2020 International Conference on Intelligent Biology and Medicine (ICIBM 2020)

Hosted by:

The International Association for Intelligent Biology and Medicine (IAIBM)

TABLE OF CONTENTS

Welcome to ICIBM 2020!

On behalf of all our conference committees and organizers, we welcome you to the 2020 International Conference on Intelligent Biology and Medicine (ICIBM 2020), hosted by the International Association for Intelligent Biology and Medicine (IAIBM). We have been closely monitoring and assessing the COVID-19 pandemic situation. After very careful evaluation and discussions, the conference organizers have decided to have the ICIBM 2020 as a virtual conference with complementary registration.

Given the rapid innovations in the fields of bioinformatics, systems biology, intelligent computing, and medical informatics, and their importance to scientific research and biomedical 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 2020 to the next level. In this year, we will have presentations selected from a substantial number of outstanding manuscripts and abstracts that span a diverse array of research subjects. 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 2020's program. We'd like to extend our thanks to our sponsors for making this event possible, including National Science Foundation. 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 2020 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 2020. We look forward to your participation in all our conference has to offer!

Sincerely,

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

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

Xinghua Mindy Shi, PhD
ICIBM Program Co-Chair
Associate Professor,
Dept. of Computer and Information Sciences
College of Science and Technology
Temple University

Li Shen, PhD
ICIBM Program Co-Chair
Professor of Informatics
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine
University of Pennsylvania

ACKNOWLEDGEMENTS

General Chairs

Zhongming Zhao, The University of Texas Health Science Center at Houston, USA
Kai Wang, Children's Hospital of Philadelphia, USA

Steering Committee

Yidong Chen, University of Texas Health Science Center at San Antonio
Kun Huang, Indiana University
Xiaohua Tony Hu, Drexel University
Yi Xing, University of Pennsylvania, USA
Jason Moore, University of Pennsylvania, USA
Sudhir Kumar, Temple University, USA

Program Committee

Xinghua (Mindy) Shi, Co-Chair, Temple University, USA
Li Shen, Co-Chair, University of Pennsylvania, USA
Yongsheng Bai, University of Michigan, USA
Xiao Chang, Children's Hospital of Philadelphia, USA
Bin Chen, Michigan State University, USA
Li Chen, Indiana University School of Medicine, USA
Yidong Chen, UT Health San Antonio, USA
Jianlin Cheng, University of Missouri Columbia, USA
Lijun Cheng, Ohio State University, USA
Zechen Chong, The University of Alabama at Birmingham, USA
Yulin Dai, University of Texas Health Science Center at Houston, USA
Yang Ding, Beijing Institute of Radiation Medicine, China
Abolfazl Doostparast, "University of California, Santa Barbara, USA
Lana Garmire, University of Michigan, USA
Jun-Tao Guo, University of North Carolina at Charlotte, USA
Yan Guo, University of New Mexico, USA
Zhi Han, Indiana University School of Medicicne, USA
Matthew Hayes, Xavier University of Louisiana, USA
Zhe He, Florida State University, USA
Ting Hu, Queen's University, Canada
Heng Huang, University of Pittsburgh & JD Digits, USA
Kun Huang, Indiana University School of Medicine, USA
Tao Huang, Chinese Academy of Sciences, China
Weichun Huang, National Institutes of Health, USA
Yufei Huang, University of Texas at San Antonio, USA
Zhi-Liang Ji, Xiamen University, China
Peilin Jia, UT Health, USA
Victor Jin, UT Health San Antonio, USA
Yufang Jin, University of Texas at San Antonio, USA
Travis Johnson, The Ohio State University, USA

Dokyoon Kim, University of Pennsylvania, USA
Rui Kuang, University of Minnesota Twin Cities, USA
Aimin Li, Xi'an University of Technology, China
Fuhai Li, Washington University School of Medicine, USA
Jia Li, Texas A&M University, USA
Tao Li, Nankai University, China
Li Liao, University of Delaware, USA
Honghuang Lin, Boston University School of Medicine, USA
Ke, Liu, Michigan State University, USA
Xiaoming, Liu, University of South Florida, USA
Yaping, Liu, Cincinnati Children's Hospital Medical Center, USA
Zhandong Liu, Baylor College of Medicine, USA
Shuang Luan, University of New Mexico, USA
Qin Ma, The Ohio State University, USA
Maciej Pietrzak, The Ohio State University, USA
Mirjana Maletic-Savatic, 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
Tabrez Anwar Shamim Mohammad, Greehey Children's Cancer Research Institute, USA
Zhengqing Ouyang, University of Massachusetts, Amherst, USA
Hatice Gulcin Ozer, The Ohio State University, USA
Ranadip Pal, Texas Tech University, USA
Jiang Qian, Johns Hopkins University, USA
Guimin Qin, Xidian University, China
Boshu Ru, Merck & Co. Inc, USA
Jianhua Ruan, The University of Texas at San Antonio, USA
Wei Shao, Indiana University Bloomington, USA
Yang Shen, Texas A&M University, USA
Xiaofeng Song, Nanjing University of Aeronautics & Astronautics, China
Fengzhu Sun, University of Southern California, USA
Zhifu Sun, Mayo Clinic, USA
Wing-Kin Sung, Nuational University of Singapore, Singapore
Manabu Torii, Kaiser Permanente, USA
Fuchiang Rich Tsui, Children's Hospital of Philadelphia and University of Pennsylvania, USA
Jiandong Wang, University of South Carolina, USA
Jiayin Wang, Xi'an Jiaotong University, China
Junbai Wang, Radium Hospital, Norway
Kai Wang, Children's Hospital of Philadelphia, USA
Yufeng Wang, University of Texas at San Antonio, USA
Chaochun Wei, Shanghai Jiao Tong University, China
Lei Wei, Roswell Park Comprehensive Cancer Center, USA
Yingying Wei, The Chinese University of Hong Kong, Hong Kong
Jia Wen, University of North Carolina at Chapel Hill, USA
Chunhua Weng, Columbia University, USA
Huanmei Wu, Indiana University - Purdue University Indianapolis, USA

Yonghui Wu, University of Florida, USA
Junfeng Xia, Anhui University, China
Lei Xie, City University of New York, USA
Hua Xu, The University of Texas School of Biomedical Informatics at Houston, USA
Min Xu, Carnegie Mellon University, USA
Jianhua Xuan, Virginia Tech, USA
Yu Xue, Huazhong University of Science and Technology, China
Jingwen Yan, Indiana University - Purdue University Indianapolis, USA
Sungroh, Yoon, Seoul National University, Korea
Liang, Zhan, University of Pittsburgh, USA
Chi Zhang, Indiana University School of Medicicne, USA
Han Zhang, Nankai University, China
Rui Zhang, Michigan Tech, USA
Shaojie Zhang, University of Central Florida, USA
Wei Zhang, University of Central Florida, USA
Ying Zhang, University of Rhode Island, USA
Zuoyi Zhang, Indiana University, USA
Zhongming Zhao, UT Health, USA
Jim Zheng, U Texas Health Science Houston, USA
Wanding Zhou, Children's Hospital of Philadelphia, USA
Yunyun Zhou, University of Mississippi, USA
Dajiang Zhu, University of Texas at Arlington, USA

Publication Committee
Yan Guo, Co-Chair, University of New Mexico, USA
Wei Zhang, Co-Chair, University of Central Florida, USA

Workshop/Tutorial Committee
Yulin Dai, Chair, University of Texas Health Science Center at Houston

Publicity Committee
Yu Xue, Co-Chair, Huazhong University of Science and Technology, China
Kwangsik Nho, Co-Chair, Indiana University

Award Committee
Jinchuan Xing, Chair, Rutgers University

Trainee Committee
James Havrilla, Chair, Children's Hospital of Philadelphia

Local Organization Committee
Dokyoon Kim, Chair, University of Pennsylvania, USA
Wanding Zhou, Chair, Children's Hospital of Philadelphia

International Conference on Intelligent Biology and Medicine (ICIBM 2020)
Program at-a-glance (August 9-10, 2020)

Sunday, August 9th

Session I: Computational Genomics — Chairs: Jinchuan Xing, Wanding Zhou

Time (US EDT)	Paper ID	Paper Title	Presenter
8:00-8:15am		Introduction	Mindy Shi and Li Shen
8:15-8:30am	13	Enhanced Co-expression Extrapolation (COXEN) Gene Selection Method for Building Anti-cancer Drug Response Prediction Models	Yitan Zhu
8:30-8:45am	15	Deep learning for HGT insertion sites recognition	Chen Li
8:45-9:00am	70	Lilikoi V2: Deep-learning enabled, personalized pathway-based package for diagnosis and prognosis predictions using metabolomics data	Xinying Fang, Yu Liu
9:00-9:15am	27	Comprehensive network modeling of single cell RNA-seq data of human and mouse revealed that transcription regulation program of hematopoiesis is largely conserved across species	Shouguo Gao
9:15-9:30am	31	LongGF: computational algorithm and software tool for fast and accurate detection of gene fusion by long-read transcriptome sequencing	Yu Hu
9:30-9:45am	36	Multivariate genome wide association and network analysis of subcortical imaging phenotypes in Alzheimer's disease	Xianglian Meng
9:45-10:00am	57	A Linear Regression and Deep Learning Approach for Detecting Reliable Genetic Alterations in Cancer using DNA Methylation and Gene Expression Data	Soumita Seth
10:00-10:15am	58	Template-based prediction of protein structure with deep-learning	Haicang Zhang
10:15-10:30am	61	Predicting metabolic pathway membership with deep neural networks by integrating sequential and ontology information	Imam Cartealy
10:30-10:45am	22	Bioinformatics Analysis Revealed Novel 3'UTR Variants Associated with Intellectual Disability	Junmeng (Jasmine) Yang
10:45-11:00am		Wrap up	

Sunday, August 9th

Session II: Biomedical Informatics Chairs: Zhandong Liu, Yunyun Zhou

Time (US EDT)	Paper ID	Paper Title	Presenter
8:00-8:15pm		Introduction	Mindy Shi and Li Shen
8:15-8:30pm	14	Utilizing Deep Learning to Identify Drug Use on Twitter Data	Joseph Tassone
8:30-8:45pm	20	Stress Detection Using Deep Neural Networks	Russell Li
8:45-9:00pm	63	Fingerprint Restoration using Cubic Bezier Curve	Yilin Liu
9:00-9:15pm	43	Unsupervised Phenotyping of Sepsis Using Non-negative Matrix Factorization on Temporal Trends from a Multivariate Panel of Physiological Measurements	Menghan Ding
9:15-9:30pm	48	An Interpretable Risk Prediction Model for Healthcare with Pattern Attention	Sundreen Asad Kamal
9:30-9:45pm	50	Comparing Different Wavelet Transform on Removing Electrocardiogram Baseline Wanders and Special Trends	Fuchiang (Rich) Tsui
9:45-10:00pm	59	In Silico Ranking of Phenolics for Therapeutic Effectiveness on Cancer Stem Cells	Monalisa Mandal
10:00-10:15pm	60	SURF: Identifying and allocating resources duringOut-of-Hospital Cardiac Arrest	Gaurav Rao
10:15-10:30pm	62	Natural Language Processing (NLP) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder	Jacqueline Peng
10:30-10:45pm	71	Identifying risk factors of preterm birth and perinatal mortality using statistical and machine learning approaches	Parth Kothiya
10:45-11:00pm		Wrap up	

Monday, August 10th

Session III: Genomics and Beyond

Chairs: Yan Guo, Wei Zhang

Time (US EDT)	Paper ID	Paper Title	Presenter
8:00-8:15am		Introduction	Mindy Shi and Li Shen
8:15-8:30am	5	Pre-training of deep DNA sequence representation and its application to prediction of deoxyuridine sites for pan-cancer genome analysis	Chaochen Wu
8:30-8:45am	16	Alterations of kidney proteomic profiling revealed the molecular mechanisms of aristolochic acids nephrotoxicity	Jingjing Liu
8:45-9:00am	17	Identifying Patient-Specific Flow Of Signal Transduction Based On Multiple Non-Synonymous Alterations Using Mutational Forks Formalism	Olha Kholod
9:00-9:15am	39	Conditional transcriptional relationships may serve as cancer prognostic markers	Hui Yu
9:15-9:30am	44	The circular RNA expression profile in ovarian serous cystadenocarcinoma revealing the circRNA-miRNA complex regulatory network	Minhui Zhuang
9:30-9:45am	75	An adaptive method of defining negative mutation status for multi-sample comparison using next-generation sequencing	Lei Wei
9:45-10:00am	66	Integrative analysis of histopathological images and chromatin accessibility data for estrogen receptor-positive breast cancer	Siwen Xu
10:00-10:15am	68	A pan-kidney cancer study identifies subtype specific perturbations on pathways with potential drivers in renal cell carcinoma	Xiaohui Zhan
10:15-10:30am	72	Network-based Drug Sensitivity Prediction	Khandakar Tanvir Ahmed
10:30-10:45am	49	Pinpointing miRNA and genes enrichment over trait-relevant tissue network in Genome-Wide Association Studies	Binze Li
10:45-11:00am	3	Annotation and Extraction of Age and Temporally-Related Events from Clinical Histories	Judy Hong
11:00-11:15am		Wrap up	

