHMP4B, and LANCL1, exhibited statistically significant SNP-CpG interactions (FDR = 0.05). Of these, three have previously been implicated in asthma risk, PF4, ATF3, and TPRA1. Follow-up analysis revealed statistically significant pairwise SNP-CpG interactions for several of these genes, including SCARNA18, LHX6, and LOC101928523, (P-Values = (1.33E-04, 8.21E-04, 1.11E-03), respectively).
Conclusions: Joint effects of genetic and epigenetic variation may play an important role in asthma pathogenesis. Statistical methods that simultaneously account for multiple variations across chromosomal regions may be needed to detect these types of effects on a genome-wide scale.

DLAD4U: deriving and prioritizing disease lists from PubMed literature

Junhui Shen[1], Suhas Vasaikar[2,3], Bing Zhang[2,3*]

[1]Information Center, Beijing University of Chinese Medicine, Beijing, China
[2]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas, USA
[3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA

*Corresponding author

Background: Due to recent technology advancements, disease related knowledge is growing rapidly. It becomes nontrivial to go through all published literature to identify associations between human diseases and genetic, environmental, and life style factors, disease symptoms, and treatment strategies. Here we report DLAD4U (Disease List Automatically Derived For You), an efficient, accurate and easy-to-use disease search engine based on PubMed literature.

Results: DLAD4U uses the eSearch and eFetch APIs from the National Center for Biotechnology Information (NCBI) to find publications related to a query and to identify diseases from the retrieved publications. The hypergeometric test was used to prioritize identified diseases for displaying to users. DLAD4U accepts any valid queries for PubMed, and the output results include a ranked disease list, information associated with each disease, chronologically-ordered supporting publications, a summary of the run, and links for file export. DLAD4U outperformed other disease search engines in our comparative evaluation using selected genes and drugs as query terms and manually curated data as “gold standard”. For 100 genes that are associated with only one disease in the gold standard, the Mean Average Precision (MAP) measure from DLAD4U was 0.77, which clearly outperformed other tools. For 10 genes that are associated with multiple diseases in the gold standard, the mean precision, recall and F-measure scores from DLAD4U were always higher than those from other tools. The superior performance of DLAD4U was further confirmed using 100 drugs as queries, with an MAP of 0.90.

Conclusions: DLAD4U is a new, intuitive disease search engine that takes advantage of existing resources at NCBI to provide computational efficiency and uses statistical analyses to ensure accuracy. DLAD4U is publicly available at http://dlad4u.zhang-lab.org/.

A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records

Shanta Chowdhury[1], Xishuang Dong[1], Lijun Qian[1], Xiangfang Li[1*], Yi Guan[2], Jinfeng Yang[3], Qiubin Yu[4]

[1]Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT), Department of Electrical and Computer Engineering, Prairie View A&M University, Texas A&M University System, Prairie View, Texas 77446, USA.
[2]Schools of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.
[3]Schools of Software, Harbin University of Science and Technology, Harbin, China.
[4]Second Affiliated Hospital of Harbin Medical University, Harbin, China.

*Corresponding author

Background: Electronic Medical Record (EMR) comprises patients’ medical information gathered by medical stuff for providing better health care. Named Entity Recognition (NER) is a sub-field of information extraction aimed at identifying specific

entity terms such as disease, test, symptom, genes etc. NER can be a relief for healthcare providers and medical specialists to extract useful information automatically and avoid unnecessary and unrelated information in EMR. However, limited resources of available EMR pose a great challenge for mining entity terms. Therefore, a multitask bi-directional RNN model is proposed here as a potential solution of data augmentation to enhance NER performance with limited data.

Methods: A multitask bi-directional RNN model is proposed for extracting entity terms from Chinese EMR. The proposed model can be divided into a shared layer and a task specific layer. Firstly, vector representation of each word is obtained as a concatenation of word embedding and character embedding. Then

Bi-directional RNN is used to extract context information from sentence. After that, all these layers are shared by two different task layers, namely the

parts-of-speech tagging task layer and the named entity recognition task layer. These two tasks layers are trained alternatively so that the knowledge learned from named entity recognition task can be enhanced by the knowledge gained from parts-of-speech tagging task.

Results: The performance of our proposed model has been evaluated in terms of micro average F-score, macro average F-score and accuracy. It is observed that the proposed model outperforms the baseline model in all cases. For instance, the micro average F-score and the macro average F-score are improved by 2.41% and 4.16%, respectively, and the overall accuracy is improved by 5.66%.

Conclusions: In this paper, a novel multitask bi-directional RNN model is proposed for improving the performance of named entity recognition in EMR. Evaluation results using real datasets demonstrate the effectiveness of the proposed model.

iMEGES: integrated Mental-disorder GEnome score for prioritizing the susceptibility genes for mental disorders in personal genomes

Atlas Khan[1], Qian Liu[2], Kai Wang[2,3*]

[1] Division of Nephrology, Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY, 10032, USA

[2] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[3] Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

*Corresponding author

Background: Many rare and common genetic variants, including SNVs (single nucleotide variants) and SVs (structural variants), are associated with mental disorders, yet more remain to be discovered. Powerful integrative methods are needed to systematically prioritize variants and genes that confer susceptibility to mental disorders

in genomes of individual patients and to facilitate the development of personalized treatment or therapeutic approaches.

Methods: Leveraging deep neural network on the TensorFlow framework, we developed a computational tool, integrated Mental-disorder GEnome Score (iMEGES), to analyze whole genome/exome sequencing data on personal genomes. iMEGES takes as the input whole-genome variants and clinical phenotype terms of an individual with mental disorders, and then integrates contributions from coding, non-coding, SVs, known brain expression quantitative trait locus (eQTLs), and epigenetic information from PsychENCODE, to prioritize a list of variants and genes that may be of relevance to the phenotypes.

Results: iMEGES was evaluated on multiple datasets of mental disorders, and it achieved improved performance than competing approaches when large training data is available.

Conclusion: iMEGES can be used in population studies to help prioritize novel genes or variants that are associated with disease susceptibility, and also on individual patients to help identify genes or variants with major effect sizes for mental disorders.

Concurrent Sessions-Medical Informatics
Monday, June 11, 2018
3:00 PM - 5:20 PM
Hollywood

Building a high performance computing infrastructure for cancer research

W Jim Zheng

School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA, 77030

The causes of cancer are now studied extensively using the latest technologies that can generate huge amounts of data. These technologies have shifted the focus of cancer research from data generation to data analysis, demanding novel ways to sift through the sea of cancer research data to find cures. As the only free-standing biomedical informatics school in the nation, the faculty, trainees, and research staff at the School of Biomedical Informatics (SBMI) are at the forefront of developing these computational methods to analyze cancer research data. However, many of these methods are the results of academic research and not readily available to the cancer research community. To address this gap, we propose a Data Science and Informatics Core for Cancer Research (DSICCR) to translate the cutting-edge data science and informatics research at SBMI to easily accessible, high-quality, and user-friendly software and services to advance cancer research. The DSICCR will build a “big data” infrastructure for cancer research, provide data science and informatics services to cancer researchers, and educate cancer researchers about the latest data science and informatics methods and their application in cancer research. DSICCR will significantly advance cancer research through the application of cutting-edge data science, and thereby help to find cures for cancer and reduce cancer deaths in Texas.

Dynamic Prediction of Hospital Admission with Medical Claim Data

Tianzhong Yang[1,2,#], Yang Yang[1,#,*], Yugang Jia[1], Xiao Li[1,2]

[1]Philips Research North America, Cambridge, Massachusetts, 02141, United States

[2]Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, Houston, Texas, USA, 77030

[#]Contributed equally
[*]Co-corresponding authors:

Background: Congestive heart failure is one of the most common reasons those aged 65 and over are hospitalized in the United States, which has caused a considerable economic burden. The precise prediction of hospitalization caused by congestive heart failure in the near future could prevent possible hospitalization, optimize the medical resources and better meet the healthcare needs of patients.

Methods: To fully utilize the monthly-updated claim feed data released by The Centers for Medicare and Medicaid Services (CMS), we present a dynamic random survival forest model adapted for periodically updated data to predict the risk of adverse events. We apply our model to dynamically predict the risk of hospital admission among patients with congestive heart failure identified using the Accountable Care Organization Operational System Claim and Claim Line Feed data from Feb 2014 to Sep 2015. We benchmark the proposed model with two commonly used models in medical application literature: the cox proportional model and logistic regression model with L-1 norm penalty.

Results: Results show that our model has high Area-Under-the-ROC-Curve across time points and C-statistics. In addition to the high performance, it provides a measure of variable importance and individual-level instant risk. Conclusion: We present an efficient model adapted for periodically updated data such as the monthly updated claim feed data released by CMS to predict the risk of hospitalization. In addition to processing big-volume periodically updated stream-like data, our model can capture event onset information and time-to-event information, incorporate time-varying features, provide insights of variable importance and have good prediction power. To the best of our

knowledge, it is the first work combining sliding window technique with random survival forest. The model achieves remarkable performance and it could easily be deployed to monitor patients in real time.

Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements

Lindsay P. Zimmerman[1], Paul A. Reyfman[1], Angela D. R. Smith[1], Zexian Zeng[1], Abel Kho[1], L. Nelson Sanchez-Pinto[1], Yuan Luo[1*]

[1]Northwestern University, Evanston, Illinois, United States of America

*Corresponding author

Background: The development of acute kidney injury (AKI) during an intensive care unit (ICU) admission is associated with increased morbidity and mortality.
Methods: Our objective was to develop and validate a data driven multivariable clinical predictive model for early detection of AKI among a large cohort of adult critical care patients. We utilized data form the Medical Information Mart for Intensive Care III (MIMIC-III) for all patients who had a creatinine measured for 3 days following ICU admission and excluded patients with pre-existing condition of Chronic Kidney Disease and Acute Kidney Injury on admission. Data extracted included patient age, gender, ethnicity, creatinine, other vital signs and lab values during the first day of ICU admission, whether the patient was mechanically ventilated during the first day of ICU admission, and the hourly rate of urine output during the first day of ICU admission.
Results: Utilizing the demographics, the clinical data and the laboratory test measurements from Day 1 of ICU admission, we accurately predicted max serum creatinine level during Day 2 and Day 3 with a root mean square error of 0.224 mg/dL. We demonstrated that using machine learning models (multivariate logistic regression, random forest and artificial neural networks) with demographics and physiologic features can predict AKI onset as defined by the current clinical guideline with a competitive AUC (mean AUC 0.783 by our all-feature, logistic-regression model), while previous models aimed at more specific patient cohorts.
Conclusions: Experimental results suggest that our model has the potential to assist clinicians in identifying patients at greater risk of new onset of AKI in critical care setting. Prospective trials with independent model training and external validation cohorts are needed to further evaluate the clinical utility of this approach and potentially instituting interventions to decrease the likelihood of developing AKI.

Epileptic foci localization based on mapping the synchronization of dynamic brain network

Tian Mei[1,2], Xiaoyan Wei[1], Ziyi Chen[3], Xianghua Tian[4], Nan Dong[1], Dongmei Li[5] ,Yi Zhou[1*],

[1]Department of Biomedical Engineering, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong Province, China
[2]Department of Information, Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, 510655, Guangdong Province, China
[3]Department of Neurology, First Affiliated Hospital of Sun Yat-sen University, Guangzhou, 510080, Guangdong Province, China
[4]Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, 830011, Xinjiang Uygur Autonomous Region, China
[5]College of Public Health, Xinjiang Medical University, Urumqi 830011, Xinjiang Uygur Autonomous Region, China

*Corresponding author

Background:
Characterizing the synchronous changes of epileptic seizures in different stages between different regions is profound to understand the transmission pathways of epileptic brain network and epileptogenic foci. There is currently no adequate quantitative calculation method for describing the propagation pathways of EEG signals in the brain network from short and long term. The goal of this study is to explore the innovative method to locate epileptic foci, mapping synchronization in the brain networks based on electroencephalogram.

Methods:
Mutual information was used to analyse the short-term synchronization in the full electrodes; while nonlinear dynamics quantifies the statistical independencies in the long –term among all electrodes. Then graph theory based on complex network was employed to construct dynamic brain network for epilepsy patients when they were awake, asleep and in seizure, anylysing the changing topology indexes.

Results:
Epileptic network achieved a high degree of nonlinear synchronization compared to awake time. and the main path of epileptiform activity was revealed by searching core nodes. The core nodes of brain network were in connection with the onset zone. Seizures always happened with a high degree distribution.

Conclusion:

This study indicated the path of EEG synchronous propagation in seizures, and core nodes could locate the epileptic foci accurately in some epileptic patents .

Gene Fingerprint model for Literature based detection of the associations among complex diseases: A case study of COPD

Guocai Chen[1,#], Yuxi Jia[1,2,#], Lisha Zhu[1], Ping Li[3], Lin Zhang[4], Cui Tao[1,*], W. Jim Zheng[1,*]

[1] The University of Texas School of Biomedical Informatics, 7000 Fannin St Suite 600, Houston, TX 77030, United States

[2] Department of Medical Informatics, School of Public Health, Jilin University, Changchun, Jilin 130021, China

[3] Department of Development Pediatrics, The Second Affiliated Hospital of Jilin University, Changchun, Jilin 130041, China.

