APBC 2023

The 21th Asia Pacific Bioinformatics Conference

April 14-16, 2023

Changsha, Hunan, China

The APBC 2023 will take place in Information Building, School of Computer Science and Engineering, Central South University.

APBC 2023

Chairs & Committees

Conference Chairs

Jianxin Wang (Central South University, China)

Yi Pan (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China)

Phoebe Chen (La Trobe University, Australia)

Program Committee Chairs

Min Li (Central South University, China)

Feng Luo (Clemson University, USA)

Local Organization Chairs

Hong-Dong Li (Central South University, China)

Ruiqing Zheng (Central South University, China)

Min Zeng (Central South University, China)

Yang Gao (Central South University, China)

Publication Chairs

Fei Guo (Central South University, China)

Jin Liu (Central South University, China)

Committee Members

Min Li (Central South University)
Sun Kim (Seoul National University)
Gabriel Valiente (Technical University of Catalonia)
Jijun Tang (University of South Carolina)
Ruiqing Zheng (Central South University)
Renchu Guan (Jilin University)
Shaoliang Peng (Hunan University)
Yang Gao (Central South University)
Lenore Cowen (Tufts University)
Giltae Song (Pusan National University)
Shichao Kan (Central South University)
Min Zeng (Central South University)
Jin Liu (Central South University)
Lingzhi Zhu (Hunan Institute of Technology)
Mengyun Yang (Shaoyang University)
Chengqian Lu (Xiangtan University)
Phoebe Chen (La Trobe University)
Ying Zheng (Changsha University of Science &Technology)
Wei Lan (Guangxi University)
Rui Yin (University of Florida)
Xiaoqing Peng (Central South University)
Ying An (Central South University)
Yinglei Lai (George Washington University)
Hulin Kuang (Central South University)
Louxin Zhang (National University of Singapore)
Minzhu Xie (Hunan Normal University)
Zengyou He (Dalian University of Technology)
Xiangmao Meng (Xiangtan University)
Shuai Cheng Li (City University of Hong Kong)
Wing-Kin Sung (National University of Singapore)
Hong-Yu Zhang (Huazhong Agricultural University)
Pawel Gorecki (University of Warsaw)
Hsuan-Cheng Huang (National Yang-Ming University)
Mingfu Shao (Carnegie Mellon University)
Ju Xiang (Central South University)
Manuel Lafond (Université de Sherbrooke)
Weichuan Yu (The Hong Kong University of Science and Technology)
Fengfeng Zhou (Jilin University)
Wei Peng (Kunming University of Science and Technology)
Juan Liu (Wuhan University)
Tatsuya Akutsu (Kyoto University)
Min Xu (Carnegie Mellon University)
Kyungsook Han (Inha University)
Hui Jiang (University of South China)
Junwei Luo (Henan Polytechnic University)
Xingyi Li (Northwestern Polytechnical University)
Kenta Nakai (Institute of Medical Science, University of Tokyo)
Yufeng Wu (University of Connecticut)
Hsien-Da Huang (The Chinese University of Hong Kong, Shenzhen)
Lusheng Wang (City University of Hong Kong)
Jie Zheng (ShanghaiTech University)
Eric Ho (Lafayette College)
Wen Zhang (Huazhong Agricultural University)
Feng Liu (Wuhan university)
Yoshihiro Yamanishi (Kyushu Institute of Technology)
Quan Zou (Tianjin University)
Guoliang Li (Huazhong Agricultural University)
Katharina Huber (University of East Anglia)
Shihua Zhang (Academy of Mathematics and Systems Science, Chinese Academy of Sciences)
Hui Lu (University of Illinois at Chicago)
Dongbo Bu (Institute Of Computing Technology, Chinese Academy of Sciences)
Feng Luo (Clemson University)
Yu Xue (Huazhong University of Science and Technology)
Cheng Yan (Hunan University of Chinese Medicine)

Conference Program Overview

Day 1: Friday, April 14, 2023		
14:00-21:00	Registration (Fushengyuan Hotel 1st floor's lobby)	
19:00-20:30	Dinner (Fushengyuan Hotel 3rd floor)	
Day 2: Saturday, April 15, 2023		
8:30-8:45	**Symposium Opening and Welcome** (Room 108, Information Building/ Zoom 1)	
8:45-9:35	Keynote Talk 1	
9:35-10:10	Group Photo & Tea break	
10:10-11:00	Keynote Talk 2	
11:00-11:50	Keynote Talk 3	
11:50-13:30	Lunch (Fushengyuan Hotel 3rd floor)	
13:00-13:30	Poster Presentation	
13:30-15:30	Session 1: Cheminformatics and Pharmacogenomics (Room 535/Zoom 1)	Session 2: Medical Data Analysis (Zoom 2)
15:30-15:50	Tea break	
15:50-17:50	Session 3: Computational Biology of Molecular Structure and Function (Room 535/ Zoom 1)	Session 4: Network Biology and Systems Biology (Zoom 2)
18:00-19:30	Banquet (Fushengyuan Hotel 3rd floor)	
Day 3: Sunday, April 16, 2023		
8:30-10:10	Session 5: Cell Biology and Regulation (Room 535/Zoom 1)	Session 6: Next Generation Sequencing and High-throughput Method (Zoom 2)
10:10-10:30	Tea break	
10:30-12:10	Session 7: Biomedical Image Analysis (Room 535/Zoom 1)	Session 8: Genomics, Transcriptomics and Proteomics (Zoom 2)
12:10-13:30	Lunch (Fushengyuan Hotel 3rd floor)	
13:30-15:10	Session 9: Artificial Intelligence of Health Informatics (Room 535/Zoom 1)	Session 10: Electronic Medical/Health Records Mining (Zoom 2)
15:10-15:30	Tea break	
15:30-17:30	Session 11: Data Mining and Artificial Intelligence of omics data (Room 535/Zoom 1)	Session 12: Cross-Cutting Computational Methods (Zoom 2)
17:30-18:00	**Ceremony and Closing Remarks**	
18:00-19:30	Dinner (Fushengyuan Hotel 3rd floor)	