Monday, August 10th

Session IV: Bioinformatics

Chairs: Xiaoming Liu, Jingwen Yan

Time (US EDT)	Paper ID	Paper Title	Presenter
8:00-8:15pm		Introduction	Mindy Shi and Li Shen
8:15-8:30pm	10	Compositional zero-inflated network estimation for microbiome data	Min Jin Ha
8:30-8:45pm	25	Clinical connectivity map for drug repurposing: using laboratory tests to bridge drugs and diseases	Qianlong Wen
8:45-9:00pm	35	Genome-wide detection of short tandem repeat expansions by long-read sequencing	Qian Liu
9:00-9:15pm	37	Effect of APOE ε4 on Multimodal Brain Connectomic Traits: A Persistent Homology Study	Chenyuan Bian
9:15-9:30pm	46	The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data	Jessica Mar
9:30-9:45pm	51	LDscaff: LD-based scaffolding of de novo genome assemblies	Zicheng Zhao
9:45-10:00pm	53	Deep learning detection of informative features in tau PET for Alzheimer's disease classification	Taeho Jo
10:00-10:15pm	56	Auto3DCryoMap: An Automated Particle Alignment Approach for 3D cryo-EM Density Map Reconstruction	Adil Al-Azzawi
10:15-10:30pm	67	PSC: A Module Detection Method Based on Topology Potential and Spectral Clustering in Weighted Networks and Its Application in Gene Co-expression Module Discovery	Yusong Liu
10:30-10:45pm	54	Identification of miRNA-related tumorigenesis variants and genes in TCGA data sets	Jeff Zhao
10:45-11:00pm		Wrap up	

Paper 3:
Annotation and Extraction of Age and Temporally-Related Events from Clinical Histories
Judy Hong, Anahita Davoudi and Danielle Mowery

In biomedicine, age and temporal information stored within patient clinical histories can provide valuable insights for assessing a patient's disease risk, understanding the progression of their disease phenotype, and determining the therapeutic interventions taken to optimize their outcomes. However, much of this data is stored as unstructured text which cannot be directly used by clinical decision support systems. Clinical Natural Language Processing (NLP), a discipline at the intersection of medicine, computational linguistics, and computer science, can be leveraged to unlock informative clinical events and associated temporal information from clinical notes for downstream utility. As initial steps towards this goal, we created an annotation schema that expands upon existing semantic and temporal representations, conducted an annotation study to validate the utility of our expanded schema, and developed a prototypic, rule-based Named Entity Recognizer (NER) to extract clinical named entities associated with age and temporal information. We observed high inter-annotator agreement (IAA) of >80% for AGE and TIMEX3. The NLP system achieved F1 scores of 86% and 86% for those classes, respectively.

Paper 5:
Pre-training of deep DNA sequence representation and its application to prediction of deoxyuridine sites for pan-cancer genome analysis
Chaochen Wu, Guan Luo and Lei Xie

Background: Genome heterogeneity, drug resistance, and immune escape impose a great challenge in anti-cancer and anti-viral drug development. Deoxyuridine is one of major drivers for the formation of drug resistance and immune escape variants. However, our knowledge on the biological process of dexoyuridine is still limited. Thus, genome-scale characterization and analysis of deoxyuridine may shed new light on cancer and viral evolution. However, because uracils in DNA may be dynamically removed by DNA glycosylases, it is difficult to experimentally detect their potential sites and link them to normal or abnormal cellular processes. With the availability of deoxyuridine sequencing data, machine learning may provide an efficient approach to predicting dynamic deoxyuridine sites.

Results: For the first time, we developed a new deep learning model DeepdUSite which combines DNA sequence pre-training using advanced Natural Language Processing technique BERT with Convolutional Neural Network (CNN) for the prediction of deoxyuridine sites in DNA. Benchmark studies showed that DeepDuSite has a high specificity compared with other state-of-the-art methods, allowing us to reliably detect deoxyuridine sites. Furthermore, we applied DeepDuSite to pan-cancer genome analysis for revealing molecular basis of deoxcyuridine driven processes. We first identified deoxyuridine sites in the upstream CpG islands of genes. We observed that the expression of corresponding genes with the deoxyuridine site correlates with the gene expression level of MBD4, a DNA glycosylase. We then analyzed deoxyuridine sites distribution over APOBECB favored mutations and found that the 5'-UTR mutations distribute unevenly in deoxyuridine positive and negative sites. Finally, we characterized ongoing APOBEC driven mutagenesis, which may provide critical insights into understanding tumor evolution.

Conclusions: DeepDuSite can effectively and reliably predict deoxyuridine sites in DNA. Its application to genome analysis may shed new light on the deoxyuridine related cellular processes and tumor evolution. The pre-trained BERT DNA sequence model can be applied to the prediction of other coding and non-coding DNA functions. The code and data are available at https://github.com/yuanxiaoheben/DeepDuSite

Research 6:
Red Panda: A novel method for detecting variants in single-cell RNA sequencing
Adam Cornish, Shrabasti Roychoudhury, Krishna Sarma, Suravi Pramanik, Kishore Bhakat, Andrew Dudley, Nitish Mishra and Chittibabu Guda

Single-cell sequencing enables us to better understand genetic diseases, such as cancer or autoimmune disorders, which are often affected by changes in rare cells. Currently, no existing software is aimed at identifying single nucleotide variations or micro (1-50bp) insertions and deletions in single-cell RNA sequencing (scRNA-seq) data. Generating high-quality variant data is vital to the study of the aforementioned diseases, among others. In this study, we report the design and implementation of Red Panda, a novel method to accurately identify variants in scRNA-seq data. Variants were called on scRNA-seq data from human articular chondrocytes, mouse embryonic fibroblasts (MEFs), and simulated data stemming from the MEF alignments. Red Panda had the highest Positive Predictive Value at 45.0%, while other tools—FreeBayes, GATK HaplotypeCaller, GATK UnifiedGenotyper, Monovar, and Platypus—ranged from 5.8%-41.53%. From the simulated data, Red Panda had the highest sensitivity at 72.44%. We show that our method provides a novel and improved mechanism to identify variants in scRNA-seq as compared to currently-existing software.

Paper 9:
Deep learning for cancer type classification and driver gene identification
Zexian Zeng, Chengsheng Mao, Andy Vo, Xiaoyu Li, Janna Nugent, Seema A Khan, Susan E Clare and Yuan Luo

We proposed DeepCues, a deep learning model that utilizes convolutional neural networks to unbiasedly derive features from raw cancer DNA sequencing data for disease classification and relevant gene discovery. Using raw whole-exome sequencing as features, germline variants and somatic mutations, including insertions and deletions, were interactively amalgamated for feature generation and cancer prediction. We applied DeepCues to a dataset from TCGA to classify seven different types of major cancers and obtained an overall accuracy of 77.6%. We compared DeepCues to conventional methods and demonstrated a significant overall improvement (p=8.8E-25; 4.5E-25). Strikingly, using DeepCues, the top 20 breast cancer relevant genes we have identified, had a 40% overlap with the top 20 known breast cancer driver genes. These data support DeepCues as a novel method to improve the representational resolution of DNA sequencings and its power in deriving features from raw sequences for cancer type prediction, as well as the discovery of cancer relevant genes

Paper 10:
Compositional zero-inflated network estimation for microbiome data
Min Jin Ha, Junghi Kim, Jessica Galloway-Pena, Kim-Anh Do and Christine Peterson

Background: The estimation of microbial networks can provide important insight into the ecological relationships among the organisms that comprise the microbiome. However, there are a number of critical statistical challenges in the inference of such networks from high-throughput data. Since the abundances in each sample are constrained to have a fixed sum and there is incomplete overlap in microbial populations across subjects, the data are both compositional and zero-inflated.

Results: We propose the COmpositional Zero-Inflated Network Estimation (COZINE) method for inference of microbial networks which addresses these critical aspects of the data while maintaining computational scalability. COZINE relies on the multivariate Hurdle model to infer a sparse set of conditional dependencies which reflect not only relationships among the continuous values, but also among binary indicators of presence or absence and between the binary and continuous representations of the data. Our simulation results show that the proposed method is better able to capture various types of microbial relationships than existing approaches. We demonstrate the utility of the method with an application to understanding the oral microbiome network in a cohort of leukemic patients.

Conclusions: Our proposed method addresses important challenges in microbiome network estimation, and can be effectively applied to discover various types of dependence relationships in microbial communities. The procedure we have developed, which we refer to as COZINE, is available online at https://github.com/MinJinHa/COZINE.

Paper 13:
Enhanced Co-expression Extrapolation (COXEN) Gene Selection Method for Building Anti-cancer Drug Response Prediction Models
Yitan Zhu, Thomas Brettin, Yvonne Evrard, Fangfang Xia, Alexander Partin, Maulik Shukla, Hyunseung Yoo, James Doroshow and Rick Stevens

Background: Cancer is a heterogeneous disease in that patients with the same cancer histology may respond differently to a treatment. Accurate prediction of tumor response to a drug treatment is of paramount importance for precision oncology. The co-expression extrapolation (COXEN) approach has been successfully used in multiple studies to select genes for predicting the response of tumor cells to a specific drug. Here, we extend the COXEN approach to select genes that are predictive of the efficacies of multiple drugs for building general drug response prediction models that are not specific to a particular drug.

Results: We have implemented two methods to select predictive genes. The first method ranks the genes according to their prediction power for each individual drug and then takes a union of top predictive genes of all the drugs. The second method uses a linear regression model to evaluate the prediction power of a gene for all the drugs with an adjustment for the efficacy difference between drugs. Among the predictive genes, the algorithm further selects genes whose co-expression patterns are well preserved between cancer cases with drug response data available and cancer cases for which drug response needs to be predicted. To test the proposed method, we trained LightGBM regression models based on the selected genes for predicting the area under the dose response curve on two benchmark in vitro drug screening datasets. We compared the performance of prediction models built based on the genes selected by the enhanced COXEN approach to that of models built on randomly picked genes. Models built with the enhanced COXEN approach always presented a statistically significantly (p-values ≤ 0.05) improved prediction performance.

Conclusions: Our results show the benefit of using the enhanced COXEN approach for building general drug response prediction models and demonstrate the selected genes can dramatically increase the prediction power of gene expression data.

Paper 14:
Utilizing Deep Learning to Identify Drug Use on Twitter Data
Joseph Tassone, Peizhi Yan, Mackenzie Simpson, Chetan Mendhe, Vijay Mago and Salimur Choudhury

Background: The collection and examination of social media has become a useful mechanism for studying the mental activity and behavior tendencies of users. Through the analysis of a collected set of Twitter data, a model will be developed for predicting positively referenced, drug-related tweets. From this, trends and correlations can be determined.

Methods: Twitter social media tweets and attribute data were collected and processed using topic pertaining keywords, such as drug slang and use-conditions (methods of drug consumption). Potential candidates were preprocessed resulting in a dataset 3,696,150 rows. The predictive classification power of multiple methods was compared including SVM, XGBoost, and CNN-based classifiers. For the latter, a deep learning approach was implemented to screen and analyze the semantic meaning of the tweets.

Results: To test the predictive capability of the model, SVM and XGBoost were first employed. The results calculated from the models respectively displayed an accuracy of 59.33% and 54.90%, with AUC's of 0.87 and 0.71. The values show a low predictive capability with little discrimination. Conversely, the CNN-based classifiers presented a significant improvement, between the two models tested. The first was trained with 2,661 manually labeled samples, while the other included synthetically generated tweets culminating in 12,142 samples. The accuracy scores were 76.35% and 82.31%, with an AUC of 0.90 and 0.91. Using association rule mining in conjunction with the CNN-based classifier showed a high likelihood for keywords such as "smoke", "cocaine", and "marijuana" triggering a drug-positive classification.

Conclusion: Predictive analysis without a CNN is limited and possibly not useful. Attribute-based models presented little predictive capability and were not suitable for analyzing this type of data. The semantic meaning of the tweets needed to be utilized, giving the CNN-based classifier an advantage over other solutions. Additionally, commonly mentioned drugs had a level of correspondence with frequently used illicit substances, proving the practical usefulness of this system. Lastly, the synthetically generated set provided increased scores, improving the predictive capability.

Paper 15:
Deep learning for HGT insertion sites recognition

Chen Li, Jiaxing Chen and Shuaicheng Li

Background: Horizontal Gene Transfer (HGT) refers to the sharing of genetic materials between distant species that are not in a parent-offspring relationship. The HGT insertion sites are important to understand the HGT mechanisms. Recent studies in main agents of HGT, such as transposon and plasmid, demonstrate that insertion sites usually holds specific sequence features. This motivates us to find a method to infer HGT insertion sites according to sequence features.

Results: In this paper, we propose a deep residual network, DeepHGT, to recognize HGT insertion sites. To train DeepHGT, we extracted about 1.55 million sequence segments as training instances from 262 metagenomic samples, where the ratio between positive instances and negative instances is aboout 1:1. These segments are randomly partitioned into three subsets: 80% of them as the training set, 10% as the validation set, and the remaining 10% as test set. The training loss of DeepHGT is 0.4163 and the validation loss is 0.423. On the test set, DeepHGT has achieved area under curve (AUC) value of 0.8782. Furthermore, in order to further evaluate the generalization of DeepHGT, we constructed an independent test set containing of 689,312 sequence segments from another 147 gut metagenomic samples. DeepHGT has achieved the AUC value of 0.8428, which approaches the previous test AUC value. As a comparison, the gradient boosting classifier model implemented in PyFeat achieve AUC value of 0.694 and 0.686 on the above two test sets, respectively. Furthermore, DeepHGT could learn discriminant sequence features; for exmaple, DeepHGT has learned sequence pattern of palindromic subsequences as significantly (P-value=0.0182) local feature. Hence, DeepHGT is a reliable model to recognize HGT insertion site.

Conclusion: DeepHGT is the first deep learning methodology that can directly recognize HGT insertion sites on genomes according to sequence pattern.