[4] Department of Respiratory Medicine, The Second Affiliated Hospital of Jilin University, Changchun, Jilin 130041, China.

*Corresponding author

Understanding disease to disease relationships may significantly improve the treatment of diseases and uncover novel uses for existing drugs. In this paper, we introduce a mathematic model to quantitatively measure the associations between genes and diseases. We applied this approach to analyze the relationships between Chronic Obstructive Pulmonary Disease (COPD) and other diseases under the Lung diseases branch in the Medical subject heading index system and detected 4 novel diseases relevant to COPD. As judged by domain experts, the F score of our approach is up to 77.6%.

Integrating Sentence Sequence Representation and Shortest Dependency Path into a Deep Learning Framework for Relation Extraction in Clinical Text

Zhiheng Li[1], Zhihao Yang[1], Chen Shen[1], Jun Xu[2], Yaoyun Zhang[2], Hua Xu[2*]

[1]College of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning, China,

[2] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

*Corresponding author

Extracting relations between important clinical entities is critical but very challenging for natural language processing (NLP) in the medical domain. Researchers have applied deep learning-based approaches to clinical relation extraction; but most of the existing deep learning models consider sentence sequence only, without modeling syntactic structures. In this paper, we propose a novel neural approach to model Shortest Dependency Path (SDP) between target entities together with the sentence sequence for clinical relation extraction. Using the 2010 i2b2/VA relation extraction dataset, we compared our approach with other baseline methods. Our experimental results show that the proposed approach achieved significant improvements over comparable existing methods, demonstrating the effectiveness of utilizing syntactic structures in deep learning-based relation extraction.

Concurrent Session - Cancer Genomics
Tuesday, June 12, 2018
10:05 AM - 4:30 PM
Grand Imperial North

Identification of exon skipping events associated with Alzheimer's disease in the human hippocampus

Seonggyun Han[1], Jason E. Miller[2], Seyoun Byun[1], Dokyoon Kim[2,3], Shannon L. Risacher[4], Andrew J Saykin[4,5], Younghee Lee[1*], Kwangsik Nho[4,5*], for Alzheimer's Disease Neuroimaging Initiative[**]

[1]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA;
[2]Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA USA;
[3]The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA;
[4]Department of Radiology and Imaging Sciences and Indiana Alzheimer Disease Center, Indiana University School of Medicine, Indianapolis, IN, USA;
[5]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA.

*Corresponding author

Background: At least 90% of human genes are alternatively spliced. Alternative splicing has an important function regulating gene expression and miss-splicing can contribute to risk for human diseases, including Alzheimer's disease (AD).

Methods: We developed a splicing decision model as a molecular mechanism to identify functional exon skipping events and genetic variation affecting alternative splicing on a genome-wide scale by integrating genomics, transcriptomics, and neuroimaging data in a systems biology approach. In this study, we analyzed RNA-Seq data of hippocampus brain tissue from Alzheimer's disease (AD; n=24) and cognitively normal elderly controls (CN; n=50) and identified three exon skipping events in two genes (RELN and NOS1) as significantly associated with AD (corrected p-value < 0.05 and fold change > 1.5). Next, we identified single-nucleotide polymorphisms (SNPs) affecting exon skipping events using the splicing decision model and then performed an association analysis of SNPs potentially affecting three exon skipping events with a global cortical measure of amyloid-β deposition measured by [18F] Florbetapir position emission tomography (PET) scan as an AD-related quantitative phenotype. A whole-brain voxel-based analysis was also performed.

Results: Two exons in RELN and one exon in NOS1 showed significantly lower expression levels in the AD participants compared to CN participants, suggesting that the exons tend to be skipped more in AD. We also showed the loss of the core protein structure due to the skipped exons using the protein 3D structure analysis. The targeted SNP-based association analysis identified one intronic SNP (rs362771) adjacent to the skipped exon 24 in RELN as significantly associated with cortical amyloid-β levels (corrected p-value < 0.05). This SNP is within the splicing regulatory element, i.e., intronic splicing enhancer. The minor allele of rs362771 conferred decreases in cortical amyloid-β levels in the right temporal and bilateral parietal lobes.

Conclusions: Our results suggest that exon skipping events and splicing-affecting SNPs in the human hippocampus may contribute to AD pathogenesis. Integration of multiple omics and neuroimaging data provides insights into possible mechanisms underlying AD pathophysiology through exon skipping and may help identify novel therapeutic targets.

Brain-wide structural connectivity alterations under the control of Alzheimer risk genes

Jingwen Yan[1*], Vinesh Raja V[1], Zhi Huang[2], Enrico Amico[3], Kwangsik Nho[4], Shiaofeng Fang[5], Olaf Sporns[6], Yu-chien Wu[4], Andrew Saykin[4], Joaquin Goni[3], Li Shen[7], For the Alzheimer's Disease Neuroimaging Initiative

[1]Department of BioHealth Informatics, Indiana University Purdue University Indianapolis, 719 Indiana Ave, Indianapolis, USA.

[2]Electrical and Computing Engineering, Indiana University Purdue University Indianapolis, 46202 Indianapolis, USA.

[3]Industrial Engineering,Purdue University, 47096 West Lafayette, USA.

[4]Radiology and Imaging Sciences, Indiana University School of Medicine, 46202 Indianapolis, USA.

[5]Computer Science, Indiana University Purdue University Indianapolis, 46202 Indianapolis, USA.
[6]Psychological and Brain Sciences, Indiana University, 47405 Bloomington, USA.
[7]Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 19104 Philadelphia, USA.

*Corresponding author

Background: Alzheimer's disease is the most common form of brain dementia characterized by gradual loss of memory followed by further deterioration of other cognitive function. Large-scale genome-wide association studies have identified and validated more than 20 AD risk genes. However, how these genes are related to the brain-wide breakdown of structural connectivity in AD patients remains unknown.
Methods: We used the genotype and DTI data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. After constructing the brain network for each subject, we extracted three types of link measures, including fiber anisotropy, fiber length and density. We then performed a targeted genetic association analysis of brain-wide connectivity measures using general linear regression models. Age at scan and gender were included in the regression model as covariates. For fair comparison of the genetic effect on different measures, fiber anisotropy, fiber length and density were all normalized with mean as 0 and standard deviation as one. We aim to discover the abnormal brain-wide network alterations under the control of 34 AD risk SNPs identified in previous large-scale genome-wide association studies.
Results: After enforcing the stringent Bonferroni correction, rs10498633 in *SLC24A4* were found to significantly associated with anisotropy, total number and length of fibers, including some connecting brain hemispheres. With a lower level of significance at 5e-6, we observed significant genetic effect of SNPs in *APOE, ABCA7, EPHA1* and *CASS4* on various brain connectivity measures.

Context-sensitive Network Analysis Identifies Food Metabolites Associated with Alzheimer's Disease: An Exploratory Study

Yang Chen, Rong Xu

Department of Population and Quantitative Health Science, School of Medicine, Case Western Reserve University, Cleveland, Ohio, 44106, USA

Background: Diet plays an important role in Alzheimer's disease (AD) initiation, progression and outcomes. Previous studies have shown individual food-derived substances may have neuroprotective or neurotoxic effects. However, few works systematically investigate the role of food and food-derived metabolites on the development and progression of AD.

Methods: In this study, we systematically investigated 7,569 metabolites and identified AD-associated food metabolites using a novel network-based approach. We constructed a context-sensitive network to integrate heterogeneous chemical and genetic data, and to model context-specific inter-relationships among foods, metabolites, human genes and AD.

Results: Our metabolite prioritization algorithm ranked 59 known AD-associated food metabolites within top 4.9%, which is significantly higher than random expectation. Interestingly, a few top-ranked food metabolites were specifically enriched in herbs and spices. Pathway enrichment analysis shows that these top-ranked herb-and- spice metabolites share many common pathways with AD, including the amyloid processing pathway, which is considered as a hallmark in AD-affected brains and has pathological roles in AD development.

Conclusions: Our study represents the first unbiased systems approach to characterizing the effects of food and food-derived metabolites in AD pathogenesis. Our ranking approach prioritizes the known AD-associated food metabolites, and identifies interesting relationships between AD and the food group "herbs and spices". Overall, our study provides intriguing evidence for the role of diet, as an important environmental factor, in AD etiology.

Using natural language processing and machine learning to identify breast cancer local recurrence

Zexian Zeng[1], Sasa Espino[2], Ankita Roy[2], Xiaoyu Li[3], Seema Khan[2], Susan Clare[2], Xia Jiang[4], Richard Neapolitan[1], Yuan Luo1*

1Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, 60611, USA
2Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, 60611, USA
3Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, 02115, USA
4Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, 15213, USA

*Corresponding author

Background: Identifying breast cancer local recurrence in clinical data sets is important for clinical research and practice. Developing a model using natural language processing and machine learning to identify local recurrences in breast cancer patients can reduce the time-consuming work in manually chart review.

Methods: We design a novel concept-based filter and a prediction model to detect local recurrences using EHRs. In the training dataset, we manually review a development corpus of 50 progress notes and extract partial sentences that indicate breast cancer local

recurrence. We process these partial sentences to obtain a positive set of concepts using MetaMap. We apply MetaMap on patients' progress notes and only retain concepts that fall within the positive concept set. These features together with number of pathology reports are used to train a support vector machine to identify local recurrence.
Results: We compared our model with three baseline classifiers using either full MetaMap concepts, filtered MetaMap concepts, or bag of words. Our model achieved the best AUC (0.93 in cross-validation, 0.87 in held-out testing).
Conclusion: Compared to a labor-intensive chart review, our model provides an automated way to identify breast cancer local recurrences. We expect that by minimally adapting the positive concept set, this study has the potential to be replicated at other institutions with a moderate sized training dataset.

Identification of long non-coding RNA-related and – coexpressed mRNA biomarkers for hepatocellular carcinoma

Fan Zhang[1,2*], Linda Ding[3], Li Cui[4], Robert Barber[5], Bin Deng[1,2]

[1]Vermont Genetics Network, University of Vermont, Burlington, Vermont 05405 USA
[2]Department of Biology, University of Vermont, Burlington, Vermont 05405 USA
[3]School of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0606, USA
[4]Department of Neurosciences, School of Medicine, University of California, San Diego, 9500 Gilman Drive #0949, La Jolla, CA 92093, USA
[5]Department of Pharmacology and Neuroscience, University of North Texas Health Science Center, Fort Worth, USA

*Corresponding authors

Background: While changes in mRNA expression during tumorigenesis have been used widely as molecular biomarkers for the diagnosis of a number of cancers, the approach has limitations. For example, traditional methods do not consider the regulatory and positional relationship between mRNA and lncRNA. The latter has been largely shown to possess tumor suppressive or oncogenic properties. The combined analysis of mRNA and lncRNA is likely to facilitate the identification of biomarkers with higher confidence.
Results: Therefore, we have developed an lncRNA-related method to identify traditional mRNA biomarkers. First we identified mRNAs that are differentially expressed in Hepatocellular Carcinoma (HCC) by comparing cancer and matched adjacent non-tumorous liver tissues. Then, we performed mRNA-lncRNA relationship and coexpression analysis and obtained 41 lncRNA-related and -coexpressed mRNA biomarkers. Next, we performed network analysis, gene ontology analysis and pathway analysis to unravel the functional roles and molecular mechanisms of these lncRNA-related and -coexpressed mRNA biomarkers. Finally, we validated the prediction and

performance of the 41 lncRNA-related and -coexpressed mRNA biomarkers using Support Vector Machine model with five-fold cross-validation in an independent HCC dataset from RNA-seq.
Conclusions: Our results suggested that mRNAs expression profiles coexpressed with positionally related lncRNAs can provide important insights into early diagnosis and specific targeted gene therapy of HCC.

Comparison of different functional prediction scores using a gene- based permutation model for identifying cancer driver genes

Alice Djotsa Nono[1] and Xiaoming Liu[1]*

[1]Human Genetics Center, UTHealth School of Public Health, Houston, TX

*Corresponding author

Background: Identifying cancer "driver" genes (CDG) is a crucial step in cancer genomic toward the advancement of precision medicine. However, driver gene discovery is a very challenging task because we are not only dealing with huge amount of data; but we are also faced with the complexity of the disease including the heterogeneity of background somatic mutation rate in each cancer patient. It is generally accepted that CDG harbor variants conferring growth advantage in the malignant cell and they are positively selected, which are critical to cancer development; whereas, non-driver genes harbor random mutations with no functional consequence on cancer. Based on this fact, function prediction based approaches for identifying CDG have been proposed to interrogate the distribution of functional predictions among mutations in cancer genomes [1]. Assuming most of the observed mutations are passenger mutations and given the quantitative predictions for the functional impact of the mutations, genes enriched of functional or deleterious mutations are more likely to be drivers. The promises of these methods have been continually refined and can therefore be applied to increase accuracy in detecting new candidate CDGs. However, current function prediction based approaches only focus on coding mutations and lack a systematic way to pick the best mutation deleteriousness prediction algorithms for usage.
Results: In this study, we propose a new function prediction based approach to discover CDGs through a gene-based permutation approach. Our method not only covers both coding and non-coding regions of the genes; but it also accounts for the heterogeneous mutational context in cohort of cancer patients. The permutation model was implemented independently using seven popular deleteriousness prediction scores covering splicing regions (SPIDEX), coding regions (MetaLR, and VEST3) and pan-genome (CADD, DANN, Fathmm-MKL coding and Fathmm-MKL noncoding). We applied this new approach to somatic single nucleotide variants (SNVs) from whole-genome sequences of

119 breast and 24 lung cancer patients and compared the seven deleteriousness prediction scores for their performance in this study.