APBC2023 Program

Day 1: Friday, April 14, 2023	
14:00-21:00	Registration (Fushengyuan Hotel 1st floor's lobby)
19:00-20:30	Dinner (Fushengyuan Hotel 3rd floor)

Day 2: Saturday, April 15, 2023		
8:30-8:45	**Symposium Opening and Welcome** (108 Room, Information Building) • **Phoebe Chen**, Professor, La Trobe University, Australia. • **Yi Pan**, Professor, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China.	Chair: **Min Li**
8:45-9:35 Plenary Session	Keynote Talk 1 Spatiotemporal omics algorithm development, bioinformatics integration and platform application ***Professor. Xun Xu****, BGI Research.*	Chair: **Yi Pan**
9:35-10:10	Group Photo & Tea break	
10:10-11:00 Plenary Session	Keynote Talk 2 Graph representation learning of single-cell omics data ***Associate Professor. Qin Ma****, Department of Biomedical Informatics, Ohio State University.*	Chair: **Min Li**
11:00-11:50 Plenary Session	Keynote Talk 3 Enhancing Statistical Rigor in Genomics Data Science ***Professor. Jingyi Jessica Li****, Department of Statistics & Department of Human Genetics and Department of Biomathematics, University of California, Los Angeles.*	Chair: **Min Li**
11:50-13:30	Lunch (Fushengyuan Hotel 3rd floor)	
13:00-13:30	Poster Presentation	

Parallel Session	Session 1: Cheminformatics and Pharmacogenomics (Room 535/Zoom 1)	Session Chair: *Lun Hu*	Session 2: Medical Data Analysis (Zoom 2)	Session Chair: *Hongdong Li*
13:30-13:50	B5527 "RLFDR: A graph representation learning framework for drug repositioning over heterogeneous information networks"		B1075 "DR-3DUNet: A Dense Residual 3D UNet for Lung Nodule Segmentation"	
13:50-14:10	B4770 "Predicting Drug Target Interaction Based on BERT Model and Subsequence Embedding"		B8595 "MMAR-net: a Multi-stride and Multi-resolution Affine Registration network for CT images"	
14:10-14:30	B4482 "DDI-Transform: A DDI-Transform Neural Network for Predicting Drug-Drug Interaction Events"		B1661 "SimHOEPI: a resampling simulator for generating SNP data with high-order epistasis model"	
14:30-14:50	B6539 "Knowledge Enhanced Attention Aggregation Network for Medicine Recommendation"		B4510 "A Multimodal Data Extraction Pipeline for MIMIC with Parallelization"	
14:50-15:10	B7675 "DTKGIN: Predicting Drug-Target interactions based on Knowledge Graph and Intent Graph"		B3811 "ACT-Tooth: A Semi-Supervised Tooth Volume Segmentation in CBCT images based on Asymmetric CNN-Transformer Network and Cross Consistency"	
15:10-15:30	B804 "Multimodal contrastive representation learning for drug-target binding affinity prediction"		B2904 "MRUNet-3D: a Multi-stride Residual 3D UNet for Lung Nodule Segmentation"	
15:30-15:50	Tea break			
Parallel Session	Session 3: Computational Biology of Molecular Structure and Function (Room 535/ Zoom 1)	Session Chair: *Jie Li*	Session 4: Network Biology and Systems Biology (Zoom 2)	Session Chair: *Wei Lan*
15:50-16:10	B4906 "PIN: a penalized integrative deep neural network for variable selection among multiple omics datasets"		B3985 "Structure-Aware Sparse Transformer-Based Model for Predicting Drug-Target Binding Affinity"	
16:10-16:30	B832 "Improving biome labeling for tens of thousands of inaccurately annotated microbial community samples based on neural network and transfer learning"		B5528 "A feature extraction framework for discovering pan-cancer driver genes based on multi-omics data"	
16:30-16:50	B8120 "Tensor Improve Equivariant Graph Neural Network for Molecular Dynamics Prediction"		B9980 "MLRR-ATV: A Robust Manifold Nonnegative Low-Rank Representation with Adaptive Total-Variation Regularization for scRNA-seq Data Clustering"	

16:50-17:10	B2666 "POLAT: Protein function prediction based on soft mask graph network and residue-Label ATtention"	B8006 "A new network-based multi-omics pathway analysis method"
17:10-17:30	B8524 "Planning Biosynthetic Pathways of Target Molecules Based on Metabolic Reaction Prediction and AND-OR Tree Search"	B4386 "scMSSL: Multi-scale attention semi-supervised learning with deep generative models to automatically identify cell types"
17:30-17:50	B7880 "A Novel Meta Sparse Learning Method for Brain Imaging Genetics without Individual-level Data"	B3012 "SMCC: a novel clustering method for single- and multi-omics data based on co-regularized network fusion"
18:00-19:30	Banquet (Fushengyuan Hotel 3rd floor)	

Day 3: Sunday, April 16, 2023				
Parallel Session	Session 5: Cell Biology and Regulation (Room 535/ Zoom 1)	Session Chair: *Yushan Qiu*	Session 6: Next Generation Sequencing and High-throughput Method (Zoom 2)	Session Chair: *Junwei Luo*
8:30-8:50	B1439 "A novel multilevel iterative training strategy for the ResNet50 based mitotic cell classifier"		B5169 "AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes"	
8:50-9:10	B6643 "JLONMFSC: Clustering scRNA-seq data based on joint learning of non-negative matrix factorization and subspace clustering"		B8847 "TargetCall: Eliminating the Wasted Computation in Basecalling via Pre-Basecalling Filtering"	
9:10-9:30	B4979 "Deep imputation bi-stochastic graph regularized matrix factorization for single-cell RNA-sequencing data clustering"		B5582 "GenMPI: Cluster Scalable Variant Calling for Short/Long Reads Sequencing Data"	
9:30-9:50	B8532 "scLRSSC: a low rank and sparse subspace clustering framework for scATAC-seq"		B5515 "Fully automated annotation of mitochondrial genomes using a cluster-based approach with de-Bruijn graphs"	
9:50-10:10	B1652 "Gene regulatory network inference based on dynamic L0 regularization"		B2166 "Genomic Variations Explorer (GenVarX): A Toolset for Annotating Promoter and CNV Regions Using Genotypic and Phenotypic Differences"	
10:10-10:30	Tea break			