Paper 16:
Alterations of kidney proteomic profiling revealed the molecular mechanisms of aristolochic acids nephrotoxicity

Jingjing Liu, Wei Dong, Jing Wu, Jinqiang Xia, Shaofei Xie and Xiaofeng Song

Aristolochic acids (AAs), nephrotoxic components of herbs, has been previously demonstrated causing DNA damage by forming DNA-AAs adducts. However, the alterations of kidney proteome profiles underlying AAs nephrotoxicity remain elusive. In the present study, the proteomics analysis of the kidney tissues from I/R rats after AAs treatment was performed by a shotgun proteomic approach coupled with LC-MS/MS technology. A total of 1475 and 1464 proteins were identified and quantified in control and AAs dosed groups. 88 proteins were significantly dysregulated in comparison between AAs verse I/R. Epithelial cell differentiation, cell-cell adhesion, protein homodimerization were greatly affected due to AAs exposure. Based on the relative proteome quantification results, four differentially expressed proteins, cytochrome P450 26B1 isoform X1, ganglioside GM2 activator, 60S ribosomal protein L21 and Vimentin could be proteic biomarkers of aristolochic acid nephropathy (AAN).

Paper 17
Identifying Patient-Specific Flow Of Signal Transduction Based On Multiple Non-Synonymous Alterations Using Mutational Forks Formalism

Olha Kholod, Johnathan Mitchem, Dong Xu and Dmitriy Shin

Background: Identification of patient-specific flow of signal transduction due to multiple non-synonymous mutations is critical in light of improving patient outcomes in cancer cases. However, accurate estimation of kinetic parameters for deterministic modeling of mutational effects in individual patient cases remains an open problem. While probabilistic pathway topology (PT) methods are gaining an interest in the scientific community, the overwhelming majority of them do not account for network perturbation effects from multiple alterations. Here, we extend our recently introduced Mutational Forks formalism to infer patient-specific flow of signal transduction based on multiple non-synonymous mutations.

Results: We have comprehensively characterized six mutational forks. The number of mutated nodes ranged from one to four depending on the topological characteristics of a fork. Transitional confidences (TCs) have been computed for every possible combination of non-synonymous alterations in the fork. The performed analysis demonstrated a capability of the Mutational Forks formalism to follow a biologically explainable logic in the identification of high likelihood signaling routes. The findings have been largely supported by the evidence from biomedical literature. The formalism has a great chance to enable an assessment of patient-specific flow without knowledge of kinetics, by leveraging mutational information of multiple non-synonymous alterations to adjust the transitional likelihoods that are solely based on the canonical view of a disease.

Conclusions: We conclude that the extension of the developed formalism can be employed in a human-in-the-loop setting as hypothesis-generation tool for precision medicine applications.

Paper 19
intePareto: An R package for integrative analyses of RNA-Seq and ChIP-Seq data

Yingying Cao, Simo Kitanovski and Daniel Hoffmann

Background: RNA-Seq, the high-throughput sequencing (HT-Seq) of mRNAs, has become an essential tool for characterizing gene expression differences between different cell types and conditions. Gene expression is regulated by several mechanisms, including epigenetically by post-translational histone modifications which can be assessed by ChIP-Seq (Chromatin Immuno-Precipitation Sequencing). As more and more biological samples are analyzed by the combination of ChIP-Seq and RNA-Seq, the integrated analysis of the corresponding data sets becomes, theoretically, a unique option to study gene regulation. However, technically such analyses are still in their infancy.

Results: Here we introduce intePareto, a computational tool for the integrative analysis of RNA-Seq and ChIP-Seq data. With intePareto we match RNA-Seq and ChIP-Seq data at the level of genes, perform differential expression analysis between biological conditions, and prioritize genes with consistent changes in RNA-Seq and ChIP-Seq data using Pareto optimization.

Conclusion: intePareto facilitates comprehensive understanding of high dimensional transcriptomic and epigenomic data. Its superiority to a naive differential gene expression analysis with RNA-Seq and available integrative approach is demonstrated by analyzing a public dataset.

Paper 20:
Stress Detection Using Deep Neural Networks

Russell Li and Zhandong Liu

Background: Over 70% of Americans regularly experience stress. Chronic stress results in cancer, cardiovascular disease, depression, and diabetes, and thus is deeply detrimental to physiological health and psychological wellbeing. Developing robust methods for the rapid and accurate detection of human stress is of paramount importance.

Methods: Prior research has shown that analyzing physiological signals is a reliable predictor of stress. Such signals are collected from sensors that are attached to the human body. Researchers have attempted to detect stress by using traditional machine learning methods to analyze physiological signals. Results, ranging between 50% to 90% accuracy, have been mixed. A limitation of traditional machine learning algorithms is the requirement for hand-crafted features. Accuracy decreases if features are misidentified. To address this deficiency, we developed two deep neural networks: a 1-dimensional (1D) convolutional neural network and a multilayer perceptron neural network. Deep neural networks do not require hand-crafted features but instead extract features from raw data through the layers of the neural networks. The deep neural networks analyzed physiological data collected from chest-worn and wrist-worn sensors to perform two tasks. We tailored each neural network to analyze data from either the chest-worn (1D convolutional neural network) or wrist-worn (multilayer perceptron neural network) sensors. The first task was a binary classification for stress detection, in which the networks differentiated between stressed and non-stressed states. The second task was 3-class classification for emotion classification, in which the networks differentiated between baseline, stressed, and amused states. The networks were trained and tested on publicly available data collected in previous studies.

Results: The deep convolutional neural network achieved 99.80% and 99.55% accuracy rates for binary and 3-class classification, respectively. The deep multilayer perceptron neural network achieved 99.65% and 98.38% accuracy rates for binary and 3-class classification, respectively. The networks' performance exhibited significant improvement over past methods that analyzed physiological signals for both binary stress detection and 3-class emotion classification.

Conclusions: We demonstrated the potential of deep neural networks for developing robust, continuous, and noninvasive methods for stress detection and emotion classification, with the end goal of improving the quality of life.

Paper 22:
Bioinformatics Analysis Revealed Novel 3'UTR Variants Associated with Intellectual Disability

Junmeng Yang, Anna Liu and Yongsheng Bai

MicroRNAs (or miRNAs) are short nucleotide sequences (~ 17-22 bp long), playing important roles in gene regulation and could be causative of diseases through targeting genes on the 3'untranslated regions (UTRs). Variants located in genomic regions might have different biological consequences in changing gene expression. Exonic variants (e.g. coding variant and 3'UTR variant) are often causative of diseases due to their influence on gene product. Variants harbored in the 3'UTR region where miRNAs perform their targeting function could potentially alter the binding relationships for target pairs.

We employed the database of microRNA Target Site SNVs (dbMTS) to discover novel single nucleotide variants (SNVs) within miRNA-mRNA targeting pairs that we gathered from published studies. We have identified a total of 183 SNVs for the 114 pairs we have selected. Detailed examination of the three genes with their identified variants that are exclusively located in the 3'UTR through bioinformatics analysis indicated their association with intellectual disability (ID). Our result showed an exceptionally high expression of GPR88 in brain tissues based on GTEx gene expression data, while WNT7A expression data were relatively high in brain tissues when compared to other tissues. Motif analysis for 3'UTR region of WNT7A showed that five identified variants were well-conserved across three species (human, mouse, and rat); the motif that contains the variant identified in GPR88 is significant at the level of the 3'UTR of human genome. Studies of pathways, protein-protein interactions, as well as relations to diseases further suggest potential association with intellectual disability of our discovered SNVs. Our results demonstrated that 3'UTR variants could change target interactions of miRNA-mRNA pairs in the context of their association with ID. We plan to automate the methods through developing a bioinformatics pipeline for identifying novel 3'UTR SNVs harbored by miRNA-targeted genes in the future.

Paper 24:
ES-ARCNN: Predicting enhancer strength by using data augmentation and residual convolutional neural networks

Mario A Flores, Yufei Huang and Tinghe Zhang

Background: Enhancers are non-coding DNA sequences bound by proteins called transcription factors. They function as distant regulators of gene transcription and participate in the development and maintenance of cell types and tissues. Since experimental validation of enhancers is a challenging task, bioinformatics tools have been developed to predict enhancers and their strength using sequence-based methods. However, most of these tools still lack good performance. Here, we developed a model that is able to improve the prediction of enhancer strength through increased sequence variability induced by data augmentation and the employment of a residual convolutional network.

Methods: We present a method to predict Enhancers Strength by using Augmented data and Residual Convolutional Neural Network (ES-ARCNN). The method classifies enhancers as weak, strong, or non-enhancer. To train ES-ARCNN, we used two data augmentation strategies, including reverse component and random mutations (i.e., substitutions, deletions, and insertions) to previously identified enhancers to enlarge a previously identified dataset of enhancers. We further employed a residual convolutional neural network and trained it using the augmented dataset.

Results: We tested ES-ARCNN's performance by a 5-fold cross-validation. We obtained 65.12±0.74% accuracy, which is higher than other state-of-art methods. We further tested ES-ARCNN on an independent dataset and obtained 93% accuracy, representing >30% improvement over all existing algorithms. To study the differences between strong and weak enhancers, we applied transcription factor binding sites (TFBSs) enrichment analysis and found that strong enhancers harbor 2.2-fold higher density of TFBS compared to weak enhancers.

Conclusions: The data augmentation strategies together with the use of a residual convolutional network effectively improved the prediction of enhancer strength. Previously, enhancer strength has been related to distinct levels of biological activities and regulatory effects on target genes. Here we showed that from the mechanistic perspective, enhancer strength is associated with a higher density of important TFBSs in a tissue. Our sequence-based method improves the prediction of strong/weak enhancers by adding diversity to previously identified enhancers. A user-friendly web-application is also provided (http://compgenomics.utsa.edu/ES-ARCNN/).

Paper 25:
Clinical connectivity map for drug repurposing: using laboratory tests to bridge drugs and diseases

Qianlong Wen, Ruoqi Liu and Ping Zhang

Background: Drug repurposing, the process of identifying additional therapeutic uses for existing drugs, has attracted increasing attention from both the pharmaceutical industry and the research community. Many existing computational drug repurposing methods rely on preclinical data (e.g., chemical structures, drug targets), resulting in translational problems for clinical trials.

Methods: In this study, we propose a clinical connectivity map framework for drug repurposing by leveraging laboratory tests to analyze complementarity between drugs and diseases. We establish clinical drug effect vectors (i.e., drug-laboratory test associations) by applying a continuous self-controlled case series model on a longitudinal electronic health record data. We establish clinical disease sign vectors (i.e., disease-laboratory test associations) by applying a Wilcoxon rank sum test on a large-scale national survey data. Finally, we compute a repurposing possibility score for each drug-disease pair by applying a dot product-based scoring function on clinical disease sign vectors and clinical drug effect vectors.

Results: We comprehensively evaluate 392 drugs for 6 important chronic diseases (include asthma, coronary heart disease, congestive heart failure, heart attack, type 2 diabetes, and stroke). We discover not only known associations between diseases and drugs, but also many hidden drug-disease associations. For example, clopidogrel and alendronate may be repurposed as candidate drugs for diabetes and cardiovascular diseases respectively. Moreover, we are able to explain the predicted drug-disease associations via the corresponding complementarity between laboratory tests of drug effect vectors and disease sign vectors.

Conclusion: The proposed clinical connectivity map framework uses laboratory tests from electronic clinical information to bridge drugs and diseases, which is explainable and has better translational power than existing computational methods. Experimental results demonstrate the effectiveness of the proposed framework and suggest that our method could help identify drug repurposing opportunities, which will benefit patients by offering more effective and safer treatments.

Paper 27:
Comprehensive network modeling from single cell RNA sequencing of human and mouse reveals well conserved transcription regulation of hematopoiesis

Shouguo Gao, Zhijie Wu, Xingmin Feng, Sachiko Kajigaya, Xujing Wang and Neal Young

Background: Presently, there is no comprehensive analysis of the transcription regulation network in hematopoiesis.

Methods: We used single-cell RNA sequencing to profile bone marrow from human and mouse, and inferred transcription regulatory networks in each species in order to characterize transcriptional programs governing hematopoietic stem cell differentiation.

Results: We designed an algorithm for network reconstruction to conduct comparative transcriptomic analysis of hematopoietic gene co-expression and transcription regulation in human and mouse bone marrow cells. Co-expression network connectivity of hematopoiesis-related genes was well conserved between mouse and human. The co-expression network showed “small-world” and “scale-free” architecture. The gene regulatory network formed a hierarchical structure, and hematopoiesis transcription factors localized to the hierarchy’s middle level.

Conclusions: Transcriptional gene regulatory networks are well conserved between human and mouse. The hierarchical organization of transcription factors may provide insights into hematopoietic cell lineage commitment, and to signal processing, cell survival and disease initiation.

Paper 30:
Genetic-Based Hypertension Subtype IdentificationUsing Informative SNPs

Yuanjing Ma, Hongmei Jiang, Sanjiv J Shah, Donna Arnett, Marguerite R Irvin and Yuan Luo

In this work, we proposed a process to selectinformative genetic variants for identifying clinically meaningfulsubtypes of hypertensive patients. We studied 1258 AfricanAmerican (AA) and 1270 Caucasian hypertensive participantsenrolled in the Hypertension Genetic Epidemiology Network(HyperGEN) study and analyzed each race-based group sepa-rately. All study participants underwent GWAS and echocardiography. We applied a variety of statistical methods andfiltering criteria, including generalized linear models, F statistics,burden tests, deleterious variant filtering, and others to selectthe most informative hypertension-related genetic variants. Weperformed an unsupervised learning algorithm non-negativematrix factorization (NMF) to identify hypertension subtypeswith similar genetic characteristics. Kruskal-Wallis tests wereused to demonstrate the clinical meaningfulness of genetic-basedhypertension subtypes. Two subgroups were identified for bothAfrican American and Caucasian HyperGEN participants. Inboth AAs and Caucasians, indices of cardiac mechanics differedsignificantly by hypertension subtypes. African Americans tendto have more genetic variants compared to Caucasians, therefore,using genetic information to distinguish the disease subtypes forthis group of people is relatively challenging, but we were able toidentify two subtypes whose cardiac mechanics have statisticallydifferent distributions using the proposed process. The researchgives a promising direction in using statistical methods to selectgenetic information and identify subgroups of diseases, whichmay inform the development and trial of novel targeted therapies.