Conclusion: The new function prediction based approach not only predicted known cancer genes listed in the Cancer Gene Census (COSMIC database), but also new candidate CDGs that are worth further investigation. The results showed the advantage of utilizing pan-genome deleteriousness prediction scores in function prediction based methods. Although VEST3 score, a deleteriousness prediction score for missense mutations, has the best performance in breast cancer, it was topped by CADD and Fathmm-MKL coding, two pan-genome deleteriousness prediction scores, in lung cancer.

Concurrent Session - Cancer Genomics
Tuesday, June 12, 2018
2:20 PM - 4:30 PM
Santa Monica

Integrating proteomic and phosphoproteomic data for pathway analysis in breast cancer

Jie Ren[1], Bo Wang[1], Jing Li[1*]

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240 China

*Corresponding author

Background: As protein is the basic unit of cell function and biological pathway, shotgun proteomics, the large-scale analysis of proteins, is contributing greatly to our understanding of disease mechanisms. Proteomics study could detect the changes of both protein expression and modification. With the releases of large-scale cancer proteome studies, how to integrate acquired proteomic and phosphoproteomic data in more comprehensive pathway analysis becomes implemented, but remains challenging. Integrative pathway analysis at proteome level provides a systematic insight into the signaling network adaptations in the development of cancer.

Results: Here we integrated proteomic and phosphoproteomic datato perform pathway prioritization in breast cancer. We manually collected and curated breast cancer well-known related pathways from the literature as target pathways (TPs) or positive control in method evaluation. Three different strategies including Hypergeometric test based over-representation analysis, Kolmogorov-Smirnov (K-S) test based gene set analysis and topology-based pathway analysis, were applied and evaluated in integrating protein expression and phosphorylation. In comparison, we also assessed the ranking performance of the strategy using information of protein expression or protein

phosphorylation individually. Target pathways were ranked more top with the data integration than using the information from proteomic or phosphoproteomic data individually. In the comparisons of pathway analysis strategies, topology-based method outperformed than the others. The subtypes of breast cancer, which consist of Luminal A, Luminal B, Basal and HER2-enriched, vary greatly in prognosis and require distinct treatment. Therefore we applied topology-based pathway analysis with integrating protein expression and phosphorylation profiles on four subtypes of breast cancer. The results showed that TPs were enriched in all subtypes but their ranks were significantly different among the subtypes. For instance, p53 pathway ranked top in the Basal-like breast cancer subtype, but not in HER2-enriched type. The rank of Focal adhesion pathway was more top in HER2- subtypes than in HER2+ subtypes. The results were consistent with some previous researches.

Conclusions: The results demonstrate that the network topology-based method is more powerful by integrating proteomic and phosphoproteomic in pathway analysis of proteomics study. This integrative strategy can also be used to rank the specific pathways for the disease subtypes.

Modeling of Hypoxia gene expression for three different cancer cell lines

Babak Soltanalizadeh[1], Erika Gonzalez Rodriguez[2], Vahed Maroufy[1], W Jim Zheng[3], Hulin Wu[1*]

[1]Department of Biostatistics & Data Science, University of Texas Health Science Center at Houston, Houston, TX, USA
[2]Center for translational Injury Research, Department of Surgery, McGovern Medical School, UT Houston, Houston, TX, USA
[3]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

*Corresponding author

Gene dynamic analysis is essential in identifying target genes involved pathogenesis of various diseases, including cancer. Cancer prognosis is often influenced by hypoxia. We apply a multi-step pipeline to study dynamic gene expressions in response to hypoxia in three cancer cell lines: prostate (DU145), colon (HT29), and breast (MCF7) cancers. We identified 26 distinct temporal expression patterns for prostate cell line, and 29 patterns for colon and breast cell lines. The module-based dynamic networks have been developed for all three cell lines. Our analyses improve the existing results in multiple ways. It exploits the time-dependence nature of gene expression values in identifying the dynamically significant genes; hence, more key significant genes and transcription factors have been identified. Our gene network returns significant information regarding biologically important modules of genes. Furthermore, the network has potential in learning the regulatory path between transcription factors and the downstream genes. In

addition, our findings suggest that changes in genes BMP6 and ARSJ expression might have a key role in the time-dependent response to hypoxia in breast cancer.

Selecting precise reference normal tissue samples for cancer research using a deep learning approach

William Zeng[1], Benjamin S. Glicksberg[1], Yangyan Li[2], Bin Chen[1,3*]

[1]Institute for Computational Health Sciences, University of California, San Francisco, CA, USA
[2]Shandong University, China
[3]Current address: Department of Pediatrics and Human Development, Department of Pharmacology and Toxicology, Michigan State University, Grand Rapids, MI, USA

*corresponding author

Background: Normal tissue samples are often employed as a control for understanding disease mechanisms, however, collecting matched normal tissues from patients is difficult in many instances. In cancer research, for example, the open cancer resources such as TCGA and TARGET do not provide matched tissue samples for every cancer or cancer subtype. The recent GTEx project has profiled samples from healthy individuals, providing an excellent resource for this field, yet the feasibility of using GTEx samples as the reference remains unanswered.
Methods: We analyze RNA-Seq data processed from the same computational pipeline and systematically evaluate GTEx as a potential reference resource. We use those cancers that have adjacent normal tissues in TCGA as a benchmark for the evaluation. To correlate tumor samples and normal samples, we explore top varying genes, reduced features from principal component analysis, and encoded features from an autoencoder neural network. We first evaluate whether these methods can identify the correct tissue of origin from GTEx for a given cancer and then seek to answer whether disease expression signatures are consistent between those derived from TCGA and from GTEx.
Results: Among 32 TCGA cancers, 18 cancers have less than 10 matched adjacent normal tissue samples. Among three methods, autoencoder performed the best in predicting tissue of origin, with 12 of 14 cancers correctly predicted. The reason for misclassification of two cancers is that none of normal samples from GTEx correlate well with any tumor samples in these cancers. This suggests that GTEx has matched tissues for the majority cancers, but not all. While using autoencoder to select proper normal samples for disease signature creation, we found that disease signatures derived from normal samples selected via an autoencoder from GTEx are consistent with those derived from adjacent samples from TCGA in many cases. Interestingly, choosing top 50 mostly correlated samples regardless of tissue type performed reasonably well or even better in some cancers.

Conclusions: Our findings demonstrate that samples from GTEx can serve as reference normal samples for cancers, especially those do not have available adjacent tissue samples. A deep-learning based approach holds promise to select proper normal samples.

Inferring gene-disease association by an integrative analysis of eQTL GWAS and Protein-Protein 3 Interaction data

Jun Wang[1], Jiashun Zheng[2], Zengmiao Wang[3], Hao Li[1,2*] and Minghua Deng[1,4,5*]

[1]Center for Quantitative Biology, Peking University, Beijing, China
[2]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, US
[3]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, US
[4]School of Mathematical Sciences, Peking University, Beijing, China
[5]Center for Statistical Sciences, Peking University, Beijing, China

*Corresponding author

Objectives: Genome-wide association studies (GWAS) have revealed many candidate SNPs, but the mechanisms by which these SNPs influence diseases are largely unknown. In order to decipher the underlying mechanisms, several methods have been developed to predict disease-associated genes based on the integration of GWAS and eQTL data (for example Sherlock and COLOC). A number of studies have also incorporated information from gene networks into GWAS analysis to reprioritize candidate genes.

Methods: Motivated by these two different approaches, we have developed a statistical framework to integrate information from GWAS, eQTL and Protein-Protein interaction (PPI) data to predict disease-associated genes. Our approach is based on a hidden Markov random field model and we called the resulting computational algorithm GeP-HMRF (GWAS-eQTL-PPI based Hidden Markov Random Field).

Results: We compared the performance of GeP-HMRF with Sherlock, COLOC and NetWAS methods on 9 GWAS datasets using the disease-related genes in MalaCards database as the standard, and found that GeP-HMRF significantly improves the prediction accuracy. We also applied GeP-HMRF to an age-related macular degeneration disease (AMD) dataset. Among the top 50 genes predicted by GeP-HMRF, 7 are reported by the MalaCards database to be AMD-related with an enrichment p-value 3.61 × 10−119. Among the top 20 genes predicted by GeP-HMRF, CFHR1, CGHR3, HTRA1 and CFH are AMD-related in MalaCards and another 9 genes are supported by literature.

Conclusions: We built a unified statistical model to predict disease-related genes by integrating the GWAS, eQTL and PPI data. Our approach outperforms Sherlock, COLOC and NetWAS in simulation studies and 9 GWAS datasets. Our approach can be

generalized to incorporate other molecular trait data beyond eQTL and other interaction data beyond PPI.

Network-based identification of critical regulators as putative drivers of human cleft lip

Aimin Li[1,2,*], Guimin Qin[2,3,*], Akiko Suzuki[4,5], Mona Gajera[4,5], Junichi Iwata[4,5,6], Peilin Jia[2,§], Zhongming Zhao[2,6,§]

[1]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, Shaanxi, 710048, China
[2]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]School of Software, Xidian University, Xi'an, Shaanxi 710071, China
[4]Department of Diagnostic and Biomedical Sciences, School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX 77054, USA
[5]Center for Craniofacial Research, The University of Texas Health Science Center at Houston, Houston, TX 77054, USA
[6]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA

*These authors contributed equally to this work

Background: Cleft lip (CL) is one of the most common congenital birth defects with complex etiology. While genome-wide association studies (GWAS) have made significant advances in our understanding of mutations and their related genes with potential involvement in the etiology of CL, it remains unknown how these genes are functionally regulated and interact with each other in lip development. Currently, identifying the disease-causing genes in human CL is urgently needed. So far, the causative CL genes have been largely undiscovered, making it challenging to design experiments to validate the functional influence of the mutations identified from large genomic studies such as CL GWAS.

Results: Transcription factors (TFs) and microRNAs (miRNAs) are two important regulators in cellular system. In this study, we aimed to investigate the genetic interactions among TFs, miRNAs and the CL genes curated from the previous studies. We constructed miRNA-TF co-regulatory networks, from which the critical regulators as putative drivers in CL were examined. Based on the constructed networks, we identified ten critical hub genes with prior evidence in CL. Furthermore, the analysis of partitioned regulatory modules highlighted a number of biological processes involved in the pathology of CL, including a novel pathway "Signaling pathway regulating pluripotency of stem cells". Our subnetwork analysis pinpointed two candidate miRNAs, hsa-mir-27b and hsa-mir-497, activating the Wnt pathway that was associated with CL. Our results were supported by an independent gene expression dataset in CL.

Conclusions: This study represents the first regulatory network analysis of CL genes. Our work presents a global view of the CL regulatory network and a novel approach on investigating critical miRNAs, TFs and genes via combinatory regulatory networks in craniofacial development. The top genes and miRNAs will be important candidates for future experimental validation of their functions in CL.

Concurrent Session - Systems Biology
Tuesday, June 12, 2018
10:05 AM - 12:15 PM
Santa Monica

Investigation of multi-trait associations using pathway-based analysis of GWAS summary statistics

Guangsheng Pei[1], Hua Sun[1], Yulin Dai[1], Xiaoming Liu[2], Peilin Jia[1,*], Zhongming Zhao[1, 2, 3,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

*Corresponding author

Background: Genome-wide association studies (GWAS) have been successful in identifying disease- associated genetic variants. Recently, an increasing number of GWAS summary statistics have been made available to the research community, providing extensive repositories for studies of human complex diseases. In particular, cross-trait associations at the genetic level can be beneficial from large-scale GWAS summary statistics by using genetic variants that are associated with multiple traits. However, direct assessment of cross-trait associations using susceptibility loci has been challenging due to the complex genetic architectures in most diseases, calling for advantageous methods that could integrate functional interpretation and imply biological mechanisms.

Results: We developed an analytical framework for systematic integration of cross-trait associations. It incorporates two different approaches to detect enriched pathways and requires only summary statistics. We demonstrated the framework using 25 traits belonging to four phenotype groups. Our results revealed an average of 54 significantly associated pathways (ranged between 18 and 175) per trait. We further proved that pathway-based analysis provided increased power to estimate cross-trait associations

compared to gene-level analysis. Based on Fisher's Exact Test (FET), we identified a total of 24 (53) pairs of trait-trait association at adjusted $p_{\mathrm{FET}} < 0.001$ ($p_{\mathrm{FET}} < 0.01$) among the 25 traits. Our trait-trait association network revealed not only many relationships among the traits within the same group but also novel relationships among traits from different groups, which warrants further investigation in future.