Parallel Session	Session 7: Biomedical Image Analysis (Room 535/ Zoom 1)	Session Chair: *Yang Gao*	Session 8: Genomics, Transcriptomics and Proteomics (Zoom 2)	Session Chair: *Mengyun Yang*
10:30-10:50	B2086 "COVID-FPT: Feature Pyramid Transformer for COVID-19 Detection on CXR Images"		B7071 "Prediction of CRISPR/Cas9 Editing Repair Outcomes Using Blended Machine Learning and Distributed Hyperparameter Optimization"	
10:50-11:10	B4580 "SemiCTrans: Semi-Supervised Medical Image Segmentation Framework combining CNN and Transformer"		B5970 "Synthesis Cost-Optimal Targeted Mutant Protein Libraries"	
11:10-11:30	B8098 "GFDet: Multi-level Feature Fusion Network for Caries Detection Using Dental Endoscope Images"		B4168 "CLCAP: Contrastive Learning Improves Antigenicity Prediction for Influenza A Virus Using Convolutional Neural Networks"	
11:30-11:50	B6529 "Automatic Coarse-to-Refinement Kidney Segmentation in Ultrasound Images"		B2310 "Accurately Predicting Methylation Site Using Deep Convolutional Neural Network"	
11:50-12:10	B3401 "Multi-scale Self-attention Multiple Instance Learning for Whole Slide Image Classification"		B3469 "iLncDA-RSN: identification of lncRNA-disease associations based on reconstructed similarity network"	
12:10-13:30	Lunch (Fushengyuan Hotel 3rd floor)			
Parallel Session	Session 9: Artificial Intelligence of Biomedical and Health Informatics (Room 535/ Zoom 1)	Session Chair: *Yijia Zhang*	Session 10: Electronic Medical/Health Records Mining (Zoom 2)	Session Chair: *Yin Yu*
13:30-13:50	B4408 "A New Parallel Spiking Neural Network Simulator Using Sunway TaihuLight"		B502 "Can ECG Be Reconstructed From IPPG?"	
13:50-14:10	B4041 "Language Model based on Deep Learning Network for Biomedical Named Entity Recognition"		B1182 "DAMS-Net: Dual Attention and Multi-Scale Information Fusion Network for 12-lead ECG Classification"	
14:10-14:30	B8100 "Re-examine Statistical Relationships among Dietary Fats, Risk Factors, and Cardiovascular Disease Risks based on Two Crucial Dataset"		B1402 "Pipelined Biomedical Event Extraction Rivaling Joint Learning"	
14:30-14:50	B2701 "MMR: A Multi-view Merge Representation Model for Chemical-Disease Relation Extraction"		B283 "Multi-Branch Transformer Fusion Network for End-to-End Motor Imagery Electroencephalogram Decoding"	
14:50-15:10	B9348 "KD_ConvNeXt: Knowledge distillation-based images classification of lung tumor surgical specimen sections"		B3565 "ResNet Meets PCBAM and SE for Medical Visual Question Answering"	

15:10-15:30	Tea break			
Parallel Session	Session 11: Data Mining and Artificial Intelligence of omics data (Room 535/Zoom 1)	Session Chair: *Cheng Yan*	Session 12: Cross-Cutting Computational Methods (Zoom 2)	Session Chair: *Jianhong Cheng*
15:30-15:50	B6730 "HKFGCN: A novel multiple kernel fusion framework on graph convolutional network to predict microbe-drug associations"		B3039 "Tuning Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding"	
15:50-16:10	B784 "Predicting associations between drugs and G Protein-Coupled Receptors using a multi-graph convolutional network"		B4185 "LPI-IBWA: Predicting lncRNA-protein interactions based on improved Bi-Random walk algorithm"	
16:10-16:30	B2038 "BERT-5mC: an interpretable model for predicting 5-methylcytosine sites of DNA based on BERT"		B7076 "Efficient sequencing data compression and FPGA acceleration based on a two-step framework"	
16:30-16:50	B4052 "DLP: duplex link prediction via subspace segmentation for predicting drug-miRNA associations"		B2850 "Multi-scale DCNN with Dynamic Weight and Part Cross-entropy Loss for Skin Lesion Diagnosis"	
16:50-17:10	B886 "Multi-modal Epileptic Seizure Detection Based on LogWT-CNN and MFCC" (online)		B2213 "Predicting the human miRNA-disease associations based on Non-linear Gaussian Profile Kernel Similarity"	
17:10-17:30			B2804 "A Hierarchical Model Based on Semantic Relation Extraction of Small Cell Lung Cancer Patents"	
17:30-18:00	**Ceremony and Closing Remarks**			
18:00-19:30	Dinner (Fushengyuan Hotel 3rd floor)			

Keynote Speakers

Xun Xu

Topic: Spatiotemporal omics algorithm development, bioinformatics integration and platform application

Abstract

Spatial omics technologies generate comprehensive spatially resolved datasets, broadening the understanding of organ development, tumor heterogeneity, cancer evolution, and other issues. The data features of Spatial omics datasets differ from single-cell RNA sequencing datasets, not only the extra spatial information, but also different gene distribution. ST-specific algorithms and tools that utilize spatial information and data features need to be developed to investigate and interpret data from a spatial perspective. This talk will introduce the efforts we have made for interpreting the spatial omics datasets, including the algorithms, tools and platform. This talk will first describe some of our recent work for generating higher accuracy cell-level gene expression, data quality improvement, finer cell clustering and annotation, and further data interpretation. Then, this talk will introduce the tools developed to better analyze and understand the spatial transcriptomics datasets with novel function modules and complete analysis workflow. Finally, this talk will present the online analysis platform, named Stomics Cloud, which includes the basic analysis for generating the spatial gene expression, advanced analysis for interpreting the datasets, personal analysis with abundant analysis notebook and docker images.