Paper 31:
LongGF: computational algorithm and software tool for fast and accurate detection of gene fusion by long-read transcriptome sequencing
Qian Liu, Yu Hu, Andres Stucky, Li Fang, Jiang F. Zhong and Kai Wang

Long-read RNA-Seq techniques can generate reads that encompass a large proportion or the entire mRNA or cDNA molecules, so they are expected to address inherited limitations of short-read RNA-Seq techniques that generate only 50-150bp reads. However, there is a general lack of software tools for gene fusion detection from long-read RNA-seq data, which takes into account of the higher error rates and the possible alignment errors for long-read data. In this study, we proposed a fast computational tool, LongGF, to efficiently detect candidate gene fusion from long-read RNA-seq data, including cDNA sequencing data and direct mRNA sequencing data. We evaluated LongGF on tens of simulated long-read RNA-seq data sets, and demonstrated its superior performance in gene fusion detection. We also tested LongGF on a Nanopore direct mRNA sequencing data set generated on a mixture of 10 cancer cell lines, and found that LongGF reliably detects known gene fusions. Finally, we tested LongGF on a cDNA long-read sequencing data set on cancer, and pinpointed the exact location of a translocation in base resolution, which was further validated by Sanger sequencing. In summary, LongGF will greatly facilitate the discovery of candidate gene fusion events from long-read RNA-Seq data, especially in cancer samples. LongGF is publicly available at https://github.com/WGLab/LongGF.

Paper 34:
Predicting Mortality in Critically Ill Patients with Diabetes Using Machine Learning and Clinical Notes
Jiancheng Ye, Liang Yao, Jiahong Shen, Rethavathi Janarthanam and Yuan Luo

Objective: Diabetes mellitus is a prevalent metabolic disease characterized by chronic hyperglycemia. However, few studies have used predictive modeling to uncover associations between comorbidities on ICU patients and diabetes. This study aimed to use Unified Medical Language System (UMLS) resources, involving machine learning and natural language processing approaches to predict risk of mortality.

Materials and Methods: We conducted a secondary analysis of Medical Information Mart for Intensive Care III (MIMIC-III) data. Different machine learning modeling and natural language processing were applied. Mortality classification was based on the combination of knowledge-guided and rule-based features. UMLS entity embedding and convolutional neural network (CNN) with word embeddings were applied.

Results: Concept Unique Identifiers (CUIs) with entity embeddings are useful to build clinical text representations. Different machine learning modeling and natural language processing were applied. The best configuration yielded a competitive AUC of 0.97.
Discussion and Conclusion: UMLS resources and clinical notes are powerful and important tools to predict mortality in diabetic patients in the critical care setting. Knowledge-guided CNN model is effective (AUC=0.97) for learning hidden features.

Paper 35:
Genome-wide detection of short tandem repeat expansions by long-read sequencing

Qian Liu, Yao Tong and Kai Wang

Short tandem repeat (STR), or "microsatellite", is a tract of repetitive DNA in which certain DNA motifs (typically <10 base pairs) are repeated multiple times. STRs are abundant throughout the human genome, and specific repeat expansions may be associated with human diseases. Long-read sequencing coupled with bioinformatics tools enable the estimation of repeat counts for STRs. However, with the exception of a few well known disease-relevant STRs, the normal range of repeat counts for most STRs in human populations are not well known, preventing the prioritization of STRs that may be associated with human diseases. In this study, we extend a computational tool RepeatHMM to infer normal ranges of 432,604 STRs using 21 human genomes with whole-genome long-read sequencing data, and build a genomic-scale database called RepeatHMM-DB with normal repeat ranges for these STRs. On 13 well-known repeats, the inferred repeat ranges provide good estimation to repeat ranges in prior knowledge. This database, together with a repeat expansion estimation tool such as RepeatHMM, enables genomic-scale scanning of repeat regions in a newly sequenced genome to identify candidates of pathogenic repeat expansions. To demonstrate the usefulness of RepeatHMM-DB, we evaluated the inferred normal repeat range on the CAG repeats of ATXN3 for 20 patients and 5 normal individuals, and we correctly classified each individual. In summary, RepeatHMM-DB is expected to facilitate large-scale prioritization and identification of disease-relevant tandem repeats for patients with undiagnosed diseases that may be caused by repeat expansions. RepeatHMM-DB is incorporated into RepeatHMM and is available at https://github.com/WGLab/RepeatHMM.

Paper 36:
Multivariate genome wide association and network analysis of subcortical imaging phenotypes in Alzheimer's disease

Xianglian Meng, Jin Li, Qiushi Zhang, Feng Chen, Chenyuan Bian, Xiaohui Yao, Jingwen Yan, Zhe Xu, Shannon L. Risacher, Andrew J. Saykin, Hong Liang and Li Shen

Background: Genome-wide association studies (GWAS) have identified many individual genes associated with brain imaging quantitative traits (QTs) in Alzheimer's disease (AD). However single marker level association discovery may not be able to address the underlying biological interactions with disease mechanism.

Results: In this paper, we used the MGAS (Multivariate Gene-based Association test by extended Simes procedure) tool to perform multivariate GWAS on eight AD-relevant subcortical imaging measures. We conducted multiple iPINBPA (integrative Protein-Interaction-Network-Based Pathway Analysis) network analyses on MGAS findings using protein-protein interaction (PPI) data, and identified five Consensus Modules (CMs) from the PPI network. Functional annotation and network analysis were performed on the identified CMs. The MGAS yielded significant hits within APOE, TOMM40 and APOC1

genes, which were known AD risk factors, as well as a few new genes such as LAMA1, XYLB, HSD17B7P2, and NPEPL1. The identified five CMs were enriched by biological processes related to disorders such as Alzheimer's disease, Legionellosis, Pertussis, and Serotonergic synapse.

Conclusions: This study provides novel insights into the molecular mechanism of Alzheimer's Disease and will be of value to novel gene discovery and functional genomic studies.

Paper 37:
Effect of APOE _4 on Multimodal Brain Connectomic Traits: A Persistent Homology Study

Jin Li, Chenyuan Bian, Dandan Chen, Xianglian Meng, Haoran Luo, Hong Liang and Li Shen

Background: Although genetic risk factors and network-level neuroimaging abnormalities have shown effects on cognitive performance and brain atrophy in Alzheimer's disease (AD), little is understood about how apolipoprotein E (APOE) _4 allele, the best known genetic risk for AD, affect brain connectivity before the onset of symptomatic AD. This study aims to investigate APOE _4 effects on brain connectivity from the perspective of multimodal connectome.

Results: Here, we propose a novel multimodal brain network modeling framework and a network quantification method based on persistent homology for identifying APOE _4-related network differences. Specifically, we employ sparse representation to integrate multimodal brain network information derived from both the resting state functional magnetic resonance imaging (rs-fMRI) data and the diffusion tensor imaging (DTI) data. Moreover, persistent homology is proposed to avoid the ad hoc selection of a specific regularization parameter and to capture valuable brain connectivity patterns from the topological perspective. The experimental results demonstrate that our method outperforms the competing methods, and reasonably yields connectomic patterns specific to APOE _4 carriers and non-carriers.

Conclusions: We have proposed a multimodal framework that integrates structural and functional connectivity information for constructing a fused brain network with greater discriminative power. Using persistent homology to extract topological features from the fused brain network, our method can effectively identify APOE _4-related brain connectomic biomarkers.

Paper 38:
An ensemble machine learning method for cancer type classification using whole-exome sequencing mutation
Yawei Li and Yuan Luo

Background: With the improvement of next-generation DNA sequencing technology, it takes a lower cost to collect sample data. More machine learning techniques can be used to help cancer analysis and diagnosis.

Methods: We developed an ensemble machine learning system named performance-weighted-voting model for cancer type classification in 6,249 samples across 14 cancer types. Our ensemble system constituted of five weak classifiers logistic regression, support vector machine, random forest, extreme gradient boosting and multilayer perceptron neural network. We first used the cross-validation to get the predicted results of the five classifiers. The weights of five classifiers are based on the predictive performance by solving linear regression functions. The final predicted probability of the performance-weighted-voting model that belongs to a cancer type can be determined by the summation of each classifier's weight multiply its predicted probability.

Results: Using the somatic mutation number of each gene as the input features, the overall accuracy of the performance-weighted-voting model reached 71.78%, which was significantly higher than other comparison classifiers. In most cancer types, a higher tumor mutation burden can help to promote accuracy. Whereas, some factors such as common developmental origins of different cancer types, steroid hormone signaling and tobacco smoking will reduce the accuracy. Using a subset of mutations as input features according to their types was attempted in comparison to using all somatic mutations but could not achieve higher accuracy. We also tried using the cancer driver genes instead of all genes as the input feature but resulted in reduced accuracy.

Conclusion: Our study has important clinical significance for identifying the origin of cancer, especially for those whose primary cannot be determined. In addition, our model provides a good strategy for applying ensemble machine learning methods in cancer origin classification.

Paper 39:
Conditional transcriptional relationships may serve as cancer prognostic markers
Hui Yu, Limei Wang, Jin Li, Danqian Chen and Yan Guo

While most differential coexpression (DC) methods are bound to quantify a single correlation value for a gene pair under across multiple samples, a newly devised approach under the name Correlation by Individual Level Product (CILP) revolutionarily projects the summary correlation value to individual product correlation values for separate samples. CILP greatly widened DC analysis opportunities by allowing integration of non-compromised statistical methods. Here, we performed a study to verify our hypothesis that conditional relationships, i.e., gene pairs of remarkable differential coexpression, may be sought as quantitative prognostic markers for human cancers. Indeed, by integrating CILP with classical univariate survival analysis, we identified up to 244 conditional gene links as potential prognostic markers in five cancer types. In particular, five prognostic gene links for kidney renal papillary cell carcinoma tended to condense around cancer gene ESPL1, and the transcriptional synchrony between ESPL1 and PTTG1 tended to be elevated in patients of adverse prognosis. Alongside the seeking of prognostic gene links in a pan-cancer setting, we also extended the observation of global trend of correlation loss in more than ten cancer types and empirically proved DC analysis results were independent of gene differential expression in five cancer types.

Paper 40:
Assembling reads improves taxonomic classification of species

Quang Tran and Vinhthuy Phan

Background: Most current metagenomic classifiers and profilers employ short reads to classify, bin and profile microbial genomes that are present in metagenomic samples. Many of these methods adopt techniques that aim to identify unique genomic regions of genomes so as to differentiate them. Because of this, short-read lengths might be suboptimal. Longer read lengths might improve the performance of classification and profiling. However, longer reads produced by current technology tend to have a higher rate of sequencing errors, compared to short reads. It is not clear if the trade-off between longer length versus higher sequencing errors will increase or decrease classification and profiling performance.

Results: We compared performance of popular metagenomic classifiers on short reads and longer reads, which are assembled from the same short reads. When using a number of popular assemblers to assemble long reads from the short reads, we discovered that most classifiers made fewer predictions with longer reads and that they achieved higher classification performance on synthetic metagenomic data. Specifically, across most classifiers, we observed a significant increase in precision, while recall remained the same, resulting in higher overall classification performance. On real metagenomic data, we observed a similar trend that classifiers made fewer predictions. This suggested that they might have the same performance characteristics of having higher precision while maintaining the same recall with longer reads.

Conclusions: This finding has two main implications. First, it suggests that classifying species in metagenomic environments can be achieved with higher overall performance simply by assembling short reads. This suggested that they might have the same performance characteristics of having higher precision while maintaining the same recall as shorter reads. Second, this finding suggests that it might be a good idea to consider utilizing long-read technologies in species classification for metagenomic applications. Current long-read technologies tend to have higher sequencing errors and are more expensive compared to short-read technologies. The trade-offs between the pros and cons should be investigated.

Paper 42:
Characterization of genome-wide association study data reveals spatiotemporal heterogeneity of mental disorders

Yulin Dai, Timothy D. O'brien, Guangsheng Pei, Zhongming Zhao and Peilin Jia

Psychiatric disorders such as schizophrenia (SCZ), bipolar disorder (BIP), major depressive disorder (MDD), attention deficit-hyperactivity disorder (ADHD), and autism spectrum disorder (ASD) are often related to brain development. Both shared and unique biological and neurodevelopmental processes have been reported to be involved in these disorders. In this work, we developed an integrative analysis framework to seek for the sensitive spatiotemporal point during brain development underlying each disorder. Specifically, we first identified spatiotemporal gene co-expression modules for four brain regions three developmental stages (prenatal, birth to 11 years old and older than 13 years), totaling 12 spatiotemporal sites. By integrating GWAS summary statistics and the spatiotemporal co-expression modules, we characterized the risk genes and their co-expression partners for five disorders. As a result, we found that SCZ and BIP were closely clustered in 10 out of the 12 investigated spatiotemporal sites, while ASD appeared to be distantly related to the other four disorders. At the gene level, we identified several genes that were shared among the most significant modules, such as CTNNB1 and LNX1, and a hub gene, ATF2, in multiple modules. Moreover, we pinpointed two spatiotemporal points in the prenatal stage with active expression activities. Further functional analysis of the disorder-related module highlighted the apoptotic signaling pathway in for ASD and the immune-related and cell-cell adhesion function for SCZ, respectively. In summary, our study demonstrated the dynamic changes of disorder-related genes at the network level, shedding light on the spatiotemporal regulation during brain development.

Paper 43:
Unsupervised Phenotyping of Sepsis Using Non-negative Matrix Factorization on Temporal Trends from a Multivariate Panel of Physiological Measurements

Menghan Ding and Yuan Luo

Background: Sepsis is a highly lethal and heterogenous disease. Utilization of unsupervised method may identify novel clinical phenotypes that lead to targeted therapies and improved care.