Conclusions: Our study revealed that risk variants for 25 different traits aggregated in particular biological pathways and that these pathways were frequently shared among traits. Our results confirmed known mechanisms and also suggested several novel insights into the etiology of multi-traits.

Lilikoi: an R package for personalized pathway-based classification modeling using metabolomics data

Sijia Huang[1#], Ph.D, Fadhl Alakwaa[2#], Ph.D, and Lana X. Garmire[1,2*], Ph.D

1Molecular Biology and Bioengineering Graduate Program, University of Hawaii at Monoa, Honolulu, HI, USA
2 Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA

#These authors contributed equally to the work
*Corresponding author

Lilikoi (Hawaiian word for passion fruit) is a new and comprehensive R package for personalized pathway based classification modelling, using metabolomics data. Four basic modules are presented as the backbone of the package: 1) Feature mapper which standardizes the metabolite names provided by users, and map them to pathways. 2) Dimension transformation module transforms the metabolomic profiles to personalized pathway-based profiles using pathway desregulation scores (PDS). 3) Feature selection module which helps to select the significant pathway features related to the disease phenotypes, and 4) Classification and prediction module which offers various machine-learning classification algorithms. The package is freely available under the GPLv3 license through bioconductor at http://bioconductor.org/packages/release/bioc/html/lilikoi, and throught github repository at: https://github.com/lanagarmire/lilikoi

---

Gene2Vec: Distributed Representation of Genes Based on Co-Expression

Jingcheng Du[#1], Peilin Jia[#1], Yulin Dai[1], Cui Tao[1], Zhongming Zhao[1*], Degui Zhi[1*]

The University of Texas School of Biomedical Informatics, Houston, TX 77030, United States

[#]equal contribution
[*]Corresponding author

Background: Existing functional description of genes are categorical, discrete, and mostly through manual process. In this work, we explore the idea of gene embedding, distributed representation of genes, in the spirit of word embedding.
Methods & Results: From a pure data-driven fashion, we trained a 200-dimension vector representation of all human genes, using gene co-expression patterns in 984 data sets from the GEO databases. These vectors capture functional relatedness of genes in terms of recovering known pathways - the average inner product (similarity) of genes within a pathway is 1.52X greater than that of random genes. Using t-SNE, we produced a gene co-expression map that shows local concentrations of tissue specific genes. We also illustrated the usefulness of the embedded gene vectors, laden with rich information on gene co-expression patterns, in tasks such as gene-gene interaction prediction.
Conclusions: We proposed a machine learning method that utilizes transcriptome-wide gene co-expression to generate a distributed representation of genes. We further demonstrated the utility of our distribution by predicting gene-gene interaction based solely on gene names. The distributed representation of genes could be useful for more bioinformatics applications.

A robust fuzzy rule based integrative feature selection strategy for gene 2 expression data in TCGA

Shicai Fan [1,2,3,*], Jianxiong Tang[1], Mengchi Wang[4], Chunguo Wu[3]

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China 611731
[2]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, Sichuan, China 611731
[3]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China 130012
[4]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, California, USA.

Background: Lots of researches have been conducted in the selection of gene signatures that could distinguish the cancer patients from the normal. However, it is still an open question on how to extract the robust gene features.

Methods: In this work, a gene signature selection strategy for TCGA data was proposed by integrating the gene expression data, the methylation data and the prior knowledge about cancer biomarkers. Different from the traditional integration method, the expanded 450K methylation data were applied instead of the original 450K array data, and the reported biomarkers were weighted in the feature selection. Fuzzy rule based classification method and cross validation strategy were applied in the model construction for performance evaluation.

Results: Our selected gene features showed prediction accuracy close to 100% in the cross validation with fuzzy rule based classification model on 6 cancers from TCGA. The cross validation performance of our proposed model is similar to other integrative models or RNA-seq only model, while the prediction performance on independent data is obviously better than other 5 models. The gene signatures extracted with our fuzzy rule based integrative feature selection strategy were more robust, and had the potential to get better prediction results.

Conclusion: The results indicated that the integration of expanded methylation data would cover more genes, and had greater capacity to retrieve the signature genes compared with the original 450K methylation data. Also, the integration of the reported biomarkers was a promising way to improve the performance. PTCHD3 gene was selected as a discriminating gene in 3 out of the 6 cancers, which suggested that it might play important role in the cancer risk and would be worthy for the intensive investigation.

Pessimistic optimization for modeling microbial communities with uncertainty

Meltem Apaydin[1], Liang Xu[2], Bo Zeng[2], Xiaoning Qian[1*]

[1]Dept. of Electrical and Computer Engineering, Texas A&M University, 77843 College Station, USA.
[2]Dept. of Industrial Engineering, University of Pittsburgh, 15260 Pittsburgh, USA.

[*]Corresponding author

It is important to understand the complicated interactions of microbial communities who play critical roles in the ecological system, human health and diseases. Optimization-based mathematical models provide ways to analyze and obtain predictions on microbial communities. However, there are inherent model and data uncertainties from the existing knowledge and experiments about different microbial communities so that the imposed models may not exactly reflect the reality in nature. Here, addressing these challenges and aiming to have a flexible framework to model microbial communities with uncertainty, we introduce P-OptCom, an extension of an existing method OptCom, based on the ideas from the pessimistic bilevel optimization literature. This framework relies on the coordinated decision making between the single upper-(community-level) and multiple lower-level (multiple microorganisms or guilds) decision makers to support robust solutions to better approximate microbial community steady states even when the individual microorganisms' behavior deviate from the optimum in terms of their cellular

fitness criteria. We formulate the problem by considering suboptimal behavior of the individual members, and relaxing the constraints denoting the interactions within communities to obtain a model flexible enough to deal with potential uncertainties. Our study demonstrates that without experimental knowledge in advance, we are able to analyze the trade-offs among the members of microbial communities and closely approximate the actual experimental measurements.

Evaluation of top-down mass spectral identification with homologous protein sequences

Ziwei Li[1,2], Bo He[1], Qiang Kou[3], Zhe Wang[4], Si Wu[4], Yunlong Liu[2,5,*], Weixing Feng[1,*] and Xiaowen Liu[3,5*]

[1]College of Automation, Harbin Engineering University 145, Nan Tong Street 150001 Harbin, Heilongjiang, China.
[2]Department of Medical and Molecular Genetics, Indiana University School of Medicine, 410 West 10th Street, 46202 Indianapolis, IN, USA.
[3]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 719 Indiana Avenue, 46202 Indianapolis, IN, USA.
[4]Department of Chemistry and Biochemistry, University of Oklahoma, 101 Stephenson Parkway, 73019 Norman, OK, USA.
[5]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 410 West 10th Street, 46202 Indianapolis, IN, USA.

[*]Corresponding author

Background: Top-down mass spectrometry has unique advantages in identifying proteoforms with multiple post-translational modifications and/or unknown alterations. Most software tools in this area search top-down mass spectra against a protein sequence database for proteoform identification. When the species studied in a mass spectrometry experiment lacks its proteome sequence database, a homologous protein sequence database can be used for proteoform identification. The accuracy of homologous protein sequences affects the sensitivity of proteoform identification and the accuracy of mass shift localization.
Results: We tested TopPIC, a commonly used software tool for top-down mass spectral identification, on a top-down mass spectrometry data set of *Escherichia coli* K12 MG1655, and evaluated its performance using an *Escherichia coli* K12 MG1655 proteome database and a homologous protein database. The number of identified spectra with the homologous database was about half of that with the *Escherichia coli* K12 MG1655 database. We also tested TopPIC on a top-down mass spectrometry data set of human MCF-7 cells and obtained similar results.

Conclusions: Experimental results demonstrated that TopPIC is capable of identifying many proteoform spectrum matches and localizing unknown alterations using homologous protein sequences containing no more than 2 mutations.

Concurrent Session - Computational drug discovery
Tuesday, June 12, 2018
10:05 AM - 12:15 PM
Hollywood

Drug-Drug Interaction Prediction based on Co-Medication Patterns and Graph Matching

Wen-Hao Chiang[1], Li Shen[2], Lang Li[3], Xia Ning[4,*]

[1]Department of Computer & Information Science, Indiana University - Purdue University Indianapolis, Indianapolis, 46202, USA.
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, 19104, USA.
[3]Department of Biomedical Inforamtics, Ohio State University, Columbus, 43210, USA.
[4]Department of Computer & Information Science, Indiana University - Purdue University Indianapolis, Indianapolis, 46202, USA.
*Corresponding author

Background: The problem of predicting whether a drug combination of arbitrary orders is likely to induce adverse drug reactions is considered in this manuscript.
Methods: Novel kernels over drug combinations of arbitrary orders are developed within support vector machines for the prediction. Graph matching methods are used in the novel kernels to measure the similarities among drug combinations, in which drug co-medication patterns are leveraged to measure single drug similarities.
Results: The experimental results on a real-world dataset demonstrated that the new kernels achieve an area under the curve (AUC) value 0.912 for the prediction problem.
Conclusions: The new methods with drug co-medication based single drug similarities can accurately predict whether a drug combination is likely to induce adverse drug reactions of interest.

Predicting drug response of tumors from integrated genomic profiles by deep neural networks

Yu-Chiao Chiu[1], Hung-I Harry Chen[1,2], Tinghe Zhang[2], Songyao Zhang[2,3], Aparna Gorthi[1], Li-Ju Wang[1], Yufei Huang[2,4*], Yidong Chen[1,4*]

[1]Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA
[2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA
[3]Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[4]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA 15

*Corresponding authors

Background: The study of high-throughput genomic profiles from a pharmacogenomics viewpoint has provided unprecedented insights into the oncogenic features modulating drug response. A recent study screened for the response of a thousand human cancer cell lines to a wide collection of anti-cancer drugs and illuminated the link between cellular genotypes and vulnerability. However, due to essential differences between cell lines and tumors, to date the translation into predicting drug response in tumors remains challenging. Recently, advances in deep neural networks (DNNs) have revolutionized bioinformatics and introduced new techniques to the integration of genomic data. Its application on pharmacogenomics may fill the gap between genomics and drug response and improve the prediction of drug response in tumors.

Results: We proposed a DNN model to predict drug response based on mutation and expression profiles of a cancer cell or a tumor. The model contains three subnetworks, i) a mutation encoder pre-trained using a large pan-cancer dataset to abstract core representations of high-dimension mutation data, ii) a pre-trained expression encoder, and iii) a drug response predictor network integrating the first two subnetworks. Given a pair of mutation and expression profiles, the model predicts IC_{50} values of 265 drugs. We trained and tested the model on a dataset of 622 cancer cell lines and achieved an overall prediction performance of mean squared error at 1.96 (log-scale IC_{50} values). The performance was superior in prediction error or stability than two classical methods (linear regression and support vector machine) and four analog DNNs of our model, including DNNs built without TCGA pre- training, partly replaced by principal components, and built on individual types of input data. We then applied the model to predict drug response of 9,059 tumors of 33 cancer types. Using per-cancer and pan-cancer settings, the model predicted both known, including EGFR inhibitors in non-small cell lung cancer and tamoxifen in ER+ breast cancer, and novel drug targets, such as vinorelbine for *TTN*-mutated tumors. The comprehensive analysis further revealed the molecular mechanisms underlying the resistance to a chemotherapeutic drug docetaxel in a pan-cancer setting and the anti-cancer potential of a novel agent, CX-5461, in treating gliomas and hematopoietic malignancies.

Conclusions: Here we present, as far as we know, the first DNN model to translate pharmacogenomics features identified from in vitro drug screening to predict the response of tumors. The results covered both well-studied and novel mechanisms of drug resistance and drug targets. Our model and findings improve the prediction of drug response and the identification of novel therapeutic options.

Predicting Adverse Drug Reactions through Interpretable Deep Learning Framework

Sanjoy Dey[1], Heng Luo[1], Achille Fokoue[2], Jianying Hu[1], Ping Zhang[1*]

[1]Center for Computational Health, IBM Research AI, 1101 Kitchawan Road, Yorktown Heights, NY, USA.
[2]Cognitive Computing, IBM Research AI, 1101 Kitchawan Road, Yorktown Heights, NY, USA.

*Corresponding author

Background: Adverse drug reactions (ADRs) are unintended and harmful reactions caused by normal uses of drugs. Predicting and preventing ADRs in the early stage of the drug development pipeline can help to enhance drug safety and reduce financial costs.
Methods: In this paper, we developed machine learning models including a deep learning framework which can simultaneously predict ADRs and identify the molecular substructures associated with those ADRs without defining the substructures a-priori.
Results: We evaluated the performance of our model with ten different
state-of-the-art fingerprint models and found that neural fingerprints from the deep learning model outperformed all other methods in predicting ADRs. Via feature analysis on drug structures, we identified important molecular substructures that are associated with specific ADRs and assessed their associations via statistical analysis.
Conclusions: The deep learning model with feature analysis, substructure identification, and statistical assessment provides a promising solution for identifying risky components within molecular structures and can potentially help to improve drug safety evaluation.