Introduction

Xun Xu, PhD, Director of BGI-Research, Doctoral supervisor of University of Chinese Academy of Sciences. Dr. Xu is appointed as Vice-Chairman of International Organization for Standardization/Biotechnology Committee (ISO/TC276), and on the National Committee for Standardization of Biological Specimens. In addition, he is also elected as Director-at-Large – China of International Society for Biological and Environmental Repositories (ISBER).

His research focuses on the sequencing technologies, single-cell sequencing, stereo-seq technologies and applications of sequencing technologies in precision medicine and biodiversity. Dr. Xu has authored or co-authored 300+ scientific papers published in top international peer-reviewed journals including Nature, Science, and Cell. The number of citations using his published papers in the past five years has reached more than 24,000 times. He’s also Selected as a "Highly Cited Scientist" by Clarivate Analytics for eight consecutive years.

Dr. Xu has either led or participated in several major national scientific research projects and several international projects, including the National 863 Sequencer Project, the National Key R&D Program, and the National Development and Reform Commission's Industrial Agglomeration Project. He has been awards with many honors, such as “Grand Challenges-Young Scientist” by MOST and Bill & Melinda Gates Foundation, “Guangdong Science and Technology Award”, “Natural Science Award of MOE”.

Qin Ma

Topic: Graph Representation Learning of Single-Cell Omics Data

Abstract

Artificial Intelligence(AI) and single-cell studies have been making waves in the science and technology communities. AI offers a broad range of methods that can be used to investigate diverse data- and hypothesis-driven questions in single-cell biology (Ma, Q., Xu, D. Deep learning shapes single-cell data analysis. Nat Rev Mol Cell Biol, 2022). The highly heterogeneous nature of single-cell data can be analyzed across a wide range of research topics by generalizing deep-learning model design and optimization in a hypothesis-free manner. This talk will introduce in-house graph representation learning methods for single-cell omics data to discover underlying mechanisms in diverse biological systems.

Introduction

Qin Ma is currently and the Chief of the Bioinformatics and Computational Biology Section in the Department of Biomedical Informatics, Ohio State University (OSU), and Leader of the Immuno-Oncology Informatics Group Pelotonia Institute for Immuno-Oncology at The OSU Comprehensive Cancer Center. He received his Ph.D. in Operational Research from Shandong University and then did his postdoc at the University of Georgia, specializing in high-throughput sequencing data mining and modeling. He established his bioinformatics research lab and moved on to the field of single-cell sequencing data analyses at South Dakota State University. Currently, his lab focuses on developing computational methods to discover heterogeneous transcriptional regulatory mechanisms from single-cell multi-omics data, with a particular interest in deploying deep learning methods in cancer research. Lab website: https://u.osu.edu/bmbl/.

Jingyi Jessica Li

Topic: Enhancing Statistical Rigor in Genomics Data Science

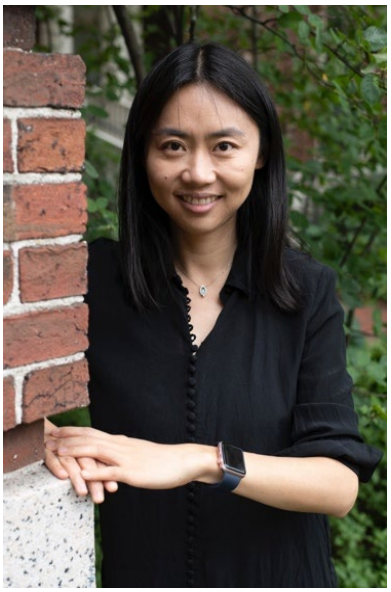
Abstract

The rapid development of genomics technologies has propelled fast advances in genomics data science. While new computational algorithms have been continuously developed to address cutting-edge biomedical questions, a critical but largely overlooked aspect is the statistical rigor. In this talk, I will introduce our recent work that aims to enhance statistical rigor in three analytical tasks where p-values may be ill-posed.

1. ChIP-seq peak calling;
2. RNA-seq differentially expressed (DE) gene detection;
3. single-cell RNA-seq DE gene detection following cell pseudotime inference or clustering.

I will also introduce our single-cell and spatial omics simulator scDesign3 as a tool for computational method benchmarking, real data interpretation, and in silico control generation.

Introduction



Dr. Jingyi Jessica Li is a Professor in the Department of Statistics at the University of California, Los Angeles, where she holds secondary appointments in the Departments of Biostatistics, Computational Medicine, and Human Genetics. She leads a research group called the Junction of Statistics and Biology, which focuses on developing interpretable statistical methods for biomedical data analysis. Dr. Li received her Ph.D. from the University of California, Berkeley, and her B.S. from Tsinghua University. Her research interests include quantifying the central dogma, extracting hidden information from bulk, single-cell, and spatial multi-omics data, and ensuring statistical rigor in biomedical data analysis. She emphasizes using in silico negative controls to avoid false discoveries. Dr. Li has received multiple awards in recognition of her work, including the NSF CAREER Award, Sloan Research Fellowship, Johnson & Johnson WiSTEM2D Math Scholar Award, MIT Technology Review 35 Innovators Under 35 China, Harvard Radcliffe Fellowship, COPSS Emerging Leader Award, and ISCB Overton Prize.

Accepted Papers

B283. Multi-Branch Transformer Fusion Network for End-to-End Motor Imagery Electroencephalogram Decoding

Qiying Cheng[1], Jingjing Luo[1], Hongbo Wang[1], Qiang Du[1], Youhao Wang[1] and Yang Li[2]

[1]Fudan University
[2]Beihang University

Abstract:
The brain-computer interface (BCI) technology has been extensively studied for stroke rehabilitation worldwide. In the BCI scheme, the decoding techniques for motor imagery electroencephalogram (MI-EEG) signals have built a bridge from biosignals to rehabilitation machines. Most existing deep learning decoding networks are convolution-based, focusing on local features. However, since EEG is non-stationary, its intra-sequence interdependencies of local features are influenced by the brain condition and the surrounding environment. Current architectures have limitations in adaptively learning global-wise interdependencies, bringing them a natural bottleneck in performance. This paper introduces an end-to-end multi-branch transformer network (MBTFN) that focuses on and fuses full-sequence length temporo-spatio-spectral information by self-attention mechanism and adaptively captures key elements in EEG signals. MBTFN employs transformer encoder structures to learn full-trial attention patterns of local features from multiple frequency bands in parallel and utilize a further multi-head self-attention to acquire fused features. The proposed model achieves an average accuracy (kappa value) of 86.93% (0.8249) on the BCIC IV 2a dataset, with the highest reaching 98.22% (0.9574), reaching a new state-of-the-art (SOTA). This approach explores a novel network structure that may be more suitable for MI-EEG decoding, which helps to improve the BCI system's performance.