Methods: Our objective was to derive clinically relevant sepsis phenotypes from a multivariate panel of physiological data using Subgraph-Augmented Non-negative Matrix Factorization (SANMF). We utilized data from Medical Information Mart for Intensive Care III (MIMIC-III) for patients who were admitted into intensive care unit (ICU) with sepsis. Data extracted contains patient demographics, physiological records, sequential organ failure assessment (SOFA) scores, and comorbidities. We applied frequent subgraph mining to extract subgraphs from physiological time series, and performed non-negative matrix factorization over subgraphs to derive patient clusters as phenotypes. Finally, we profiled identified phenotypes over demographics, physiological patterns, disease trajectories, comorbidities and outcomes, and performed functional validations on clinical implications of these phenotypes.

Results: We analyzed a cohort of 5782 patients, derived three novel phenotypes of distinct clinical characteristics and demonstrated their independent association with patient outcomes. Subgroup 1 is the less severe and deadly, and smallest-in-size group (30-day mortality, 17%; n = 1218, 21%), characterized by old age (mean age, 73 years), male majority (male-to-female ratio, 59-41), and complex chronic conditions. Subgroup 2 is the most severe and deadliest, and second-in-size group (30-day mortality, 28%; n = 2036, 35%), characterized by male majority (male-to-female ratio, 60-40), severe organ dysfunction or failure multiplied by a wide range of comorbidities, and uniquely high incidence in coagulopathy and liver disease. Subgroup 3 is the least deadly and largest group (30-day mortality, 10%; n = 2528, 44%), characterized by low age (mean age, 60 years), balanced gender ratio (male-to-female ratio, 50-to-50), least complicated conditions, and uniquely high incidence in neurologic disease. These phenotypes are validated to be independent predictors of clinical outcomes including mortality and length of stay.

Conclusion: Our results suggest that these phenotypes could be used inform targeted therapies based on phenotypic heterogeneity, and design algorithms to detect, monitor, and intervene for sepsis management.

Paper 44:
The circular RNA expression profile in ovarian serous cystadenocarcinoma revealing the circRNA-miRNA complex regulatory network

Minhui Zhuang, Jian Zhao, Jing Wu, Shilong Fu, Ping Han and Xiaofeng Song

Background: Ovarian serous cystadenocarcinoma is one of the most serious gynecological malignancies. Circular RNA (circRNA) is one kind of noncoding RNAs with a covalently closed continuous loop structure. Abnormal circRNA expression might be associated with tumorigenesis, due to its complex biological mechanisms such as functioning as microRNA (miRNA) sponge. However, the circRNA expression profile in ovarian serous cystadenocarcinoma and their association is still unclear. The main purpose of this study is to reveal the circRNA expression profile in ovarian serous cystadenocarcinoma.

Results: We collected six specimens from three patients with ovarian serous cystadenocarcinoma and adjacent normal tissues, and 15092 unique circRNAs were identified in these tissues. About 46% of them were not recorded in the public database. Then we reported 353 differentially expressed circRNAs and the occurrence of oncogenes and tumor-suppressor genes in them. Furthermore, a conjoint analysis with relevant mRNAs was carried out, and consistent changes were found between circRNAs and their homologous mRNAs. In the end, the construction of a circRNA-miRNA network suggested 4 special circRNAs as potential biomarkers.

Conclusions: Our study revealed the circRNA expression profile in the tissues of patients with ovarian serous cystadenocarcinoma. The differential expression of circRNAs was considered to be associated with ovarian serous cystadenocarcinoma in the enrichment analysis, and co-expression analysis with relevant mRNAs and miRNAs showed their latent regulatory network. We also constructed a complex circRNA-miRNA interaction network and then demonstrated the potential function of certain circRNAs in the future diagnosis and treatment.

Paper 46:
The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data.
Laurence de Torrente, Samuel Zimmerman, Masako Suzuki, Maximilian Christopeit, John Greally and Jessica Mar

Background. In genomics, we often assume that continuous data, such as gene expression, follow a specific kind of distribution. However we rarely stop to question the validity of this assumption, or consider how broadly applicable it may be to all genes that make up the transcriptome. Our study investigated the prevalence of non-Normally distributed genes in different tumor types from the Cancer Genome Atlas (TCGA).

Results. Surprisingly, less than 50% of all genes were Normally-distributed, with other distributions including Gamma, Bimodal, Cauchy, and Lognormal were represented. Relevant information about cancer biology was captured by the genes with non-Normal gene expression. When used for classification, the set of non-Normal genes were able to discriminate between cancer patients with poor versus good survival status.

Conclusions. Our results highlight the value of studying a gene's distribution shape to model heterogeneity of transcriptomic data. These insights would have been overlooked when using standard approaches that assume all genes follow the same type of distribution in a patient cohort.

Paper 48:
An Interpretable Risk Prediction Model for Healthcare with Pattern Attention
Sundreen Asad Kamal, Changchang Yin, Buyue Qian and Ping Zhang

Background: The availability of massive amount of data enables the possibility of clinical predictive task. Deep learning methods have achieved promising performance on the task. However, most existing methods suffer from limitations: (i) Some existing methods embed Boolean value medical events (e.g. diagnosis code), but ignore real value medical events (e.g., lab tests and vital signs). (ii) There are lots of missing value for real value events, many methods impute the missing value and then train their models based on the imputed values. The models' performance are highly dependent on imputation accuracy. (iii) Existing interpretable models can illustrate which medical events are conducive to the output results, but are not able to compute medical event patterns' contributions.

Methods: In this study, we propose a new interpretable Pattern Attention model with Value Embedding (PAVE) to predict the disease risks. PAVE takes the embedding of various medical events, their values and the corresponding occurring time as inputs, leverage self-attention mechanism to attend to meaningful medical event patterns to do the risk prediction task. Because only the observed values are embedded into vectors, we don't need to impute the missing values. Moreover, the self-attention mechanism is helpful for the model interpretability, which means the proposed model can output which medical event patterns cause high risks.

Results: We conduct sepsis prediction and mortality prediction experiments on public available dataset MIMIC-III and our proprietary EHR dataset. The experimental results show that PAVE outperforms existing models. Moreover, by analyzing the self-attention weights, our model outputs meaningful medical event patterns related to mortality.

Conclusions: PAVE learns effective medical event representation by incorporating the values and occurring time, which can improve the risk prediction performance. Moreover, the presented self-attention mechanism can not only capture patients' health state information, but also output the contributions of various medical event patterns, which pave the way for interpretable clinical risk predictions.

Paper 49:
Pinpointing miRNA and genes enrichment over trait-relevant tissue network in Genome-Wide Association Studies

Binze Li, Julian Dong, Lulu Shang, Xiang Zhou and Yongsheng Bai

Background: Understanding gene regulation is important but difficult. Elucidating tissue-specific gene regulation mechanism is even more challenging and requires gene co-expression network assembled from protein-protein interaction, transcription factor and gene binding, and post-transcriptional regulation (e.g., miRNA targeting) information. The miRNA binding affinity could therefore be changed by SNP(s) located at the 3' untranslated regions (3'UTR) of the target messenger RNA (mRNA) which miRNA(s) interacts with. Genome-Wide Association Study (GWAS) has reported significant numbers of loci hosting SNPs associated with many traits. The goal of this study is to pinpoint GWAS functional variants located in 3'UTRs and elucidate if the genes harboring these variants along with their targeting miRNAs are associated with genetic traits relevant to certain tissues.

Results: By employing GWAS, MIGWAS, CoCoNet, ANNOVAR, and DAVID bioinformatics software and utilizing the gene expression database (eg. GTEx data), we have identified a list of miNRAs and targeted genes harboring 3'UTR variants, which could contribute to trait-relevant tissue over miRNA-target gene network. Our result demonstrated that strong association exists between traits and tissues, and, in particular, the Primary Biliary Cirrhosis (PBC) trait has the most significant p-value for all 180 tissues among all 43 traits used for this study. We reported SNPs located in 3'UTR regions of genes (SFMBT2, ZC3HAV1, and UGT3A1) targeted by miRNAs for PBC trait and its tissue association network. After performing Gene Ontology (GO) analysis for PBC trait, we have also identified a very important miRNA targeted gene over miRNA-target gene network, PFKL, which encodes the liver subunit of an enzyme.

Conclusion: The non-coding variants identified from GWAS studies are casually assumed to be not critical to translated protein product. However, 3' untranslated regions (3'UTRs) of genes harbor variants that can often change the binding affinity of targeting miRNAs that play important roles in protein translation degree. Our study has shown that GWAS variants could play important roles on miRNA-target gene networks by contributing to the association between traits and tissues. Our analysis expands our knowledge on trait-relevant tissue network and paves the way for future human disease studies.

Paper 50:
Comparing Different Wavelet Transform on Removing Electrocardiogram Baseline Wanders and Special Trends

Chao Chen Chen, Fuchiang Rich Tsui and Fuchiang Rich Tsui

Motivation: Electrocardiogram (ECG) signal, an important indicator for heart problems, is commonly corrupted by a low-frequency baseline wander (BW) artifact, which may cause interpretation difficulty or inaccurate analysis. Unlike current state-of-the-art approach using band-pass filters, wavelet transform can accurately capture both time and frequency information of a signal. However, extant literature is limited in applying wavelet transform (WT) for baseline wander removal. In this study, we aimed to evaluate 5 wavelet families with a total of 14 wavelets for removing ECG baseline wanders from a semi-synthetic dataset.

Methods: We created a semi-synthetic ECG dataset based on a public QT Database on Physionet repository with ECG data from 105 patients. The semi-synthetic ECG dataset comprised ECG excerpts from the QT database superimposed with artificial baseline wanders. We extracted one ECG excerpt from each of 105 patients, and the ECG excerpt comprised 14 seconds of randomly selected ECG data. 12 baseline wanders were manually generated, including sinusoidal waves, spikes and step functions. We implemented and evaluated 14 commonly used wavelets up to 12 WT levels. The evaluation metric was mean-square-error (MSE) between the original ECG excerpt and the processed signal with artificial BW removed.

Results: Among the 14 wavelets, Daubechies-3 wavelet and Symlets-3 wavelet with 7 levels of WT had best performance, MSE=0.0044. The average MSEs for sinusoidal waves, step, and spike functions were 0.0271, 0.03011, 0.02523 respectively. For artificial baseline wanders with spikes or step functions, wavelet transforms in general had lower performance in removing the BW; however, WTs accurately located the temporal position of an impulse edge.

Conclusions: We found WTs in general accurately removed various baseline wanders. Daubechies-3 and Symlets-3 wavelets performed best. The study could facilitate future real-time processing of streaming ECG signals.

Paper 51:
LDscaff: LD-based scaffolding of de novo genome assemblies
Yingxiao Zhou, Zicheng Zhao, Shuai Wang, Xiuqing Zhang, Changfa Wang and Shuaicheng Li

Genome assembly is fundamental for de novo genome analysis. Hybrid assembly, utilizing short reads, long reads and Hi-C data, increases both contiguity and accuracy. While such approaches require extra costly sequencing efforts, the information provided millions of existed whole-genome sequencing data have not been fully utilized to resolve the task of scaffolding. Genetic recombination patterns in population data indicate non-random association among alleles at different loci, can provide physical distance signals to guide scaffolding.

In this paper, we propose LDscaff for draft genome assembly incorporating linkage disequilibrium information in population data. We evaluated the performance of our method with both simulated data and real data. We simulated scaffolds by splitting the pig reference genome and reassembled them. Gaps between scaffolds were introduced ranging from 0kb to 100kb. The genome misassembly rate is 2.43% when there is no gap. Then we implemented our method to refine the giant panda genome and the donkey genome, which are purely assembled by NGS data. After LDscaff treatment, the resulting panda assembly has scaffold N50 of 3.6 Mb, 2.5 times larger than the original N50 (1.3 Mb). The re-assembled donkey assembly has an improved N50 length of 32.1Mb from 23.8 Mb.

Paper 52:
Prognostic Value and Co-expression Patterns of Metabolic Pathways in Cancers
Yan Guo, Dan Zhang and Ni Xie

Abnormal metabolic pathways have been considered as one of the hallmarks of cancer. While numerous metabolic pathways have been studied in various cancers, the direct link between metabolic pathway gene expression and cancer prognosis has not been established. Using two recently developed bioinformatics analysis methods, we evaluated the prognosis potential of metabolic pathway expression and tumor-vs-normal dysregulations for up to 29 metabolic pathways in 33 cancer types. Results show that increased metabolic gene expression within tumors corresponds to poor cancer prognosis. Meta differential co-expression analysis identified four metabolic pathways with significant global co-expression network disturbance between tumor and normal samples. Differential expression analysis of metabolic pathways also demonstrated strong gene expression disturbance between paired tumor and normal samples. Taken together, these results strongly suggested that metabolic pathway gene expressions are disturbed after tumorigenesis. Within tumors, many metabolic pathways are upregulated for tumor cells to activate corresponding metabolisms to sustain the required energy for cell division

Paper 53:
Deep learning detection of informative features in tau PET for Alzheimer's disease classification

Taeho Jo, Kwangsik Nho, Shannon Risacher and Andrew Saykin

Background: Alzheimer's disease (AD) is the most common type of dementia, characterized by severe cognitive impairment and memory loss. Many clinical trials of therapies for AD have failed, and there is currently no cure, prevention, or treatment for the disease. Biomarkers for early detection and mechanistic understanding of disease course are critical for drug development and clinical trials. Amyloid has been the focus of most biomarker research. Here, we developed a deep learning-based framework to identify informative features for AD classification using tau positron emission tomography (PET) scans that yields AD probability scores based on molecular neuroimaging data.