Predict drug sensitivity of cancer cells with pathway activity inference

Xuewei Wang[1], Zhifu Sun[1], Michael Zimmermann[1], Andrej Bugrim[2], Jean-Pierre Kocher[1]

[1]Division of Biomedical and Statistical Informatics, Mayo Clinic, Rochester, MN, USA
[2]Silver Beach Analytics, Inc, St Joseph, MI, USA

*Corresponding author

Predicting cellular responses to drugs has been a major challenge for personalized drug therapies. Recent pharmacogenomic studies profiled the sensitivities of heterogeneous cell lines to numerous drugs, and provided valuable data resources to develop and validate computational approaches for the prediction of drug responses. Most of current approaches predict drug sensitivity by building prediction models with individual genes, which oftentimes suffer from practical limitations such as difficulty to interpret biological relevance. As an alternative, drug response of cancer cells could be predicted based on pathway activity scores derived from gene expression. In this study, pathway-based prediction models were built with four approaches inferring pathway activity in unsupervised manner, including competitive scoring approaches (*DiffRank* and *GSVA*) and self-contained scoring approaches (*PLAGE* and *Z-score*). It's the first time to compare such unsupervised pathway scoring approaches in the context of drug sensitivity prediction. Our analysis on all the 24 drugs from Cancer Cell Line Encyclopedia (CCLE) demonstrated that pathway-based models achieved better predictions for 14 out of the 24 drugs, while taking much less features as inputs. Further investigation focusing on drug-related genes (targets, transporters and metabolic enzymes) indicated that pathway-models indeed captured pathways involving drug-related genes for majority of drugs, whereas gene-models failed to identify drug targets in most cases. Among the four approaches, competitive scoring (*DiffRank* and *GSVA*) tend to provide more accurate predictions and captured more pathways involving drug-related genes than self-contained scoring (PLAGE and Z-Score). Detailed interpretation of top pathways from *DiffRank* (the top method) has demonstrated multiple merits of pathway-based approaches toward drug sensitivity prediction, particularly the ability to identify pathways relevant to drug mechanisms. Taken together, pathway-based modeling with inferred pathway activity is a promising alternative to predict drug response and provide mechanistic insights into the mechanisms of drug actions.

Application of Transfer Learning for Cancer Drug Sensitivity Prediction

Saugato Rahman Dhruba[1], Raziur Rahman[1], Kevin Matlock[1], Souparno Ghosh[2], Ranadip Pal[1*]

1 Department of Electrical and Computer Engineering, Texas Tech University, 1012 Boston Ave, 79409 Lubbock, TX, USA.
2 Department of Mathematics and Statistics, Texas Tech University, 1108 Memorial Circle, 79409 Lubbock, TX, USA

*Corresponding author

Background: In precision medicine, scarcity of suitable biological data often hinders the design of an appropriate predictive model. In this regard, large scale pharmacogenomics studies, like CCLE and GDSC hold the promise to mitigate the issue. However, one cannot directly employ data from multiple sources together due to the existing distribution shift in data. One way to solve this problem is to utilize the transfer learning methodologies tailored to fit in this specific context.

Results: In this paper, we present two novel approaches for incorporating information from a secondary database for improving the prediction in a target database. The first approach is based on latent variable cost optimization and the second approach considers polynomial mapping between the two databases. Utilizing CCLE and GDSC databases, we illustrate that the proposed approaches accomplish a better prediction of drug sensitivities for different scenarios as compared to the existing approaches.

Conclusion: We have compared the performance of the proposed predictive models with database-specific individual models as well as existing transfer learning approaches. We note that our proposed approaches exhibit superior performance compared to the abovementioned alternative techniques for predicting sensitivity for different anti-cancer compounds, particularly the nonlinear mapping model shows the best overall performance.

Large-scale mining disease comorbidity relationships from post-market drug adverse events surveillance data

Chunlei Zheng[1] and Rong Xu[1*]

[1]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, 2103 Cornell Road, 44106 Cleveland, USA

*Corresponding author

Background: Systems approaches in studying disease relationship have wide applications in biomedical discovery, such as disease mechanism understanding and drug discovery. The FDA Adverse Event Reporting System contains rich information about patient diseases, medications, drug adverse events and demographics of 17 million case reports. Here, we explored this data resource to mine disease comorbidity relationships using association rule mining algorithm and constructed a disease comorbidity network.

Results: We constructed a disease comorbidity network with 1,059 disease nodes and 12,608 edges using association rule mining of FAERS (14,157 rules). We evaluated the performance of comorbidity mining from FAERS using known disease comorbidities of multiple sclerosis (MS), psoriasis and obesity that represent rare, moderate and common disease respectively. Comorbidities of MS, obesity and psoriasis predicted from our network achieved precisions of 58.6%, 73.7%, 56.2% and recalls 87.5%, 69.2% and 72.7% separately. We performed comparative analysis of the disease comorbidity network with disease semantic network, disease genetic network and disease treatment network. We show that (1) disease comorbidity clusters exhibit significantly higher semantic similarity

than random network (0.18 vs 0.10); (2) disease comorbidity clusters share significantly more genes (0.46 vs 0.06); and (3) disease comorbidity clusters shares significantly more drugs (0.64 vs 0.17). Finally, we demonstrate that the disease comorbidity network has potential in uncovering interesting disease patterns using asthma as a case study.

Conclusions: Our study presents the first attempt to build a disease comorbidity network from FDA Adverse Event Reporting System. This network shows well correlated with disease semantic similarity, disease genetics and disease treatment, which has great potential in disease genetics prediction and drug discovery.

Concurrent Session - International PI Talk
Tuesday, June 12, 2018
2:20 PM - 4:30 PM
Grand Imperial North

Bioinformatics research on potential ability of circular RNA encoding protein

Xiaofeng Song

Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Circular RNA (circRNAs) is a type of RNA molecule which, unlike all other known RNAs, forms a covalently closed loop structure with no 5’ to 3’ polarity and is commonly generated through back-splicing of pre-mRNA. Although circRNAs are widely expressed in human tissues, in particular in the brain, their biological functions remain largely uncharacterized. CircRNAs were initially regarded as a noncoding RNAs, but previous results have indicated that artificial synthetic circRNAs containing an IRES (internal ribosome entry site) element are shown to be able to encode proteins, which implies the potential coding ability of natural circRNAs. Accordingly, we established a database, named circRNADb, for human circRNAs with protein-coding annotations. In this database, we firstly evaluated the protein-coding potential for each circRNAs and then predicted the IRES position for those circRNAs with coding potential. Two circRNAs (circ-SHPRH and circ-FBXW7) derived from circRNADb were experimentally validated to encode proteins by other research group. Furthermore, their IRESs derived from circRNADb were also proved to be true by the bicistronic reporter assay. Beside them, some other circRNAs were also experimentally verified to encode proteins in an IRES-dependent manner. Although only a few protein-coding circRNAs has been found so far, their encoded proteins are found to play important roles in cellular responses to environmental stress, myogenesis, and the suppression of glioma tumorigenesis. Thus, to gain more insights into the functions of circRNAs, comprehensive identification of protein-coding circRNAs is the first step. As the translation of circRNAs is driven by IRES, the identification of IRES is important for the prediction of coding circRNAs. Since experimental validation of IRES elements is time-consuming and labor-intensive, there is an urgent need for computational methods to identify IRES. Therefore, we proposed a novel method called IRESfinder specific for identifying cellular IRESs, and it outperforms the other methods.

The impact of genetic admixture and natural selection on driving population differences in East Asia

Shuhua Xu

Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai 200031, China.

Genetic admixture in human, the result of inter-marriage among people from different well-differentiated populations, has been a common phenomenon throughout the history of modern humans. Understanding of population admixture dynamics is important not only to mapping human complex traits/diseases but also to other applications, such as elucidating population history and detecting natural selection signatures. In this talk, I will briefly present our recent progresses in analyzing human population admixture in East Asia. In particular, I would propose that gene flow (or admixture), as an evolutionary driving force and also an under-investigated one, played an important role in shaping human genetic diversity and population differentiation in East Asia.

Identification of functional PTM events in autophagy

Yu Xue[1,*], Yongbo Wang[1], and Wankun Deng[1]

[1]Key Laboratory of Molecular Biophysics of Ministry of Education, College of Life Science and Technology and the Collaborative Innovation Center for Biomedical Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China.

Autophagy is a highly conserved process for degrading cytoplasmic contents, determines cell survival or death, and regulates the cellular homeostasis. Besides core ATG proteins, numerous regulators together with various post-translational modifications (PTMs) are also involved in autophagy. Recent studies demonstrated that the dysregulation of macroautophagy/autophagy is involved in human diseases such as cancers and neurodegenerative disorders. Thus, autophagy has become a promising therapeutic target for biomedical design. Here, we developed a database of The Autophagy, Necrosis,

ApopTosis OrchestratorS (THANATOS, http://thanatos.biocuckoo.org), containing 191,543 proteins potentially associated with autophagy and cell death pathways in 164 eukaryotes. We performed an evolutionary analysis of core ATG genes, and observed that ATGs required for the autophagosome formation are highly conserved across eukaryotes. Further analyses revealed that known cancer genes and drug targets were over-represented in human autophagy proteins, which were significantly associated in a number of signaling pathways and human diseases. By re-constructing a human kinase-substrate phosphorylation network for core ATG proteins, our results confirmed that phosphorylation play a critical role in regulating autophagy. Using this data resource, we performed a quantitative phosphoproteomic profiling to delineate the phosphorylation signalling networks regulated by 2 natural neuroprotective autophagy enhancers, corynoxine (Cory) and corynoxine B (Cory B). We developed a novel algorithm of in silico Kinome Activity Profiling (iKAP) to predict that Cory or Cory B potentially regulates different kinases. We discovered 2 kinases, MAP2K2/MEK2 (mitogen-activated protein kinase kinase 2) and PLK1 (polo-like kinase 1), to be potentially upregulated by Cory, whereas the siRNA-mediated knockdown of Map2k2 and Plk1 significantly inhibited Cory-induced autophagy. Furthermore, Cory promoted the clearance of Alzheimer disease-associated APP (amyloid beta [A4] precursor protein) and Parkinson disease-associated synuclein alpha (SNCA/α-synuclein) by enhancing autophagy, and these effects were dramatically diminished by the inhibition of the kinase activities of MAP2K2 and PLK1. Taken together, our study not only provided bioinformatics resources and approaches for analyzing PTMs in autophagy, but also identified the important role of MAP2K2 and PLK1 in neuronal autophagy.

Tumor heterogeneity in hepatocellular carcinoma and intrahepatic cholangiocarcinoma

Ruibin Xi

School of Mathematical Sciences, Peking University, Beijing, 100871, China.
Center for Statistical Science, Peking University, Beijing, 100871, China

Abstract: Tumor heterogeneity is an emerging theme that we only start to understand recently. Many human tumors display genetic and epigenetic heterogeneity. These heterogeneities may have important implications in tumor progression, therapeutic effect and tumor prognosis. Here, we used multi-regional exome sequencing to study the tumor heterogeneity of hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC). Our analysis revealed that HCC and ICC tumors have significant genetic heterogeneity. For many patients, the tumor cells from different regions often have different sets of targetable mutations. Further experimental analysis show that these subregional tumors respond differently to target therapies due to their genetic heterogeneity.

Abstract 7

Study on risk prediction model of type 2 diabetes

Xin Zhang[1,2], Yun Liu[1,2,*], Tingyu Xu[1,2]

[1]Department of Information, the First Affiliated Hospital, Nanjing Medical University, No.300 Guang Zhou Road, Nanjing, Jiangsu, 210029, China

[2]Institute of Medical Informatics and Management, Nanjing Medical University, No.300 Guang Zhou Road, Nanjing, Jiangsu, 210029, China

*Corresponding author

Objective: The paper studies the retrospective data of patients with type 2 diabetes mellitus. We used the artificial intelligence methods to analyze the clinical big data, and established a diabetes risk prediction model.

Method: From the Clinical Data Repository (CDR) of the First Affiliated Hospital of Nanjing Medical University, the data of 140,000 patients with type 2 diabetes and the same number of non-diabetic patients from 2005 to 2015 was extracted. Two-thirds of the data formed the training set, and another one-third of the data formed the test set. Using the basic information, diagnosis, treatment, test results and follow-up information, we carried out data mining with the improved support vector machine (SVM) algorithm and convolutional neural network (CNN) algorithm. Then we established the prediction model and tested the model.

Result: The accuracy of the model was 70.27%, and the area under the ROC curve was 79.40%.

Conclusion: The risk prediction model of type 2 diabetes based on the general population in Jiangsu Province of China was established successfully. This model was validated to provide a reference for clinical practice. In the future, if we combine clinical data with genomics data, the accuracy of the risk prediction model would be improved.