B502. Can ECG Be Reconstructed From IPPG?

Xiaoxing Yang[1], Xinchun He[1] and Lin Lin[1]

[1]Shenzhen Technology University

Abstracts:
Image Photoplethysmography (IPPG) is a non-contact technique for measuring pulse signals that can be acquired by ordinary color cameras, which avoids direct contact between the human body and the equipment, and is more suitable for remote long-term monitoring. Reconstruction of electrocardiogram (ECG) signals by indirectly measuring IPPG signals will make it possible to monitor the health of the body through ECGs for a long time in daily life. Therefore, we attempt to study whether it is possible to reconstruct ECG signals using IPPG in this paper. Firstly, we design a dual channel acquisition platform of IPPG signal and ECG signal, in order to obtain original IPPG and ECG signals. Secondly, after preprocessing original IPPG and ECG signals, data is divided into training and testing data. Subsequently, the training data is fed into signal decomposition - long short-term memory networks, in order to obtain a model that can reconstruct ECG signals from IPPG signal. Finally, the obtained model is used to reconstruct the ECG signal waveform from the testing IPPG signal, and the reconstructed ECG is compared with the reference ECG. Experimental results show the feasibility of ECG waveform reconstruction from IPPG, which can greatly expand the breadth of ECG applications.

B784. Predicting Associations Between Drugs and G Protein-Coupled Receptors Using a Multi-Graph Convolutional Network

Yuxun Luo[1], Shasha Li[2] , Li Peng[1], Pingjian Ding[3] and Wei Liang[1]

[1]Hunan University of Science and Technology
[2]The University of Hong Kong
[3]University of South China

Abstracts:
Developing new drugs is an expensive, time-consuming process that frequently involves safety concerns. By discovering novel uses for previously verified drugs, drug repurposing helps to bypass the time-consuming and costly process of drug development. As the largest family of proteins targeted by verified drugs, G protein-coupled receptors (GPCR) is vital to efficiently repurpose drugs by inferring its associations with drugs. Drug repurposing may be sped up by computational models that predict the strength of novel drug-GPCR pairs interaction. To this end, a number of models have been put forth. In existing methods, however, drug structure, drug-drug interactions, GPCR sequence, and subfamily information couldn't simultaneously be taken into account to detect novel drugs-GPCR relationships. In this study, based on a multi-graph convolutional network, an end-to-end deep model was developed to efficiently and precisely discover latent drug-GPCR relationships by combining data from multi-sources. We demonstrated that our model based on multi-graph convolutional networks outperformed rival deep learning techniques as well as non-deep learning models in terms of inferring drug-GPCR relationships. Our results indicated that integrating data from multi-sources can lead to further advancement. Meanwhile, graph representation of interactions among biomedical entities is suitable for discovering novel relationships between drugs and GPCR.

B804. Multimodal Contrastive Representation Learning for Drug-Target Binding Affinity Prediction

Linlin Zhang[1], Chunping Ouyang[1], Yongbin Liu[1] and Zheng Gao[2]

[1]University of South China
[2]Indiana University Bloomington

Abstract:
In the biomedical field, the efficacy of most drugs is demonstrated by their interactions with targets, meanwhile, accurate prediction of the strength of drug-target binding is extremely important for drug development efforts. Traditional bioassay-based drug-target binding affinity (DTA) prediction methods cannot meet the needs of drug R&D in the era of big data. Recent years we have witnessed significant success on deep learning-based models for drug-target binding affinity prediction task. However, these models only considered a single modality of drug and target information, and some valuable information was not fully utilized. In fact, the information of different modalities of drug and target can complement each other, and more valuable information can be obtained by fusing the information of different modalities. In this paper, we introduce a multimodal information fusion model for DTA that is called FMDTA, which fully considers drug/target information in both string and graph modalities and balances the feature representations of different modalities by a comparative learning approach. In addition, we exploited the alignment information of drug atoms and target residues to capture the positional information of string patterns, which can extract more useful feature information in SMILES and target sequences. Experimental results on two benchmark datasets show that FMDTA outperforms the state-of-the-art model, demonstrating the feasibility and excellent feature capture capability of FMDTA.

B832. Improving Biome Labeling for Tens of Thousands of Inaccurately Annotated Microbial Community Samples Based on Neural Network and Transfer Learning

Nan Wang[1], Teng Wang[1] and Kang Ning[1]

[1]huazhong university of science and technology

Abstracts:
Microbiome samples are accumulating at a fast speed, leading to millions of accessible microbiome samples in the public databases. However, due to the lack of strict meta-data standard for data submission and other reasons, there is currently a non-neglectable proportion of microbiome samples, especially those collected from the environment, in the public databases that have no annotations about where these samples were collected, how they were processed and sequenced, etc., among which the missing information about collection niches (biome) is one of the most prominent. The lack of sample biome information has created a bottleneck for the mining of the microbiome data, making it difficult in applications such as sample source tracking and biomarker discovery, especially in environmental scientific research. Here, we have designed Meta-Sorter, a neural network and transfer learning enabled AI method for improving the biome labeling of thousands of microbial community samples without detailed biome information. The results have shown that out of 16,507 samples that have no detailed biome annotations, 96.65% could be correctly classified, largely solving the missing biome labeling problem. Intriguingly, we succeeded in classifying certain representative environmental samples, which were sampled from benthic and water column but vaguely labeled as "Marine" in MGnify, in more details and with high fidelity. What's more, many of successfully predicted sample labels were from studies that involved human-environment interactions, for which we could also clearly differentiated samples from environment or human. Taken together, we have improved the completeness of biome label information for thousands of microbial community samples, facilitating sample classification from millions of microbiome samples, making knowledge discovery in multiple disciplines especially environmental researches more accurate.