Methods: We analysed [18F]flortaucipir PET image data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. We first developed an image classifier to distinguish AD from cognitively normal (CN) older adults by training a 3D convolutional neural network (CNN)-based deep learning model on tau PET images (N=132; 66 CN and 66 AD), then applied the classifier to images from individuals with mild cognitive impairment (MCI; N=168). In addition, we applied a layer-wise relevance propagation (LRP)-based model to identify informative features and to visualize classification results. We compared these results with those from whole brain voxel-wise between-group analysis using conventional Statistical Parametric Mapping (SPM12).

Results: The 3D CNN-based classification model of AD from CN yielded an average accuracy of 90.8% based on five-fold cross-validation. The LRP model identified the brain regions in tau PET images that contributed most to the AD classification from CN. The top identified regions included the hippocampus, parahippocampus, thalamus, and fusiform. The LRP results were consistent with those from the voxel-wise analysis in SPM12, showing significant focal AD associated regional tau deposition in the bilateral temporal lobes including the entorhinal cortex. The AD probability scores calculated by the classifier were correlated with brain tau deposition in the medial temporal lobe in MCI participants (r=0.43 for early MCI and r=0.49 for late MCI).

Conclusion: A deep learning framework combining 3D CNN and LRP algorithms can be used with tau PET images to identify informative features for AD classification and may have application for early detection during prodromal stages of AD.

Paper 54:
Identification of miRNA-related tumorigenesis variants and genes in TCGA data sets

Chang Li, Brian Wu, Han Han, Yongsheng Bai and Xiaoming Liu

microRNAs (miRNAs) are a class of small non-coding RNA that can down-regulate their targets by selectively bind to the 3' untranslated region (3'UTR) of most messenger RNAs (mRNAs) in human. Single nucleotide variants (SNVs) located on miRNA target sites (MTS) can disrupt the binding of targeting miRNAs. Anti-correlated miRNA-mRNA pairs between normal and tumor tissues obtained from the Cancer Genome Atlas (TCGA) project can reveal important information behind these SNVs on MTS and their associated oncogenesis. In this study, using previously identified anti-correlated miRNA-mRNA pairs in 15 TCGA cancer types and publicly available variant annotation databases, namely dbNSFP and dbMTS, we identified multiple functional variants and their gene products that could be associated with various types of cancers. Their potential roles in cancers were explored and analyzed. We found two genes from dbMTS and 33 from dbNSFP passing our stringent filtering criteria (e.g. pathogenicity). Specifically, from dbMTS, we identified two genes (BMPR1A and XIAP) that were associated with diseases that increased risk of cancer in patients; From dbNSFP, we identified 33 genes that were likely pathogenic and had a potential causative relationship with cancer. This study provides a novel way of utilizing TCGA data and integrating multiple publicly available databases to explore cancer genomics.

Paper 56:
Auto3DCryoMap: An Automated Particle Alignment Approach for 3D cryo-EM Density Map Reconstruction
Adil Al-Azzawi, Anes Ouadou, Ye Duan and Jianlin Cheng

An essential process for determining protein structures from cyro-EM data is a 3D density map reconstruction. Cryo-EM data generated by electron tomography (ET) contains images for individual protein particles in different orientations and tilted angles. Individual cryo-EM particles can be aligned to reconstruct a 3D density map of a protein structure. However, low contrast and high noise in particle images make it challenging to build 3D density maps at intermediate resolution (1–3_Å). To overcome this problem, we propose a fully automated cryo-EM 3D density map reconstruction approach based on deep learning particle picking. It uses Deep learning Approach to automatically pick the particles from the micrographs and classify them into top view or side-view. A perfect 2D particle mask is fully automatically generated for every single particle. Then, it uses a computer vision image alignment algorithm (image registration) to fully automatically align the particle masks. It calculates the difference of the particle image orientation angles to align the original particle image. Finally, it reconstructs a localized 3D density map between every two single-particle images that have the largest number of corresponding features. The localized 3D density maps are then averaged to reconstruct a final 3D density map. The constructed 3D density map results illustrate the potential to determine the structures of the molecules using few samples of good particles. Also, using the localized particle samples (with no background) to generate the localized 3D density maps can improve the process of the resolution evaluation in experimental maps of cryo-EM. Tested on two widely used datasets, Auto3DCryoMap can reconstruct good 3D density maps using only a few thousand protein particle images, which is much smaller than hundreds of thousands of particles required by the existing methods. We design a fully automated approach for cryo-EM 3D density maps reconstruction based on deep supervised and unsupervised learning approaches (Auto3DCryoMap). Instead of increasing the signal-to-noise ratio by using 2D class averaging, our approach used the perfect 2D mask to produce locally aligned particle images. The experimental results show that the Auto3DCryoMap can accurately align structural particle shapes. Also, it can construct a decent 3D density map from only a few thousand aligned particle images while the existing tools require hundreds of thousands of particle images and reconstructs a better 3D density map. In the future, we plan to extend our methods to reconstruct 3D density maps of particles with irregular and complicated particle shapes.

Paper 57:
An integrative approach for detecting reliable alterations in cancer using linear regression and deep learning for DNA methylation and gene expression data
Saurav Mallik, Soumita Seth, Tapas Bhadra and Zhongming Zhao

DNA methylation change has been useful for cancer biomarker discovery, classification, and potential treatment development. So far, existing methods use either differentially methylated CpG sites or combined CpG sites, namely differentially methylated regions, that can be mapped to genes. However, such methylation signal mapping has limitations.
To address these limitations, in this study, we introduced a combinatorial framework using linear regression, differential expression, deep learning method for correct biological interpretation of DNA methylation data through integrating DNA methylation data and corresponding TCGA gene expression data for uterine cervical cancer data. First, we pre-filtered outliers from the data set and then determined the predicted gene expression value from the pre-filtered methylation data through linear regression. We identified differentially expressed genes (DEGs) by 16,315, Student`s t-test. Then we performed Gene Set Enrichment Analysis (GESA) of DEGs by DAVID software. Furthermore, we applied a well-known deep learning method, ""deepnet"" to classify the cervical cancer label of those DEGs to determine minimum learning error rate through 10-fold cross validation and different parameter setting in "deepnet" tool [e.g., (i) number of epochs, (ii) number of hidden layers and (iii) type of activation function]. We applied a Uterine cervical cancer DNA methylation data (NCBI accession id: GSE30760) which consisted of 27,578 features and 215 samples including 63 cancer and 152 normal samples. After linear regression and differential expression analysis, we obtained 16,315 DEGs having pvalue< 0:05. Next, we performed KEGG pathway and Gene Ontology analysis on those DEGs using FDR < 0:01 through DAVID database. In deep learning analysis, we finally got minimum learning error rate 0.02273 for Uterine cervical cancer which is highly significant (< 0:05).

Conclusions: Our proposed framework integrated linear regression, differential expression, deep learning method to interpret DNA methylation data correctly instead of using single differential methylation analysis or differentially methylated region finding for any kind of cancer or tumor methylation data.

Paper 58:
Template-based prediction of protein structure with deep-learning
Haicang Zhang and Yufeng Shen

Accurate prediction of protein structure is fundamentally important to understand biological function of proteins. Template-based modeling, including protein threading and homology modeling, is a popular method for protein tertiary structure prediction. However, accurate template-query alignment and template selection are still very challenging, especially for the proteins with only distant homologs available. We propose a new template-based modelling method called ThreaderAI to improve protein tertiary structure prediction. ThreaderAI formulates the task of aligning query sequence with template as the classical pixel classification problem in computer vision and naturally applies deep residual neural network in prediction. ThreaderAI first employs deep learning to predict residue-residue aligning probability matrix by integrating sequence profile, predicted sequential structural features, and predicted residue-residue contacts, and then builds template-query alignment by applying a dynamic programming algorithm on the probability matrix. We evaluated our methods both in generating accurate template-query alignment and protein threading. Experimental results show that ThreaderAI outperforms currently popular template-based modelling methods HHpred, CNFpred, and the latest contact-assisted method CEthreader, especially on the proteins that do not have close homologs with known structures. In particular, in terms of alignment accuracy measured with TM-score, ThreaderAI outperforms HHpred, CNFpred, and CEthreader by 56%, 13%, and 11%, respectively, on template-query pairs at the similarity of fold level from SCOPe data. And on CASP13's TBM-hard data, ThreaderAI outperforms HHpred, CNFpred, and CEthreader by 16%, 9% and 8% in terms of TM-score, respectively. These results demonstrate that with the help of deep learning, ThreaderAI can significantly improve the accuracy of template-based structure prediction, especially for distant-homology proteins."

Paper 59:
In Silico Ranking of Phenolics for Therapeutic Effectiveness on Cancer Stem Cells

Monalisa Mandal, Sanjeeb Kumar Sahoo, Priyadarshan Patra, Saurav Mallik and Zhongming Zhao

Background: Cancer stem cells (CSCs) have features such as the ability to self-renew, differentiate into defined progenies and initiate the tumor growth. Treatments of cancer include drugs, chemotherapy and radiotherapy or a combination. However, treatment of cancer by various therapeutic strategies often fail. One possible reason is that the nature of CSCs that have stem-like properties, which make it more dynamic and complex and may cause the therapeutic resistance. Another limitation is the side effects associated with the treatment of chemotherapy or radiotherapy. To explore better or alternative treatment options the current study aims to investigate the natural drug-like molecules that can be used as CSC-targeted therapy. Among various natural products, anticancer potential of phenolics is well established. We collected the 21 phytochemicals from phenolic group and their interacting CSC genes from the publicly available databases. Then a bipartite graph is constructed from the collected CSC genes along with their interacting phytochemicals from phenolic group as other. The bipartite graph is then transformed into weighted bipartite graph by considering the interaction strength between the phenolics and the CSC genes. The CSC genes are also weighted by two scores, namely, DSI (Disease Specificity Index) and DPI (Disease Pleiotropy Index). For each gene. its DSI score reflects the specific relationship with the disease and DPI score reflects the association with multiple diseases. Finally, a ranking technique is developed based on PageRank (PR) algorithm for ranking the phenolics.

Results: We collected 21 phytochemicals from phenolic group and 1118 CSC genes. The top ranked phenolics have been evaluated by their molecular and pharmacokinetics properties and disease association networks. We selected top five ranked phenolics (Resveratrol, Curcumin, Quercetin, EpigallocatechinGallate, and Genistein) for further examination of their oral bioavailability through molecular properties, drug likeness through pharmacokinetic properties, and associated network with CSC genes.

Conclusion: Our PR ranking based approach is useful to rank the phenolics that are associated with CSC genes. Our results suggested some phenolics are potential molecules for CSC related cancer treatment.

Paper 60:
SURF: Identifying and allocating resources duringOut-of-Hospital Cardiac Arrest
Gaurav Rao, Salimur Choudhury, Pawan Lingras, David Savage and Vijay Mago

When an Out-of-Hospital Cardiac Arrest (OHCA) incident is reported to emergency services, the agent dispatches Emergency Medical Services (EMS) to the location and activates Responder Network System (RNS), if available. The RNS notifies all the registered users in the vicinity of the cardiac arrest patient by sending alerts to their mobile devices, which contains the location of the emergency. The notified users then find a publicly available Automated External Defibrillator (AED) and carry it to the emergency for early resuscitation of the patient. The currently implemented RNS has some drawbacks and can be enhanced to reduce the time taken by the notified users to reach the emergency,thus increasing the survival chances of the patients. In this research, the following changes are made in the RNS notification process: a) notify only those users who can reach the emergency before the arrival of EMS, b) provide the nearest AED location to the user c) keep the travel time minimum while assigning an AED toa user. These changes are achieved by using Bipartite Matching and Integer Linear Programming. However, these approaches take a longer processing time;therefore, a new approach Preprocessed Integer Linear Programming is proposed that solves the problem in a significantly lower time than the other two approaches and provides the most optimal solution.

Paper 61:
Predicting metabolic pathway membership with deep neural networks by integrating sequential and ontology information
Imam Cartealy and Li Liao

Inference of protein's membership in metabolic pathways has become an important task in functional annotation of protein. The membership information can provide valuable context to the basic functional annotation and also aid reconstruction of incomplete pathways. Previous works have shown success of inference by using various similarity measures of gene ontology. In this work, we set out to explore integrating ontology and sequential information to further improve the accuracy. Specifically, we developed a neural network model with an architecture tailored to facilitate the integration of features from different sources. Furthermore, we built models that are able to perform predictions from pathway-centric or protein-centric perspectives. We tested the classifiers using 5-fold cross validation for all metabolic pathways reported in KEGG database and shown that our method outperforms the existing methods significantly in the pathway-centric mode, and in the protein-centric mode, our method either outperforms or performs comparably with a suite of existing GO term based semantic similarity methods.

Paper 62:

Natural Language Processing (NLP) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder

Jacqueline Peng, Mengge Zhao, James Havrilla, Cong Liu, Chunhua Weng, Whitney Guthrie, Robert Schultz, Kai Wang and Yunyun Zhou

Natural language processing (NLP) tools can facilitate the extraction of biomedical concepts from unstructured free texts, such as research articles or clinical notes. The NLP software tools CLAMP, cTAKES, and MetaMap are among the most widely used tools to extract biomedical concept entities. However, their performance in extracting disease-specific terminology from literature has not been compared extensively, especially for complex neuropsychiatric diseases with a diverse set of phenotypic and clinical manifestations. To address this problem, we used autism spectrum disorder (ASD) as a case study, with the goal of evaluating these NLP tools in extracting ASD-specific vocabulary from 545 full-text articles and 20,424 abstracts from PubMed. Our results show that CLAMP had the best performance in extracting disease-specific entities, among the three NLP tools. Furthermore, since CLAMP uses the conditional random field (CRF) machine learning algorithm for entity extraction, we retrained the CRF model on PubMed articles and compared its performance to the deep learning algorithm BioBERT. BioBERT performed comparably well to the CLAMP CRF model in learning from an existing terminology set, but is better at finding additional ASD-relevant terms. The preliminary ASD terms extracted from the PubMed literature in this study can be used to facilitate the precise diagnosis of ASD and improve our understanding of the phenotypic manifestations of ASD. The analysis protocols used in this study can be utilized in other neuropsychiatric or neurodevelopmental disorders that lack well-defined terminology sets to describe their phenotypic presentations.