Accurate whole transcriptome assembly using genome-free approach in systematic exploration of under-developed model organism

Yangmei Qin[1], Fan Lin[2], Zhi-Liang Ji[1,*]

[1]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, Fujian, P R China
[2]Software School, Xiamen University, Xiamen 361005, Fujian, P R China

*Corresponding author

For the organisms that have no genome or partial genome, it is a big challenge to achieve a reliable transcriptome for systematic study. To solve this problem, we developed a computational workflow for constructing an integrative cDNA library from heterogeneous RNA-Seq and EST data as an alternative reference to genome, using Amphioxus as a model. Transcriptome assembled by referring to the library showed superior to that by genome-based or de novo methods. For the first time, we identified 52,896 novel transcripts and 35,454 potential isoforms for Amphioxus. On top of the library-based transcriptomes, we illustrated how transcription factors ARNT2/SIM1 switch their isoform roles in regulation of downstream target genes.

Furthermore, we depicted seven activity modes of transcription factors in Amphioxus embryo development. In summary, building an integrative cDNA library can be a practical solution in replace of genome for large scale -omics studies. This method will be particularly useful for systematic and dynamic exploration of gene behavior in those under-developed model organisms, regardless of the genome quality.

Pseudo nucleotide composition-a smart sequence encoding scheme

Wei Chen[1*], Hao Lin[2*]

[1]Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China
[2]Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

*Corresponding author

With the advent of next-generation sequencing technologies, the available data of sequenced genomes and transcriptomes was dramatically expanded, and more attention has been paid on the annotation of functional elements therein. Towards this goal, many computational methods have been proposed to decode the complicate genomes. However, most of the existing methods were based on the nucleic acid composition or some short-range or local sequence order effect without taking into account the global sequence order effect. In addition, the structural and physicochemical properties of nucleotide are another layer of hidden information that is also important for decoding the genomes. It's a pity that this kind of hidden information was only considered in a relatively few studies. Thus, it's necessary to develop a new method, in which both the sequence order information and structural and physicochemical properties of nucleotide should be integrated. Keeping this in mind, we proposed a novel method, called pseudo K-tuple nucleotide composition (PseKNC), to encode genomic sequences. Compared with the conventional nucleotide composition, PseKNC has the advantage of converting sequences with different lengths to a fixed-length digital vector, while at the same time, keeping the long-range sequence order information via the structural and physicochemical properties of the constituent oligonucleotides. For the convenience of scientific community, a freely available and open source package, called PseKNC-General, was provided at http://lin-group.cn/server/pseknc. By selecting the 106 kinds of DNA structural and physicochemical properties or 12 kinds of RNA structural and physicochemical properties included in the package, users can easily generate many different modes of PseKNC (such as conventional k-tuple nucleotide compositions, Moreau-Broto autocorrelation coefficient, Moran autocorrelation coefficient, Geary autocorrelation coefficient, Type I PseKNC, and Type II PseKNC) for DNA or RNA sequences, respectively. The performance and smartness of PseKNC have been demonstrated by its applications in identifying functional elements in genomes. It is

anticipated that the PseKNC method together with the PseKNC-General open source package will become a useful tool in computational genomics.

Development of a Generalized Text Mining Framework for Characterizing Low Back Pain in Primary Care: A Pilot Study

Akshay Rajaram MMI[1], Michael Judd[2], Farhana H. Zulkernine PhD PEng[2], David Barber MD[3], Brent Wolfrom MD MSc[3]

[1]School of Medicine, Queen’s University, Kingston, ON, Canada
[2]School of Computing, Queen’s University, Kingston, ON, Canada
[3]Department of Family Medicine, Queen’s University, Kingston, ON, Canada

Approximately 20% of Canadians report chronic low back pain (LBP).1,2 The information used to make the diagnosis is often collected over multiple patient encounters and stored as unstructured free text within the electronic medical record (EMR). In recent years, text mining (TM) has emerged as a potential, automated solution to process this invaluable information using natural language processing (NLP) and machine learning techniques. A recent review of TM case-detection algorithms proves that they can be very effective; combining structured information within the EMR and free-text for case-detection resulted in some studies having sensitivity and specificity above 90%.3

We developed a TM framework using NLP and machine learning techniques to characterize the type of LBP for a subset of primary care patients. We compared the results to a manual medical record review. We used the Centre for Effective Practice Clinically Organized Relevant Exam Back Tool as a guide in abstracting historical data and physical findings from the unstructured free text notes and to characterize the type of pain based on the information collected.

12 male patients and 8 female patients were included. The average age of included patients was 54. The duration of pain varied from 0.1 to 8.5 years. 11 patients had no functional limitation described in their record. Five patients reported being unable to work because of pain and four patients reported other limitations (e.g., walking, lifting restrictions). Chart review yielded the following pain diagnoses: 10 disc, six compressed nerve (CN), and one of each of facet, spinal stenosis, non-mechanical (NM), and mixed types. The TM system assigned the following diagnoses: 12 disc pain, five compressed nerve pain, one facet joint, one spinal stenosis, and one non-mechanical. Using disc pain as a reference, the kappa statistic was 0.80.

The strong Kappa value suggests that there was excellent agreement between the text mining system and the chart review. However, this result must be interpreted with caution given the limited size of the dataset. With such a small dataset, machine learning algorithms can be over or under trained, and statistical modelling becomes inherently difficult. Furthermore, every included patient had some type of back pain, making it impossible to calculate sensitivity and specificity and harder to judge the performance of the case-detection algorithm. Our future work with the text mining framework will address these issues and look at combining structured and unstructured data to increase diagnostic accuracy.

ProLanGO: protein function prediction using neural machine translation based on recurrent neural network

Renzhi Cao[1], Colton Freitas[1], Leong Chan[2], Miao Sun[3], Haiqing Jiang[4], Zhangxin Chen[5]

[1]Department of Computer Science, Pacific Lutheran University, WA 98447, USA
[2]School of Business, Pacific Lutheran University, WA 98447, USA
[3]Baidu Inc., Sunnyvale CA 94089, USA
[4]Hiretual Inc., CA, USA
[5]School of electronic engineering, University of Electronic Science and Technology of China, China

With the development of next generation sequencing techniques, it is fast and cheap to determine protein sequences, but relatively slow and expensive to extract useful information from protein sequences because of limitations of traditional biological experimental techniques. Protein function prediction has been a long standing challenge to fill the gap between huge amount of protein sequences and the known function. In this paper, we propose a novel method to convert the protein function problem into a language translation problem by the new proposed protein sequence language "ProLan" to the protein function language "GOLan", and build a neural machine translation model based on recurrent neural networks to translate "ProLan" language to "GOLan" language. We blindly test our method by attending the latest third Critical Assessment of Function Annotation (CAFA 3) in 2016, and also evaluate the performance of our methods on selected proteins which function has been released after CAFA competition. The good performance on the training and testing datasets demonstrates that our new proposed method is a promising direction for protein function prediction. In summary, we first time propose a method which converts protein function prediction problem to a language translation problem, and applies neural machine translation model for protein function prediction.

Joint Principal Trend Analysis for Longitudinal High-Dimensional Data

Yuping[1,2,3], Zhengqing Ouyang[4,5,6]

[1]Department of Statistics, University of Connecticut, Storrs, Connecticut, U.S.A.
[2]Center for Quantitative Medicine, University of Connecticut Health Center, Farmington, Connecticut, U.S.A.
[3]Institute for Systems Genomics, Institute for Collaboration on Health, Intervention, and Policy, CT Institute of the Brain and Cognitive Sciences, University of Connecticut, Storrs, Connecticut, U.S.A.
[4]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, U.S.A.
[5]Department of Biomedical Engineering, Institute for Systems Genomics, University of Connecticut, Storrs, Connecticut, U.S.A.
[6]Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, Connecticut, U.S.A

We consider a research scenario motivated by integrating multiple sources of information for better knowledge discovery in diverse dynamic biological processes. Given two longitudinal high-dimensional datasets for a group of subjects, we want to extract shared latent trends and identify relevant features. To solve this problem, we present a new statistical method named as joint principal trend analysis (JPTA). We demonstrate the utility of JPTA through simulations and applications to gene expression data of the mammalian cell cycle and longitudinal transcriptional profiling data in response to influenza viral infections.

HBPred: a tool to identify growth hormone-binding proteins

Hua Tang[1,*], Hao Lin[2]

[1]Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China; [2]Key Laboratory for NeuroInformation of Ministry of Education, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

*Corresponding author

Hormone-binding protein (HBP) is a kind of soluble carrier protein and can selectively and non- covalently interact with hormone. HBP plays an important role in life growth, but its function is still unclear. Correct recognition of HBPs is the first step to further study their function and understand their biological process. However, it is difficult to correctly recognize HBPs from more and more proteins through traditional biochemical experiments because of high experimental cost and long experimental period. To overcome these disadvantages, we designed a computational method for identifying HBPs accurately in the study. At first, we collected HBP data from UniProt to establish a high-quality benchmark dataset. Based on the dataset, the dipeptide composition was extracted from HBP residue sequences. In order to find out the optimal features to provide key clues for HBP identification, the analysis of various (ANOVA) was performed for feature ranking. The optimal features were selected through the incremental feature selection strategy. Subsequently, the features were inputted into support vector machine (SVM) for prediction model construction. Jackknife cross-validation results showed that 88.6% HBPs and 81.3% non-HBPs were correctly recognized, suggesting that our proposed model was powerful. This study provides a new strategy to identify HBPs. Moreover, based on the proposed model, we established a webserver called HBPred, which could be freely accessed at http://lin-group.cn/server/HBPred.

Deep architectures are not necessary for anomaly classification of matrix‐formed omics data

Hui Yu[1], Wei Yue[1], Ying-Yong Zhao[2], Yan Guo[1,*]

[1]Department of Internal Medicine, University of New Mexico, Albuquerque, NM, 87131, USA
[2]Key Laboratory of Resource Biology and Biotechnology in Western China, School of Life Sciences, Northwest University, Xi'an, Shaanxi 710069, China

* Corresponding author

Motivation: While deep learning has made breathtaking successes in tackling sequence‐based problems, its effectiveness in phenotype classification using numerical matrix‐formed omics data remains under studied. It is informative to compare deep learning neural networks with classical machine learning methods in the setting of high throughput omics data for anomaly classification purpose. Using 37 high throughput datasets, covering transcriptomes and metabolomes, we evaluated the classification power of deep learning and traditional machine learning methods. Representative deep learning methods, Multi‐Layer Perceptrons (MLP) and Convolutional Neural Networks (CNN), were deployed and explored in seeking optimal architectures for the best classification performance. Together with five classical supervised classification methods (Linear Discriminant Analysis, Multinomial Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine), MLP and CNN were comparatively tested on the 37 datasets to predict disease stages or discriminate diseased samples from normal samples.
Results: Single layer MLPs with ample hidden units outperformed deeper MLPs, and 64 to 128 hidden neurons seemed sufficient to yield highest prediction accuracy. MLPs of two hidden layers returned nearly as good performance as single layered MLPs and they were more robust than single layer MLPs in case of extremely imbalanced class composition. CNN was not conspicuous in either accuracy or robustness, even non‐comparable to most traditional methods. Summarizing the results across all 37 datasets, MLP achieved the highest accuracy among all methods tested, and it was one of the most robust methods against imbalanced class composition and inaccurate class labels. Secondary findings include that a drop‐out layer consistently promoted the classification accuracy for MLP but not for CNN, and that MLP and CNN cost tens of thousands times more computation time than Linear Discriminant Analysis.
Conclusions: The comparative results proved that well‐configured single‐layer or two‐layer MLPs are a good choice for anomaly classification of matrix‐formed omics

data. Shallow MLPs with ample hidden neurons are sufficient to achieve superior classification performance in handling numerical matrix - formed omics data. Conclusions and guidance generated from this study are helpful for improving future neural network applications on omics data matrices.

Undisclosed Pure Natural Molecule (Drug-A) Exhibits Anti-Cancer Activity on Breast, Colorectal and Brain Cancer Cell Lines.