B886. Multi-Modal Epileptic Seizure Detection Based on LogWT-CNN and MFCC

Chen Chen[1], Zhihan Zuo[1], Wei Tang[1], Yuchun Fang[1], Limin Hou[1] and Changfeng Chai[2]

[1]School of Computer Engineering and Science, Shanghai University
[2]Changhai Hospital, Shanghai

Abstract:
Most traditional methods of epileptic seizure detection depended on handcrafted electroencephalogram (EEG) features. Recently, the development of deep learning provides a new direction in EEG annotation. In this article, we propose the logarithmic wavelet transform (LogWT) to obtain the time-frequency images, enhance feature annotation with a pre-trained convolutional neural network (CNN), and create a new EEG representation method named LogWT-CNN. The LogWT serves to convert 1-dimensional EEG signals into 2-dimensional representations to make it possible for the CNN model to extract robust features. For epileptic seizure detection, we propose to combine two modalities, i.e., the LogWT-CNN and the Mel Frequency Cepstral Coefficient (MFCC) feature, through mapping them into independent prototype spaces. Experiment results show that the proposed LogWT-CNN reaches an average accuracy of 98.39% by cross-validation, exceeding other benchmark algorithms. Moreover, the fusion of LogWT-CNN and MFCC modalities gains an even higher accuracy of 98.65%.

B1075. DR-3DUNet: A Dense Residual 3D UNet for Lung Nodule Segmentation

Kafui Efio-Akolly[1], Hao Gui[1], Fei Luo[1], Ronald Bbosa[1], Feng Liu[1] and Yi Ping Phoebe Chen[2]

[1]School of Computer Science, Wuhan University
[2]Computer Science and Information Technology, La Trobe University, Melbourne, Australia

Abstracts:
The assessment of lung cancer requires a precise segmentation of lung nodules in Computed Tomography (CT) images.However, due to the difficulty in establishing the boundaries of some types of lung nodules with irregular shapes, varying densities and textures, as well as their similarities with the neighboring surroundings in the CT image, make it a complex task for existing methods to accurately segment certain types of lung nodules. This study proposes the DR-3DUNet to improve the general segmentation performance of the existing state-of-the-art 3D UNet model on lung nodules as well as improving the segmentation accuracy on those nodules with irregular shapes and structures. The proposed model adopts the structure of the 3D UNet model and is comprised of a Dense Residual Block (DR-B) in the skip connections. The DR-B is based on a three-dimensional convolution neural network with dense residual connections. It focuses on finer details from the image by enhancing and extracting essential features. The proposed model is evaluated on the LUNA16 dataset. When compared with the existing state-of-the-art models under the same experimental conditions, the results show that the proposed DR-3DUNet performs competitively and more efficiently with a DSC score of 83.02% and an ASD score of 0.37mm. The results also demonstrate that our proposed model performs effectively on the different lung nodule types especially on the small and sub-solid nodules which are more challenging to segment. The ablation study also reveals that the segmentation performance of the modified 3D UNet model significantly improves with the inclusion of the proposed DR-B module in the skip connections which signifies its importance.

B1182. DAMS-Net: Dual Attention and Multi-Scale Information Fusion Network for 12-lead ECG Classification

Rongzhou Zhou[1], Junfeng Yao[1], Qingqi Hong[1], Yuan Zheng[1] and Liling Zheng[2].

[1]Xiamen University
[2]First Hospital of Quanzhou Affiliated to Fujian Medical University

Abstracts:
Automated 12-lead electrocardiographic (ECG) classification algorithms play an important role in the diagnosis of clinical arrhythmias. Current methods that perform well in the field of automatic ECG classification are usually based on Convolutional Neural Networks (CNN) or Transformer. However, due to the intrinsic locality of convolution operations, CNN can't extract long-dependence between series. On the other side, the Transformer design includes a built-in global self-attention mechanism, but it doesn't pay enough attention to local features. In this paper, we propose DAMS-Net, which combines the advantages of Transformer and CNN, introducing a spatial attention module and a channel attention module using a CNN-Transformer hybrid encoder to adaptively focus on the significant features of global and local parts between space and channels. In addition, our proposal fuses multi-scale information to capture high and low-level semantic information by skip-connections. We evaluate our method on the 2018 Physiological Electrical Signaling Challenge dataset, and our proposal achieves a precision rate of 83.6%, a recall rate of 84.7%, and an F1-score of 0.839. The classification performance is superior to all current single-model methods evaluated in this dataset. The experimental results demonstrate the promising application of our proposed method in 12-lead ECG automatic classification tasks. Code will come soon at https://github.com/cdmc-zrz/DAMS-Net

B1402. Pipelined Biomedical Event Extraction Rivaling Joint Learning

Pengchao Wu[1], Xuefeng Li[1], Jinghang Gu[2], Longhua Qian[1] and Guodong Zhou[1].

[1]Soochow University
[2]the Hong Kong Polytechnic University

Abstracts:
Biomedical event extraction is an information extraction task to obtain events from biomedical text, whose targets include the type, the trigger, and the respective arguments involved in an event. Traditional biomedical event extraction usually adopts a pipelined approach, which contains trigger identification, argument role recognition, and finally event construction either using specific rules or by machine learning. In this paper, we propose an n-ary relation extraction method based on the BERT pre-training model to construct Binding events, in order to capture the semantic information about an event's context and its participants. The experimental results show that our method achieves promising results on the GE11 and GE13 corpora of the BioNLP shared task with F1 scores of 63.14% and 59.40%, respectively. It demonstrates that by significantly improving the performance of Binding events, the overall performance of the pipelined event extraction approach or even exceeds those of current joint learning methods.