Paper 63:

Fingerprint Restoration using Cubic Bezier Curve

Yanglin Tu, Zengwei Yao, Jiao Xu, Yilin Liu, Zhe Zhang and Shuaicheng Li

Background: Fingerprint biometrics play an essential role in the authentication. It remains a challenge to match fingerprints with the informative minutiae or ridges missing. Many fingerprints cannot be effectively identified due to incompleteness and other reasons.

Result: In this work, we modeled the fingerprints with Bezier curves and proposed a novel algorithm to detect and restore fragmented ridges in fingerprints. Our model proposes that the control points of the Bezier curve can be used to describe the fingerprint. This method can help to describe fingerprints. Compared with the commonly used picture format, we save 89% of space in this form. Using our algorithm to restore the image,100% is considered to be the same fingerprint as the original image. We can also use it to reconstruct incomplete fingerprints, which can improve the matching score 22.34% to 60.57%, and compared our method with the one proposed by Feng. Under the unified standard, Feng's results are unsatisfactory, or Even worse than the matching results of the incomplete fingerprint. On the contrary, more incomplete fingerprints can be identified after restoration by our method, the False Acceptance Rate is 4.59%and the False Reject Rate is 2.83%.

Conclusions: Experimental results show that the proposed algorithm can successfully repair and reconstruct ridges in single or multiple damaged regions of incomplete fingerprint images, and hence improve the accuracy of fingerprint matching.

Paper 65:
Tissue Classification Using Landmark and Non-landmark Gene Sets for Feature Selection

Carly Clayman, Alakesh Mani, Suraj Bondugula and Satish Srinivasan

Dimensionality reduction methods such as principal component analysis (PCA) are used to select relevant features, and k-means clustering performs well when applied to data with low effective dimensionality. The L1000 dataset, containing gene microarray data from 978 landmark genes has been previously shown to predict expression of ~81% of the remaining 21,290 target genes with low error. Groups within the L1000 dataset were characterized using microarray data to assess whether 978 landmark genes would improve clustering samples into groups based on distinct tissue types, compared to a random set of 978 genes. The 978 landmark genes better differentiated k-means clusters, relative to 978 non-landmark genes. These results suggest that the 978 landmark genes better represent the overall genetic profile of these heterogeneous samples. Future studies will implement predictive analytics techniques to further investigate interaction of microarray data and other sample characteristics such as tumor stage.

Paper 66:
Integrative analysis of histopathological images and chromatin accessibility data for estrogen receptor-positive breast cancer

Siwen Xu, Yunlong Liu, Weixing Feng, Kun Huang, Zixiao Lu, Qianjin Feng, Wei Shao, Christina Yu and Jill Reiter

Background: Existing studies have demonstrated that the integrative analysis of histopathological images and genomic data can be used to better understand the onset and progression of many diseases, as well as identify new diagnostic and prognostic biomarkers. However, since the development of pathological phenotypes are influenced by a variety of complex biological processes, complete understanding of the underlying gene regulatory mechanisms for the cell and tissue morphology is still a challenge. In this study, we systematically explored the relationship between the chromatin accessibility changes and the epithelial tissue proportion in histopathological images for estrogen receptor (ER) positive breast cancer.

Results: We firstly utilized our previously established whole slide image processing pipeline based on deep learning to perform global segmentation of epithelial and stromal tissues. Using canonical correlation analysis, we detected 436 potential regulatory regions that exhibited significant correlation between quantitative chromatin accessibility changes and the epithelial tissue proportion across 54 patients ($r > 0.5$, $FDR < 0.05$). By integrating ATAC-seq data with matched RNA-seq data, we found that these 436 regulatory regions were associated with 74 potential target genes. After functional enrichment analysis, we observed that these potential target genes were enriched in cancer-associated pathways. We demonstrate that using the gene expression signals and the epithelial tissue proportion extracted from our integration framework could stratify patient prognoses more accurately, outperforming predictions based on only omics or image features.

Conclusion: We conclude that our integrative analysis is a useful strategy for identifying potential regulatory regions in the human genome that are associated with tumor tissue quantification. This study will enable efficient prioritization of genomic regulatory regions identified by ATAC-seq data for further studies to validate their causal regulatory function. Ultimately, identifying epithelial tissue proportion-associated regulatory regions will further our understanding of the underlying molecular mechanisms of disease and inform the development of potential therapeutic targets.

Paper 67:
TPSC: A Module Detection Method Based on Topology Potential and Spectral Clustering in Weighted Networks and Its Application in Gene Co-expression Module Discovery

Yusong Liu, Christina Y. Yu, Wei Shao, Jie Hou, Weixing Feng, Jie Zhang, Xiufen Ye and Kun Huang

Background: Gene co-expression network is one of the most widely studied networks in biomedical area with algorithms such as WGCNA and lmQCM been developed to detect co-expressed modules. However, this kind of algorithms have limitations such as insufficient granularity and unbalanced module size, which prevent us from obtaining complete information of interest. In addition, it is difficult to incorporate prior knowledge for current co-expression module detection algorithms.

Results: In this paper, we propose a novel module detection algorithm based on topology potential and spectral clustering algorithm to detect co-expressed modules in gene co-expression networks. By testing on TCGA data, our novel method can provide more complete coverage of genes, more balanced module size and finer granularity than current methods in detecting modules with significant overall survival difference. In addition, the proposed algorithm can identify modules by incorporating prior knowledge.

Conclusion: In summary, we developed a method to obtain as much as possible information from networks with increased input coverage and the ability to detect more size-balanced and granular modules. In addition, our method can integrate data from different sources. Our proposed method performs better than current methods with complete coverage of input genes and finer granularity. Moreover, this method is designed not only for gene co-expression networks but can also be applied to any general fully connected weighted networks.

Paper 68:
A pan-kidney cancer study identifies subtype specific perturbations on pathways with potential drivers in renal cell carcinoma

Xiaohui Zhan, Yusong Liu, Christina Y. Yu, Tian-Fu Wang, Jie Zhang, Dong Ni and Kun Huang

Background: Renal cell carcinoma (RCC) is a complex disease and is comprised of several histological subtypes, the most frequent of which are clear cell renal cell carcinoma (ccRCC), papillary renal cell carcinoma (PRCC) and chromophobe renal cell carcinoma (ChRCC). While lots of studies have been performed to investigate the molecular characterizations of different subtypes of RCC, our knowledge regarding the underlying mechanisms are still incomplete. As molecular alterations are eventually reflected on the pathway level to execute certain biological functions, characterizing the pathway perturbations is crucial for understanding tumorigenesis and development of RCC.

Results: In this study, we carried out a pathway-based analysis to explore the mechanisms of various RCC subtypes with TCGA RNA-seq data. Both commonly altered pathways and subtype-specific pathways were detected. To identify the distinctive characteristics of each subtype, we focused on subtype-specific perturbed pathways. Specifically, we identified the upstream regulators based on differentially expressed genes within subtype-specific pathways. Further analysis indicated that most of the common upstream regulators are recurrent regulators affecting a majority of altered pathways, while presenting different expression patterns among various RCC subtypes. Moreover, we also evaluated the relationships between perturbed pathways and clinical outcome. Prognostic pathways were identified and their roles in tumor development and progression were inferred.

Conclusions: In summary, we evaluated the relationships among pathway perturbations, upstream regulators and clinical outcome for differential subtypes in RCC. We hypothesized that the alterations of common upstream regulators as well as subtype-specific upstream regulators work together to affect the downstream pathway perturbations and drive cancer initialization and prognosis. Our findings not only increase our understanding of the mechanisms of various RCC subtypes, but also provide targets for personalized therapeutic intervention.

Paper 69:
Differential alternative splicing between hepatocellular carcinoma with normal and elevated serum alpha-fetoprotein

Young-Joo Jin, Habtamu Aycheh, Seonggyun Han, John Chamberlin, Jaehang Shin, Seyoun Byun and Younghee Lee

Background: Serum alpha-fetoprotein (AFP) is the approved serum marker for hepatocellular carcinoma (HCC) screening. However, not all HCC patients show high (≥20 ng/mL) serum AFP, and the molecular mechanisms of HCCs with normal (<20 ng/mL) serum AFP remain to be elucidated. Therefore, we aimed to identify biological features of HCCs with normal serum AFP by investigating differential alternative splicing (AS) between HCCs with normal and high serum AFP.

Methods: We performed a genome-wide survey of AS events in 249 HCCs with normal (n=131) and high (n=118) serum AFP levels using RNA-sequencing data obtained from The Cancer Genome Atlas.
Results: In group comparisons of RNA-seq profiles from HCCs with normal and high serum AFP levels, 161 differential AS events (125 genes; _PSI>0.05, FDR<0.05) were identified to be alternatively spliced between the two groups. Those genes were enriched in cell migration or proliferation terms such as "the cell migration and growth-cone collapse" and "regulation of insulin-like growth factor (IGF) transport and uptake by IGF binding proteins." Most of all, two AS genes (FN1 and FAM20A) directly interact with AFP; these relate to the regulation of IGF transport and post-translational protein phosphorylation. Interestingly, 42 genes and 27 genes were associated with gender- and vascular invasion (VI), respectively, but only eighteen genes were significant in survival analysis. We especially highlight that FN1 exhibited increased differential expression of AS events (_PSI>0.05), in which exons 25 and 33 were more frequently skipped in HCCs with normal (low) serum AFP compared to those with high serum AFP. Moreover, these events were gender- and VI-dependent.

Conclusion: We found that AS may influence the regulation of transcriptional differences inherent in the occurrence of HCC maintaining normal rather than elevated serum AFP levels.

Paper 70:
Lilikoi V2: Deep-learning enabled, personalized pathway-based package for diagnosis and prognosis predictions using metabolomics data

Xinying Fang, Yu Liu, Zhijie Ren, Fadhl Alakwaa, Yuheng Du, Qianhui Huang and Lana X. Garmire

Previously we developed Lilikoi, a personalized pathway-based method to classify diseases using metabolomics data. Given the new trends of computation in the metabolomics field, here we report the next version of Lilikoi as a significant upgrade. The new Lilikoi v2 R package has implemented a deep-learning method for classification, in addition to popular machine learning methods. It also has several new modules, including the most significant addition of prognosis prediction, implemented by Cox-PH model and the deep-learning based Cox-nnet model. Additionally, Lilikoi v2 supports data preprocessing, exploratory analysis, pathway visualization and metabolite-pathway regression. In summary, Lilikoi v2 is a modern, comprehensive package to enable metabolomics analysis in R programming environment.

Paper 71:
Identifying risk factors of preterm birth and perinatal mortality using statistical and machine learning approaches

Parth Kothiya, Huanmei Wu, David Haas, Shelley Burnham, Bobbie Ray and Sara Quinney

Background: The association of pregnancy outcomes with maternal health and lifestyle is well studied. However, investigations on an independent basis, such as alcohol use alone or drug use alone has not been well explored to fullest. Our project aims to analyze these parameters on an independent basis and make the analysis more user-friendly. This is crucial in order to manage information and perform an analysis on multiple factors.

Methods: We used the MySQL database to organize and store data. The predictive model was built using MANOVA and logistic regression to predict factors that were associated on pregnancy outcomes including preterm birth and perinatal morbidity.

Results: The pregnancy database efficiently stores information and will continue collecting information. The key feature that significantly impacts pregnancy outcomes has been identified with statistical significance.

Conclusions: Multiple factors, including alcohol use, family medical history, tobacco use, patient's medical history, and drug for global impact on pregnancy outcome. Durg use such as marijuana and cocaine has significant impact on preterm birth pregnancy outcome. The Coordinate chart with effect index as an output makes it easier to analyze impact. The predicted impact factors cross validated with ROC curve and AUC(>0.5).

Paper 72:
Network-based Drug Sensitivity Prediction

Khandakar Tanvir Ahmed, Sunho Park, Qibing Jiang, Taehyun Hwang and Wei Zhang

Background: Drug sensitivity prediction and drug responsive biomarker selection on high-throughput genomic data is a critical step in drug discovery. Many computational methods have been developed to serve this purpose including several deep neural network models. However, the modular relations among genomic features have been largely ignored in these methods. To overcome this limitation, the role of the gene co-expression network on drug sensitivity prediction is investigated in this study.

Methods: In this paper, we first introduce a network-based method to identify representative features for drug response prediction by using the gene co-expression network. Then, two graph-based neural network models are proposed and both models integrate gene network information directly into neural network for outcome prediction. Next, we present a large-scale comparative study among the proposed network-based methods, canonical prediction algorithms (i.e., Elastic Net, Random Forest, Partial Least Squares Regression, and Support Vector Regression), and deep neural network models for drug sensitivity prediction. All the source code and processed datasets in this study are available at https://github.com/compbiolabucf/drug-sensitivity-prediction.

Results: In the comparison of different feature selection methods and prediction methods on a non-small cell lung cancer (NSCLC) cell line RNA-seq gene expression dataset with 50 different drug treatments, we found that (1) the network-based feature selection method improves the prediction performance compared to Pearson correlation coefficients; (2) Random Forest outperforms all the other canonical prediction algorithms and deep neural network models; (3) the proposed graph-based neural network models show better prediction performance compared to deep neural network model; (4) the prediction performance is drug dependent and it may relate to the drug's mechanism of action.