Alrfaei, Bahauddeen[1]; Al-Akiel, Maaged[2]; Albahkali, Sarah[1]; Bader, Ammar[3]; Al-Hujaily Ensaf[4]; Halwani, Majed[5*]

[1]Stem Cell & Regenerative Medicine Department, King Abdullah International Medical Research Center, King Saud bin Abdulaziz University for health sciences, Riyadh, Saudi Arabia
[2]Department of Clinical Laboratory, College of Applied Medical Sciences, King Saud bin Abdulaziz University for Health Sciences, King Abdullah International Medical Research Center
[3]Department of Pharmacognosy, Faculty of Pharmacy, Umm Al-Qura University, Makkah, Saudi Arabia.
[4]Medical Genomics Research Department, King Abdullah International Medical Research Center, King Saud bin Abdulaziz University for health sciences, Riyadh, Saudi Arabia
[5]Nanomedicine Department, King Abdullah International Medical Research Center, King Saud bin Abdulaziz University for health sciences, Riyadh, Saudi Arabia

*Corresponding author

Cancer is one of the most intriguing diseases for oncologist. Even though, there are some therapeutic improvements in some cancer therapies such as breast and colorectal cancer, others still far from satisfaction such as brain cancer. Conventional therapies are still struggling to keep up with malignancies and resistance. Introduction of new treatment or management to cancer patients are highly needed. Drug-A is a natural volatile molecule that showed anti-inflammatory, anti-oxidant and anti-microbial effects. In This research we report its ability to inhibit growth of malignant cells such as breast, colorectal and brain cancer. Different concentrations of Drug-A (0.02 mg, 0.04 mg, 0.08 mg, 0.16 mg, 0.33 mg, 0.67 mg, 1.35 mg, 2.7 mg and 5.4 mg) were incubated with the following cancer cell lines: HT29 (colorectal), MCF7 (breast), U87 (brain glioblastoma), and Daoy (brain medulloblastoma). MTT assay detected less viability or metabolic activities when cells are treated with Drug-A. IC50 was averaged between cell lines at 168 ng/ml. Drug-A has the potential to be used as anticancer treatment

especially that it has inhibitory effects upto 60%. Nonetheless, further studies are required such as synergistic outcome in combination with existing chemotherapeutic drugs. Brain malignancies are the most difficult to treat in humans due variety of reasons which include challenging surgical resection, high resistance to chemotherapy, and fast relapse. Current treatment offers less than two years of life span extension. Obviously, new methods for patient's managements and therapies are required. Since Temozolomide is being used as standard of care in treating GBM, we investigated synergistic dose between TMZ and Drug-A. We found that 0.297mM from each drug deliver synergistic effect with maximized growth inhibition. These data show new potential of Drug-A involving cancer treatment.

En-CNN: Predicting DNA binding sites in proteins by ensemble convolutional neural network

Yongqing Zhang[1,2], Shengjie Ji[1], Yuan Feng[1]
[1]School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

*Corresponding author

Protein-DNA interactions play crucial roles in various crucial process, which are essential for gene regulation. The prediction of DNA- binding sites in proteins is critical and challenging for understanding gene regulation, especially in the post-genome era where large volumes of protein sequences have quickly accumulated. In this study, we report a new predictor, En-CNN, for predicting DNA-binding sites form protein primary sequences. En-CNN uses a protein's evolutionary information and sequence feature as two base features and employs sampling strat- egy to deal with the class imbalance problem. Multiple initial predictor with CNN as classifiers are trained by applying ANASYN and a ran- dom under-sampling technique to the original dataset. The final ensem- ble predictor is obtained by majority voting. It predicts DNA-binding sites with 79.48% sensitivity, 92.33% specificity, 90.69% accuracy and 0.632 MCC when tested on a dataset with 543 protein-DNA complexes. Compared with a recently published method, TargetDNA, our method predicts DNA-binding sites with a 0.33 better MCC value when tested on the same dataset. Thus, our prediction method will be useful in finding such DNA-binding sites.

Predicting the clinical pathogenicity of mutations in Lamin A/C to identify laminopathic patients in a large healthcare population

Joseph Park[1], Michael Levin[1], Renae Judy[1], Rachel Kember[1], Anjali T. Owens[2], Scott M. Damrauer[3], Daniel J. Rader[1,2,4]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
[2]Penn Cardiovascular Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
[3]Department of Vascular Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
[4]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Background: Mutations in Lamin A/C (LMNA) lead to several rare cardiovascular diseases including dilated cardiomyopathy and familial partial lipodystrophy type 2, for which genetic testing is often overlooked in the clinical setting. There has been a recent focus to better understand the genetics of cardiovascular disease by studying large heterogeneous healthcare populations from an unbiased, genotype-first standpoint. In this realm, our study associates predicted deleterious variants in LMNA with diverse clinical phenotypes captured in electronic health records (EHR).
Methods: We recently constructed a database called the Penn Medicine Biobank, comprised of whole-exome sequencing data and EHR data for over 10,000 patients. We conducted gene-based burden tests of association for rare, computationally predicted pathogenic mutations in LMNA using Phenome-Wide Association Studies (PheWAS). Additionally, we conducted regression analyses on echocardiography and serum laboratory measurements extracted from the EHR to further investigate the clinical characteristics of mutation carriers.
Results: Gene-based burden association analyses showed that predicted deleterious variants in LMNA are associated with primary cardiomyopathy and cardiac conduction disorders. The association signal was most robust when interrogating predicted loss of function mutations with nonsynonymous variants predicted to be deleterious by REVEL (p=1.78E-11; N=72 carriers). Deeper interrogation of the EHR revealed that carriers had echocardiographic measurements consistent with dilated cardiomyopathy (e.g. increased left atrial volume index, decreased LVEF), as well as serum laboratory measurements consistent with fatty liver (e.g. elevated LFTs, total cholesterol, triglycerides). Despite the high prevalence of disease among these carriers, only six individuals received clinical genetic testing to confirm their laminopathies.

Conclusions: We present the first report of a genotype-first approach to examining the pleiotropic clinical effects of loss of function variants in LMNA by fully utilizing data available in the EHR beyond billing codes. Our study suggests a lack of clinical genetic testing in patients with laminopathies, and that a better understanding of LMNA loss of function variants present among healthcare populations is warranted to change medical management and avoid nonspecific diagnoses. Finally, our study represents an important intermediate step toward developing a comprehensive mutation panel for LMNA genetic testing.

Highly Sensitive and Specific Statistical Approach for Accurate Detection of Gene Fusions and Cryptic Splicing Events Using RNA-Seq

Roozbeh Dehghannasiri[1], Julia Salzman[1,2]

[1]Department of Biochemistry, Stanford University, Stanford, CA, USA,
[2]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

A salient aim of cancer genomics is to discover cancer-specific splicing events that could be resulted from genomic rearrangement and instability. Perhaps the best example of these cancer-associated RNA events is gene fusions, which are formed by joining exons from previously unrelated genes. Gene fusions are among the most powerful biomarkers and treatment targets for targeted cancer therapy. Precise detection of various RNA splicing events particularly gene fusions is crucial for answering fundamental cancer biology questions and also from a translational perspective plays a major role in developing effective cancer diagnosis and therapy techniques.

Despite rapid improvements in RNA analysis algorithms over the past decade, an unbiased and low-false-positive algorithm is still out of reach. There is an increasing agreement in the cancer genomics field that the only viable solution to filter out false positives and detect bona fide splicing events is to take advantage of the rich theory of statistics and utilize its powerful tools to develop reliable statistical RNA-Seq algorithms.

In this abstract, we propose a highly sensitive and specific statistical algorithm for detecting splicing events, namely fusion transcripts and cryptic splicing events. We first build a database of candidate junctions using split-read mapping of the reads that fail to map to the reference genome. Then the initial junction database is refined by defining a scoring function and keeping only junctions whose evaluated score is above a predetermined threshold. Then all unaligned reads from the first step are realigned to the refined junction database. The realignment information such as junction overlap, mapping quality, and alignment score are used as the predictors in a generalized linear model (GLM) to predict a per-read probability, being the estimated likelihood of true alignment for each read alignment. For each putative junction, all per-read probabilities are aggregated, and a p-value is computed by building an empirical null model. The p-value is a statistical score characterizing the probability of each junction being an artifact. Performance evaluation based on various RNA-Seq datasets demonstrates the promising performance of the proposed method in terms of achieving high sensitivity and specificity compared to other existing RNA algorithms.

Epigenomic Patterns Are Associated with Gene Haploinsufficiency and Predict Risk Genes of Developmental Disorders

Siying Chen[1], Xinwei Han[1], Yufeng Shen[1,*]

[1]Department of Systems Biology, Columbia University Medical Center, New York, NY

*Corresponding author

Haploinsufficiency is a major mechanism of genetic risk of human disease. Accurate prediction of haploinsufficient genes is essential for prioritizing and interpreting deleterious variants in genetic studies. Current methods based on mutation intolerance in population data suffer from inadequate power for genes with short transcripts or under modest selection. In this study we showed haploinsufficiency is strongly associated with regulatory complexity of gene expression measured by epigenomic profiles, and then developed a computational method (Episcore) to predict haploinsufficiency from epigenomic and gene expression data from a broad range of tissue and cell types using machine learning methods. Using data from recent exome sequencing studies of developmental disorders, Episcore achieved better performance in prioritizing loss of function de novo variants than current methods. We further showed that Episcore was complementary to mutation intolerance metrics for prioritizing loss of function variants. Our approach enables new applications of epigenomic and gene expression data and facilitates discovery of novel risk variants in studies of developmental disorders.

Comparative gene co-expression network analysis of epithelial to mesenchymal transition reveals lung cancer progression stages

Daifeng Wang[1,2], John D. Haley[2,3], Patricia Thompson[2,3]

[1]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA
[2]Stony Brook Cancer Center, Stony Brook Medicine, Stony Brook, NY, USA
[3]Department of Pathology, Stony Brook Medicine, Stony Brook, NY, USA

The epithelial to mesenchymal transition (EMT) plays a key role in lung cancer progression and drug resistance. However, the dynamics and stability of gene expression patterns as cancer cells transition from E to M at a systems level and relevance to patient outcomes are unknown. To address this, using comparative network and clustering analysis, we systematically analyzed time-series gene expression data from lung cancer cell lines H358 and A549 that were induced to undergo EMT [1]. In particular, we predicted the putative regulatory networks controlling EMT expression dynamics, especially for the EMT-dynamic genes and related these patterns to patient outcomes using data from TCGA. Also, we validated the EMT hub regulatory genes using RNAi. From the network, we identified several novel genes distinct from the static states of E or M that exhibited temporal expression patterns or 'periods' during the EMT process that were shared in different lung cancer cell lines. For example, cell cycle and metabolic genes were found to be similarly down-regulated where immune-associated genes were up-regulated after middle EMT stages. The presence of EMT-dynamic gene expression patterns supports the presence of differential activation and repression timings at the transcriptional level for various pathways and functions during EMT that are not detected in pure E or M cells. Importantly, the cell line identified EMT-dynamic genes were found to be present in lung cancer patient tissues and associated with patient outcomes. In summary, our study suggests that in vitro identified EMT-dynamic genes capture elements of gene EMT expression dynamics at the patient level. Measurement of EMT dynamic genes, as opposed to E or M only, is potentially useful in future efforts aimed at classifying patient's responses to treatments based on the EMT dynamics in the tissue.

NOTES

About IAIBM

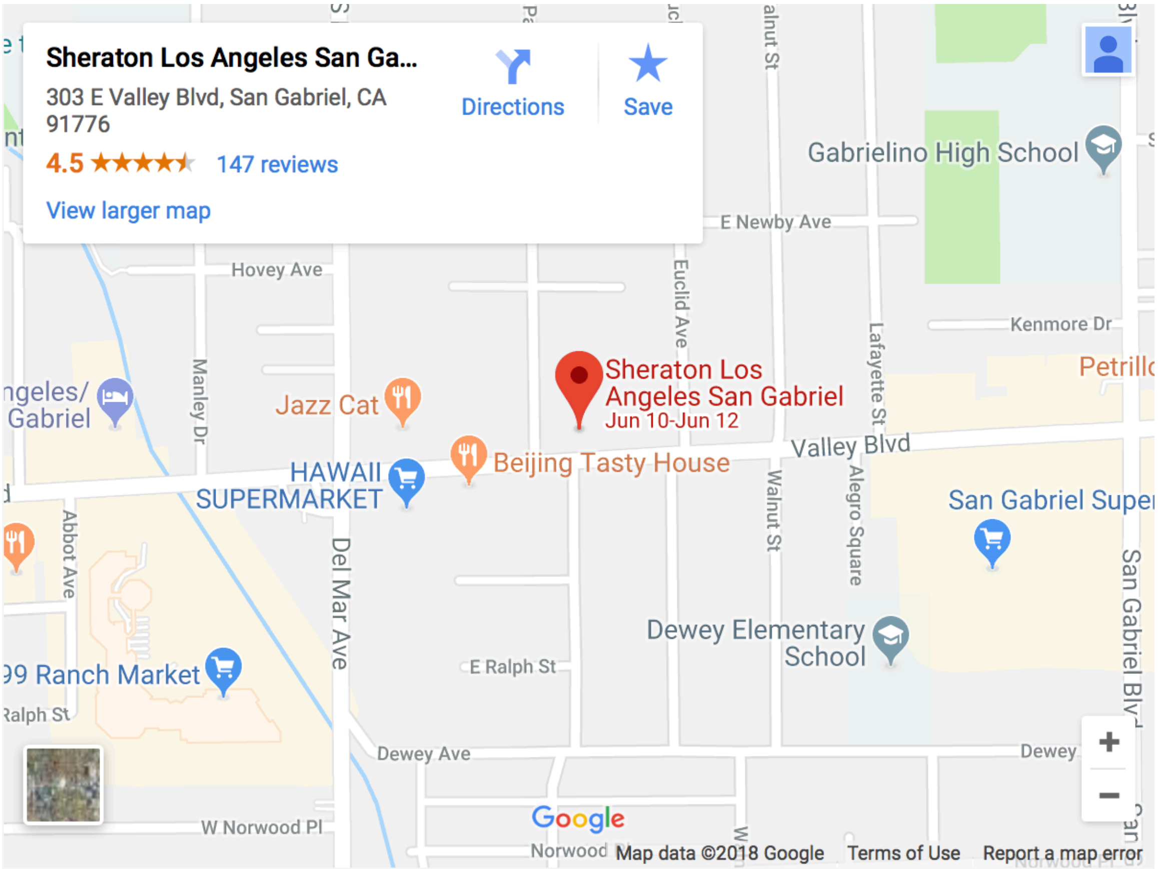
IAIBM

The International Association for Intelligent Biology and Medicine (IAIBM) is a non-profit organization. It was formed on January 19, 2018. Its mission is to promote the intelligent biology and medical science, including bioinformatics, systems biology, and intelligent computing, to a diverse background of scientists, through member discussion, network communication, collaborations, and education.