B1439. A Novel Multilevel Iterative Training Strategy for the ResNet50 Based Mitotic Cell Classifier

Yuqi Chen[1], Juan Liu[1], Peng Jiang[1] and Yu Jin[1].

[1]Wuhan University

Abstracts:
The number of mitotic cells is an important indicator of grading invasive breast cancer. It is very challenging for pathologists to identify and count mitotic cells in pathological sections with naked eyes under the microscope. Therefore, many computational models for the automatic identification of mitotic cells based on machine learning, especially deep learning, have been proposed. However, converging to the local optimal solution is one of the main problems in model training. In this paper, we proposed a novel multilevel iterative training strategy to address the problem. To evaluate the proposed training strategy, we constructed the mitotic cell classification model with ResNet50 and trained the model with different training strategies. The results showed that the models trained with the proposed training strategy performed better than those trained with the conventional strategy in the independent test set, illustrating the effectiveness of the new training strategy. Furthermore, after training with our proposed strategy, the ResNet50 model with Adam optimizer has achieved 89.26% F1 score on the public MITOSI14 dataset, which is higher than that of the state-of-the-art methods reported in the literature.

B1652. Gene Regulatory Network Inference Based on Dynamic L0 Regularization

Xiong Li[1], Xu Meng[1], Yuchao Luo[1], Xing Li[1] and Juan Zhou[1]

[1]East China Jiaotong University

Abstract:
The single-cell RNA-seq (scRNA-seq) facilitated observation of the gene expression state of individual cells. Accurate inference of gene regulatory network (GRN) from scRNA-seq data is beneficial for understanding of cellular processes. Many computational methods have been developed for GRN inference, but few methods can accurately predict the key transcription factor that affect the changes of cellular processes. We proposed L0DWGRN (https://github.com/mengxu98/L0DWGRN), a toolkit for dynamic detecting causal regulatory interactions between transcription factor (TF) and genes. First of all, we used Slingshot infer pseudotimes, and then the cell states and pseudotimes were combined to re-divide the cells with a variable-length sliding-window approach. After that, we constructed GRN across state transition stages based on L0 regularization. The experimental results on 200 different types of datasets showed that the L0DWGRN has satisfactory performance.

B1661. SimHOEPI: a Resampling Simulator for Generating SNP Data with High-Order Epistasis Model

Xinrui Cai[1], Yan Sun[1], Junliang Shang[1], Feng Li[1], Boxin Guan[1] and Yuanyuan Zhang[2].

[1]Qufu Normal University
[2]Qingdao University of Technology

Abstracts:
Background
Epistasis is a ubiquitous phenomenon in genetics, and it is considered to be one of the main factors in current efforts to detect missing heritability for complex diseases. Simulated data is crucial for evaluating epistasis detection tools in genome-wide association studies. Existing simulators suffer from two limits. One is that some do not support high-order epistasis models containing multiple single nucleotide polymorphisms, and the other is that some cannot generate simulated data independently but rely on other software.
Results
In this study, we present a simulator, SimHOEPI, that can allow selection of prevalence or heritability to calculate the high-order epistasis model with the baseline penetrance provided by user. SimHOEPI generates the simulated data by a resampling method. Thus, the minor allele frequencies (MAFs) as that observed in the real data can be preserved in the generated simulated data. In addition, SimHOEPI provides a graphical user interface in order to make it more convenient for users.
Conclusions
SimHOEPI has three main properties: preservation of realistic MAFs, accuracy epistasis model embedding, and acceptable generating time. To evaluate the applicability of these three properties, we conduct experimental verifications from different aspects. Experimental results show that SimHOEPI can generate large-scale simulated data and has certain stability in running time. Meanwhile, the simulated data generated by SimHOEPI also well preserves the MAFs of real data.

B2038. BERT-5mC: an Interpretable Model for Predicting 5-Methylcytosine Sites of DNA Based on BERT

Shuyu Wang[1], Yinbo Liu[1], Yufeng Liu[1], Yong Zhang[1] and Xiaolei Zhu[1]

[1]Anhui Agricultural University

Abstract:
DNA 5-methylcytosine (5mC) is widely identified in multicellular eukaryotes, which is involved in a variety of developmental and physiological processes and a wide range of human diseases. Therefore, it is important to accurately detect the 5mC sites. Although current sequencing technologies are capable of mapping genome-wide 5mC sites, these experimental methods are both expensive and time-consuming. To achieve a fast and accurate prediction of 5mC, we propose a new computational approach, called BERT-5mC. We first pre-trained a new BERT model using human promoter sequences, and then fine-tuned and evaluated it on the 5mC datasets. Ultimately, on the independent tests, BERT-5mC showed better performance than other 5mC predictors. In addition, we analyzed the attention weights generated by BERT to identify a number of nucleotide distributions that are closely associated with 5mC modifications.

B2086. COVID-FPT: Feature Pyramid Transformer for COVID-19 Detection on CXR Images

Chang Liu[1], Guangchao Yang[1], Yunfei Yin[1] and Faliang Huang[2].

[1]College of Computer Science, Chongqing University
[2]Guangxi Key Lab of Human-machine Interaction and Intelligent Decision, Nanning Normal University

Abstracts:
Owing to its simplicity, rapidity, and low exposure to radiation, COVID-19 detection using Chest X-ray (CXR) images has been a research hotspot. However, some important semantic information is often lost in the computation process of the existing detection models, and location information of the lesions is also ignored in the existing models, such as TNT and SwinT. This may lead to unsatisfactory detection performance. To address the issue, we propose a multi-scale feature fusion neural network, named COVID-FPT (COVID-19 Feature Pyramid Transformer). In COVID-FPT, a Channel Attention Transformer Block (CAT Block) is designed to capture the extra semantic information, and a feature pyramid structure is devised to make full use of positional information via fusing different scale feature maps. Experimental results on three benchmark datasets indicate that, COVID-FPT outperforms the state-of-the-art COVID-19 detection algorithms, and demonstrate that COVID-FPT is promising for the COVID-19 detection task.