Conclusions: Network-based feature selection method and prediction models improve the performance of the drug response prediction. The relations between the genomic features are more robust and stable compared to the correlation between each individual genomic feature and the drug response in high dimension and low sample size genomic datasets.

Paper 75:
An adaptive method of defining negative mutation status for multi-sample comparison using next-generation sequencing

Nicholas Hutson, Fenglin Zhan, James Graham, Mitsuko Murakami, Han Zhang, Sujana Ganaparti, Qiang Hu, Li Yan, Changxing Ma, Song Liu, Jun Xie and Lei Wei

Background: Multi-sample comparison is commonly used in cancer genomics studies, including profiling tumor heterogeneity and serial liquid biopsies. In these analyses, the statuses of a specific mutation in multiple samples need to be compared to understand the mechanism of cancer progression. By using next-generation sequencing (NGS), a mutation's status in a specific sample can be measured by the number of reads supporting mutant or wildtype alleles. When no mutant reads are detected in a sample, it could represent either a true negative mutation status or a false negative due to an insufficient number of total measured reads, so-called ""coverage"", at the genomic location of the mutation. To minimize the chance of false negative, we should consider the mutation status as ""unknown"" instead of ""negative"" when the coverage is inadequately low. There is no established method for determining the coverage threshold. A common solution is to require the coverage to pass a universal minimum coverage (UMC). However, this method relies on an arbitrarily chosen threshold, and it does not take into account the mutations' relative abundances in the positive samples, which can vary dramatically by the type of mutations. The result could be misclassification between negative and unknown statuses.

Methods: We propose an adaptive mutation-specific negative (MSN) method to improve the classification of negative and unknown mutation statuses. A non-positive sample for a specific mutation is compared with every known positive sample to test the null hypothesis that they may contain the same frequency of that mutation. The non-positive sample can only be claimed as “negative” when this null hypothesis is rejected with all known positive samples, otherwise, the status would be “unknown”.

Results: When evaluated on a real dual-platform single-cell sequencing dataset, the MSN method not only provided more accurate assessments of negative statuses, but also yielded three times more available data after excluding the “unknown” statuses, compared with the UMC method.

Conclusions: We developed a new adaptive method for distinguishing unknown from negative statuses in multi-sample comparison NGS data. The method can provide more accurate negative statuses than the conventional UMC method and generate a remarkably higher amount of available data by reducing unnecessary “unknown” calls.

Abstract 4:
Novel Variations in Goat PRNP Gene and Their Potential Effect on Prion Protein Stability

Eden Teferedegn, Cemal Un and Yalcin Yaman

Scrapie is a lethal neurodegenerative disease of sheep and goat caused by the misfolding of the prion protein. Variations such as M142, D145, S146, H154, Q211, and K222 were experimentally found to extend scrapie incubation period and increase resistance to scrapie in goat. We aimed to identify frequencies of scrapie related variations in two of Ethiopian goats breeds and compare them with other goat population in the world, and also to detect possible novel gene variants encoding different functional domains of goat prion protein. Besides scrapie related variations, four non-synonymous novel polymorphisms G67S, W68R, G69D, and R159H in the first octarepeat and the highly conserved C-terminus globular domain were detected. Frequencies of the resistant variants were low in the population under study. According to in-silico analysis variants 68R and 69D increase the stability of prion protein and decrease amyloidogenicity of the hotspot sequence. The relative solvent accessibility of the substituted region was higher than the wild type amino acid sequence. These novel variants could be the source of conformational flexibility that may trigger the gain/ loss of function by prion protein or are protective against PrPC to PrPSc conversion in the course of scrapie development.

Abstract 8:
Prediction of epithelial ovarian cancer in patients with pelvic mass with preoperative CT segmentation using deep learning technique

Hyun Hoon Chung, Aeran Seol, Jing Hui Jinag and Yeong Gil Shin

Objective: The aim of this study was to use deep learning technique to predict the malignancy of pelvic mass on the basis of preoperative abdominal computed tomography (CT) in patients with epithelial ovarian cancer.

Methods: Institutional review board approved this retrospective study. Preoperative CT scans in patients with suspected ovarian malignancy were performed in 135 patients, and 10 scans with no abnormal findings were used as a reference standard. The deep learning model was trained by coaching. A running model of a nonlinear neural network had local CT data read and taught the labels of that data to enable classification. The basis for discriminating normal from unusual ovaries in CT images is the size and composition of the ovaries. Because the number of data was smaller than conventional task that applies deep learning, and the data imbalance was greater, the experiment was conducted by balancing normal and unusual data with data integration techniques. The training was conducted to distinguish between ovaries and lack of sufficient learning due to the lack of data, and the parameters of the model were finely tuned to conduct normal and abnormal sorting again. The ovarian-free images were mainly selected as breast, femur or lower, and a small number of small images.

Results: Validation set consisted of 59 scans with no abnormal findings, and 41 scans with pelvic mass diagnosed as epithelial ovarian cancer. The results of the experiment show that the rate of recognition rate of the presence or absence of ovaries was 99.9%. The sensitivity of the deep learning technique was 75.6% (31 of 41), and the specificity was 67.8% (40 of 59). The accuracy of the technique was 71.0%.

Conclusion: The deep learning technique for predicting the malignancy of pelvic mass in preoperative CT showed relatively high accuracy of 71.0% with 32.2% false negativity. Results from our study provide valuable insight to make important preoperative prediction of pelvic tumor before treatment.

Abstract 11:
Space: A Missing Piece of the Dynamic Puzzle

Armin Iraji, Robyn Miller, Tulay Adali and Vince Calhoun

The brain is a multiscale dynamic system in which functional units interact and evolve at different spatial and temporal scales. Thus, analyzing the temporal reconfiguration of functional connectivity, known as dynamic functional connectivity (dFC), plays a pivotal role to study the functional interactions between neuronal populations and to characterize their roles in brain function. Recent studies support the potential neurophysiological relevance of dFC and the benefit of studying dFC properties to improve the understanding of healthy and disordered brain function. However, a focus on existing approaches shows that most focus on temporal dynamics and ignore spatial and spatiotemporal dynamics. Here, we accentuate the importance of within-subject spatial variations of functional units.

We highlight and extend recent work from our group which shows we can capture spatial dynamics using a spatially fluid chronnectome model. We showed evidence that moment-to-moment spatial reconfiguration of brain networks occurs at the finest observable scale (voxel-level) and that these changes are important and relevant for the study of brain disorders. We find that brain networks evolve spatially over time, and they transiently merge and separate from each other. We show that spatial dynamics could explain the inconsistencies observed in previous spatially static FC (sFC) studies. We show that distinct default mode network reported in sFC studies appear at different times, highlighting inconsistencies due to methods that do not account for spatial dynamics.
Another approach we discuss is the study of the dynamic properties of the brain by using a hierarchical models of brain function. In this framework, we estimate brain function within a hierarchical structure based on functional homogeneity. Different levels of the hierarchy represent different spatial scales and capture distinct dynamic information. We show evidence of spatially dynamic properties within functional domains (FDs), including changes in a region's association to a given FD from strong association to complete dissociation.

One crucial advantage of incorporating space in a dFC analysis is that it provides a new set of dynamic metrics, inaccessible using previous FC methods, to study brain function. Spatial dynamic studies also reveal unique patterns of alterations in psychosis, which are hidden and undetectable using previous sFC and dFC techniques.

In conclusion, neuroimaging research is shifting rapidly toward studying brain dynamism from the perspective of the temporal reconfiguration of functional connectivity. We suggest the incorporation of spatial dynamics into dFC analyses is a promising avenue for understanding the mechanisms and clinical implications of brain dynamism.

Abstract 26:
Gut microbiome analysis towards targeted non-invasive biomarkers for early multiple myeloma diagnosis

Xingxing Jian, Yinghong Zhu, Huihui Wan, Wen Zhou and Lu Xie

The interactions between gut microbiome and host are extensively studied and found closely related to human health. The potential of gut microbes as markers for early diagnosis has been reported in several tumors. Although efforts have been made to reveal the microbe-host interactions in multiple myeloma (MM) subjects, to date, there is no study towards characterizing gut microbiome as markers for early MM diagnosis.

In this work, based on in-house metagenomics sequencing data, 37 samples from a cohort of newly diagnosed group of MM patients and healthy controls (HCs) were analyzed. The convergent gut micrbiota composition in MM was presented owing to the tumorigenesis and progression of MM. In addition, the species with differential abundance were identified in MM, of which 11 were further verified using qPCR in new samples set. These 11 gut microbes include: MM-depleted Anaerostipes hadrus, Clostridium butyricum, and MM-enriched Citrobacter freundii, Enterobacter cloacae, Klebsiella aerogenes, Klebsiella variicola, Klebsiella pneumonia, Streptococcus salivarius, Streptococcus oralis, Streptococcus gordonii, Streptococcus mitis. Taking these characteristic species as features, 10 diagnosis models were constructed using different machine learning methods, of which 7 were validated with AUC of >0.7 in an independent validation set.

This study provides a novel insight for MM diagnosis. We first characterized gut microbiome in MM as biomarkers and reported the successful establishment of diagnosis models, indicating the potential of gut microbiome analysis towards targeted non-invasive biomarkers for early MM diagnosis.

Abstract 76:
Identifying Falls Documented in Home Health Care Clinical Notes using Natural Language Processing

Judy Hong, Yancy Lo, Kathryn H. Bowles, Margaret V. McDonald, Jason H. Moore and Danielle Mowery

Motivation: Falls are the leading cause of injuries among the elderly, particularly those in the more vulnerable home health care (HHC) population. Existing standardized fall risk assessments require additional effort to collect and have low specificity. The precision of machine learning models for predicting fall risk during HHC is limited by unbalanced outcome classes due to the short time frame of a HHC episode and the potential underreporting of falls in coded administrative data. Our short-term goal was to develop a natural language processing (NLP) approach to increase the identification of falls in HHC by including the narrative clinical notes.

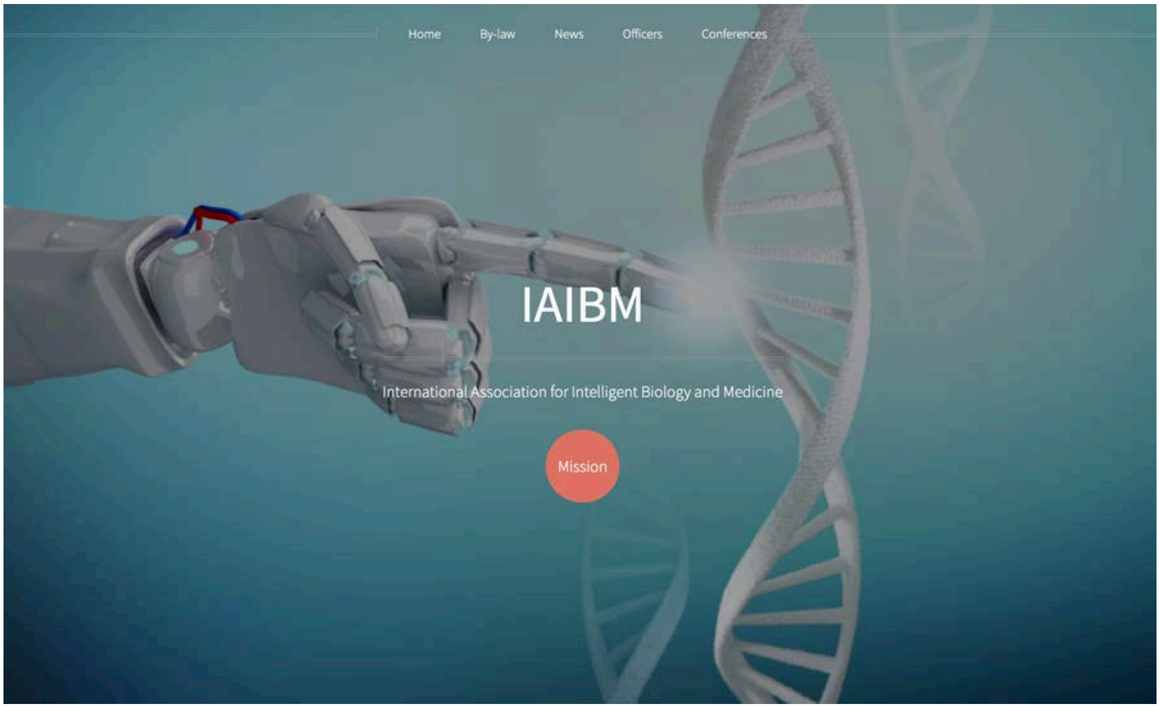
Method: After Institutional Review Board approval, we randomly selected 47,500 clinical notes from a set of 3,000 HHC patients who received services from a large, urban agency in New York City. We trained an NLP algorithm called pyConText, which leverages networkX digraphs to relate targets (positive and exclusionary terms for falls) with modifiers (negation, temporality, prevention, unrelated) to assert whether or not a fall event occurred (positive, possible, negative). We generated the pyConText initial lexicon by adapting Zhu et al. 2017 fall assessment terms and incorporating syntactic knowledge into its variants. We validated each flagged sentence using manual review.

Results: pyConText achieved promising accuracies of 0.76 (positive = “pt stated she fell out of bed”), 0.95 (possible = “patient still at risk for falls”) and 0.89 (negative = “no recent falls reported”; “wound dressing had fallen off”). For next steps, we will apply pyConText to the full 59,006 patients, estimate the number of falls omitted from the HHC structured dataset (original incidence is 5.14%), and incorporate new data into our existing machine learning fall risk prediction model that leverages the random forest algorithm (Lo et al. 2019).

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.


Home
By-law
News
Officers
Conferences
IAIBM
International Association for Intelligent Biology and Medicine
Mission