Conference Location

Sheraton Los Angeles San Gabriel
303 East Valley Boulevard San Gabriel, California 91776

The Sheraton Los Angeles San Gabriel is an ideal setting for successful scientific meetings and inspiring social events. More than 15,000 square feet of function space is available within the hotel. All venues are well equipped with state-of-the-art technology, and the outdoor pool deck can host private receptions for up to 80 people. The hotel offers several onsite dining options, including: Ba Shu Feng (Sichuan style cuisine), EST. Prime Steakhouse, and all day dining in the lobby lounge. In addition, there are over 50 Asian and Western style restaurants within walking distance of the hotel.



Sheraton Los Angeles San Ga...

303 E Valley Blvd, San Gabriel, CA 91776

4.5 ★★★★★ 147 reviews

View larger map

Airport Information

Los Angeles Intercontinental Airport (LAX) - located 30 miles from Hotel, 40-60 minute drive time. Approximate one-way Uber/Lyft fare: $30-50.
Hollywood Burbank Airport (BUR) - located 16 miles from Hotel, 20-30 minute drive time. Approximate one-way Uber/Lyft fare: $15-30.
Transportation for students - make reservation here or Call 1-800-258-3826
From LAX: ~$30 for Shared-Ride Van (60min riding time)

Hotel Information

Sheraton Los Angeles San Gabriel
303 E Valley Blvd, San Gabriel, CA 91776 1-626-639-0300
Situated in the heart of San Gabriel.

Special Acknowledgments

We are grateful for the help from the following volunteers:

Abolfazl Doostparast
Alice Nono Djotsa
Colleen E. Bersano
Guangsheng Pei
Jasmine Zhou
Jessica Li
Jie Ren
Lana Garmire
Luca Giancardo
Marcos Hernandez
Mary Same
Peipei Ping
Qian Liu
Rui Liu
Ruochen Jiang
Saurav Mallik
Shuo Li
Xinzhou Ge
Yi Xing
Yiling Chen
Youping Deng
Yulin Dai

MANY THANKS TO OUR SPONSORS!

IAIBM
International Association for
Intelligent Biology and Medicine

UTHealth
The University of Texas
Health Science Center at Houston
School of Biomedical
Informatics

CPH
Center for
Precision Health

illumina®

组 NextOmics 未来组

Humanizing Genomics
macrogen

UTHealth®
The University of Texas
Health Science Center at Houston
Data Science and Informatics Core
for Cancer Research

Novogene

IBM®

BGI
GLOBAL GENOMIC SERVICES

NSF
National Science Foundation
WHERE DISCOVERIES BEGIN

UTHealth®
The University of Texas
Health Science Center at Houston

UTHealth

Established in 1972 by The University of Texas System Board of Regents, The University of Texas Health Science Center at Houston (UTHealth) is Houston's Health University and Texas' resource for health care education, innovation, scientific discovery and excellence in patient care. The most comprehensive academic health center in The UT System and the U.S. Gulf Coast region, UTHealth is home to schools of biomedical informatics, biomedical sciences, dentistry, nursing and public health and the John P. and Kathrine G. McGovern Medical School. UTHealth includes The University of Texas Harris County Psychiatric Center and a growing network of clinics throughout the region. The university's primary teaching hospitals include Memorial Hermann-Texas Medical Center, Children's Memorial Hermann Hospital and Harris Health Lyndon B. Johnson Hospital. As a comprehensive health science university, the mission of The University of Texas Health Science Center at Houston is to educate health science professionals, discover and translate advances in the biomedical and social sciences, and model the best practices in clinical care and public health.

CPH
Center for
Precision Health

Center for Precision Health

The Center for Precision Health (CPH) is a joint enterprise bridging School of Biomedical Informatics (SBMI) and School of Public Health (SPH) at UTHealth. CPH was established in January 2016, with the funds approved by the Texas Legislature to enhance biomedical and health informatics education and research in the State of Texas in the era of big data and precision medicine. Dr. Zhao was recruited from Vanderbilt University Medical Center to serve as its founding director. A synergizing entity, CPH will build upon the core strengths of SBMI (i.e., informatics programs, centers and clinical resources) and SPH (i.e., cohort-based research, statistical genomics, computational biology, environmental, behavioral and policy research programs) and develop both independent and collaborative research programs, as well as precision health resources, for UTHealth and other Texas Medical Center (TMC) institutions. CPH faculty will actively participate in educational programs and curriculum development at

SBMI and SPH. In addition, CPH faculty will actively promote informatics and population health technology development. CPH currently has four high priority research areas: (1) Population-based Genomics for Precision Health, (2) Cancer Precision Medicine, (3) Translational Bioinformatics, and (4) Smart Clinical Trials.

About Illumina

Illumina is improving human health by unlocking the power of the genome. Our focus on innovation has established us as the global leader in DNA sequencing and array-based technologies, serving customers in the research, clinical, and applied markets. Our products are used for applications in the life sciences, oncology, reproductive health, agriculture, and other emerging segments. To learn more, visit www.illumina.com and follow @illumina.

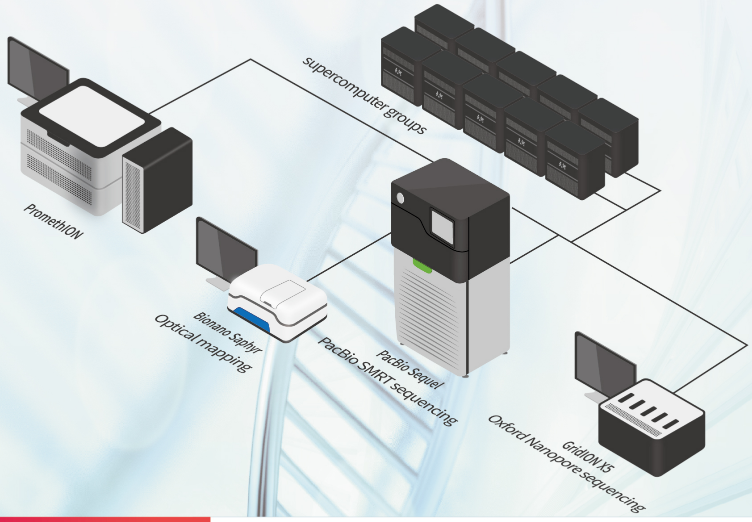
GrandOmics希望组
NextOmics未来组

Nanopore/Pacbio/ BioNano/Hi-C /NGS Service

Grandomics / NextOmics

A World Leading Third Generation Sequencing Genome Center

de novo genome sequencing&assembly	Accurate diagnosis of chromosome structural variation(sv)	Diagnosis of facioscapulohumeral muscular dystrophy (FSHD)	Full-length high resolution HLA typing	Epigenetics

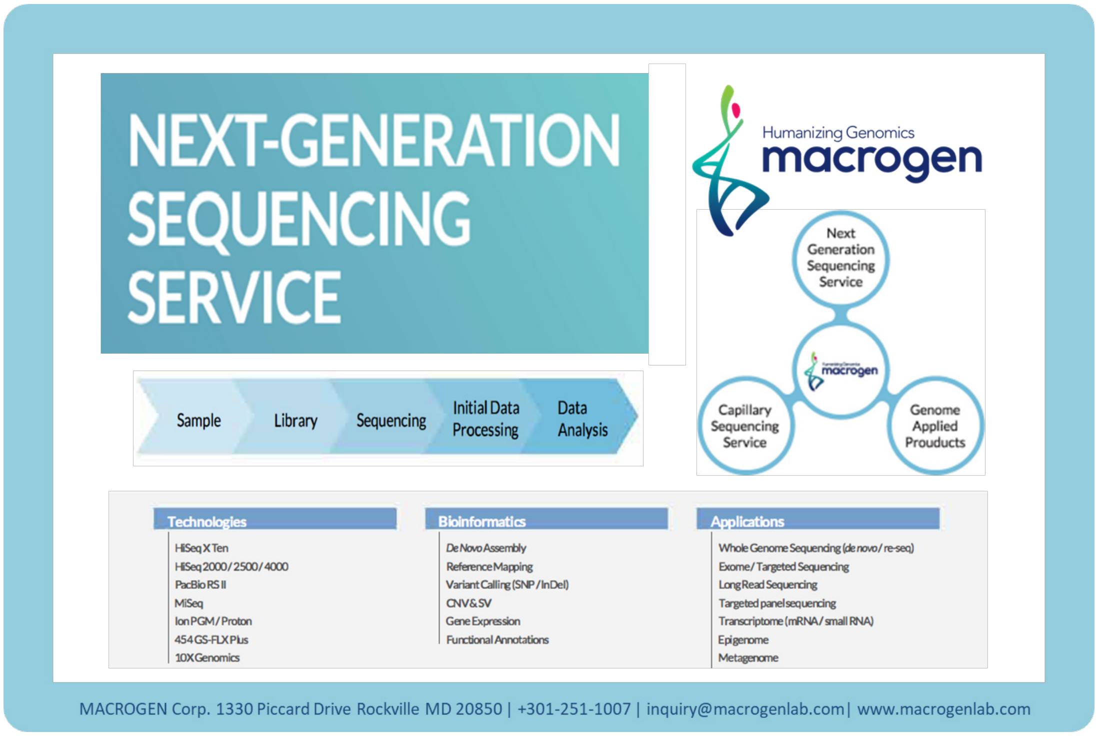


Service Applications:

- Whole genome sequencing
- Whole exome sequencing
- Clinical exome sequencing
- Whole genome CNV sequencing
- Gene panel sequencing
- Structural variation detection
- Mitochondrial sequencing
- Full-length transcriptome sequencing
- Epigenetic sequencing
- Genome database website construction
- BioNano Optical Mapping
- Meta-genomics

Add: Zhongguancun Life Science Park/Wuhan Biolake, China ; | support@nextomics.org | Customer Service Tel : +86 27 8778 2123

www.nextomics.cn
www.grandomics.com



NEXT-GENERATION SEQUENCING SERVICE

Humanizing Genomics macrogen

Technologies

- HiSeq X Ten
- HiSeq 2000/2500/4000
- PacBio RS II
- MiSeq
- Ion PGM/Proton
- 454 GS-FLX Plus
- 10X Genomics

Bioinformatics

- *De Novo* Assembly
- Reference Mapping
- Variant Calling (SNP/InDel)
- CNV & SV
- Gene Expression
- Functional Annotations

Applications

- Whole Genome Sequencing (*de novo*/re-seq)
- Exome/Targeted Sequencing
- Long Read Sequencing
- Targeted panel sequencing
- Transcriptome (mRNA/small RNA)
- Epigenome
- Metagenome

MACROGEN Corp. 1330 Piccard Drive Rockville MD 20850 | +301-251-1007 | inquiry@macrogenlab.com| www.macrogenlab.com

BGI

GLOBAL GENOMIC SERVICES

BGI was founded in 1999 to support the Human Genome Project. Over the years, BGI has grown into a leading genomic services company with global sequencing laboratories based in the US, Europe, Hong Kong, and mainland China.

Our experience in high-throughput Next Generation Sequencing and bioinformatics is second to none and positions BGI uniquely to support academia and pharmaceutical companies with highly reliable genomic data for basic and translational research, as well as pharmaceutical drug development.

To learn more about our services, please contact us at info@bgi-international.com or visit www.bgi.com/us.

IBM Research Healthcare and Life Sciences

Focus: The IBM Research Healthcare and Life Sciences team is dedicated to exploring and developing new methodologies and improving processes for a broad range of health care challenges. From how we can help in the diagnosis of diseases, to managing population health, or a better understanding the human genome, the team blends a broad set of disciplines such as biology, chemistry, data analytics, AI and medicine to pursue their work.

Team: The IBM Research Healthcare and Life Sciences global team blends scientists and experts from a broad set of disciplines such as biology, chemistry, data analytics, AI and cognitive computing, engineering and computer science, as well as medical practitioners and clinicians.

Focus Areas: computational genomics, nanobiology, healthcare informatics, multi-scale modeling, drug discovery, cognitive IoT and devices, microbiome, imaging analytics.

More details: https://www.research.ibm.com/healthcare-and-life-sciences/