B2166. Genomic Variations Explorer (GenVarX): A Toolset for Annotating Promoter and CNV Regions Using Genotypic and Phenotypic Differences

Yen On Chan[1], Jana Biova[2], Anser Mahmood[1], Nicholas Dietz[1], Kristin Bilyeu[3], Mária Škrabišová[2] and Trupti Joshi[4].
[1]University of Missouri [2]Palacky University in Olomouc, Olomouc, Czech Republic
[3]Usda [4]University of Missouri, Columbia

Abstracts:
The rapid growth of sequencing technology and its increasing popularity in biology-related research over the years has made whole-genome re-sequencing (WGRS) data become widely available. A large amount of WGRS data can unlock the knowledge gap between genomics and phenomics through gaining an understanding of the genomic variations that can lead to phenotype changes. These genomic variations are usually comprised of allele and structural changes in DNA, and these changes can affect the regulatory mechanisms causing changes in gene expression and altering the phenotypes of organisms. In this research work, we created the GenVarX toolset that is backed by transcription factor binding sequence data in promoter regions, the copy number variations data, SNPs and Indels data, and phenotypes data which can potentially provide insights about phenotypic differences and solve compelling questions in plant research. Analytics-wise, we have developed strategies to better utilize the WGRS data and mine the data using efficient data processing scripts, libraries, tools, and frameworks to create the interactive and visualization-enhanced GenVarX toolset that encompasses both promoter regions and copy number variation analysis components. The main capabilities of the GenVarX toolset are to provide easy-to-use interfaces for users to perform queries, visualize data, and interact with the data. Based on different input windows on the user interface, users can provide inputs corresponding to each field and submit the information as a query. The data returned on the results page is usually displayed in a tabular fashion. In addition, interactive figures are also included in the toolset to facilitate the visualization of statistical results or tool outputs. Currently, the GenVarX toolset supports soybean, rice, and Arabidopsis. The researchers can access the soybean GenVarX toolset from SoyKB via https://soykb.org/SoybeanGenVarX/, rice GenVarX toolset, and Arabidopsis GenVarX toolset from KBCommons web portal with links https://kbcommons.org/system/tools/GenVarX/Osativa and https://kbcommons.org/system/tools/GenVarX/Athaliana, respectively.

B2213. Predicting the Human miRNA-disease Associations Based on Non-linear Gaussian Profile Kernel Similarity

Haittao Zou[1], Xiaolan Xie[1], Boya Ji[2] and Shaoliang Peng[2].
[1]Guilin University of Technology [2]Hunan University

Abstracts:
In view of the fact that the traditional methods of determining potential miRNA-disease associations tend to be destructive, labor-intensive, time-consuming, and associated with practice effects, a number of computational methods are being developed to address the burden on biological researchers. In this study, we introduced the computational strategy of non-linear gaussian profile kernel similarity and proposed a novel deep-learning method called NGPKS to engage the in-depth understanding of miRNA-disease associations. More specifically, NGPKS comprehensively integrates the miRNA functional similarity and disease semantic similarity information. Then, the gaussian interaction profile kernel similarity algorithm was utilized to capture the structural information between miRNAs and diseases. Finally, a deep learning framework was constructed for modeling the integration of two types of similarity features. We used three model validation strategies, including five-fold cross-validation, comparison with the state-of-the-art methods, and ablation experiments were used to check the predictive ability of our model. Besides, we conducted case studies for two common diseases. As a result, there are 50 (Colon Cancer), and 47(Lymphoma) among the top 50 predicted miRNAs validated through experiments. Therefore, we could conclude that NGPKS is an effective method to predict potential miRNA-disease associations.

B2310. Accurately Predicting Methylation Site Using Deep Convolutional Neural Network

Austin Spadaro, [1] Alok Sharma[2] and Iman Dehzangi[1].

[1]Rutgers University
[2]Riken

Abstracts:
Protein lysine methylation is a particular type of post translational modification that plays an important role in both histone and non-histone function regulation in proteins. Deregulation caused by lysine methyltransferases has been identified as the cause of several diseases including cancer as well as both mental and developmental disorders. Identifying lysine methylation sites is a critical step in both early diagnosis and drug design. This study proposes a new Machine Learning method called CNN-Meth for predicting lysine methylation sites using a convolutional neural network (CNN). Our model is trained using evolutionary, structural, and physicochemical-based presentation along with binary encoding. Unlike previous studies, instead of extracting handcrafted features, we use CNN to automatically extract features from different presentations of amino acids to avoid information loss. To the best of our knowledge, automated feature extraction from these representations of amino acids as well as CNN as a classifier have never been used for this problem. Our results demonstrate that CNN-Meth can significantly outperform previous studies found in the literature for predicting Methylation sites. It achieves 96.0%, 85.1%, 96.4%, and 0.65 in terms of Accuracy, Sensitivity, Specificity, and Matthew's Correlation Coefficient (MCC), respectively. CNN-Meth and its source code are publicly available at https://github.com/MLBC-lab/CNN-Meth

B2666. POLAT: Protein Function Prediction Based On Soft Mask Graph Network and Residue-Label ATtention

Yang Liu[1], Yi Zhang[1], Zihao Chen[1], Yi Zhang[1] and Jing Peng[1].

[1]Wuhan University of Technology

Abstracts:
Motivation: Elucidating the function of proteins is a central problem in biochemistry, genetics, and molecular biology. The huge gap between the sequence and functional data of proteins makes it a critical task to develop computational protein function prediction methods. Recent breakthroughs in protein structure prediction methods, coupled with the strong correlation between protein structure and function, make it a realistic idea to use protein structure to predict function. However, existing structure-based methods ignore that different residues contribute differently to the function and give no consideration to the correlation between protein residues and functions. How to effectively use the relationship between protein residues and functional information to predict protein function is still a problem to be solved.
Result: We proposed a Protein function prediction method based On soft mask graph networks and residue-Label ATtention(POLAT), which could combine sequence features, predicted structure features, and functional information to get an accurate prediction. We use soft mask graph networks to adaptively extract the residues relevant to functions. A residue-label attention mechanism is adopted to get the protein-level encoded feature of a protein, which is concatenated with a protein embedding and fed into a dense classifier to get the probabilities of each function. POLAT gets 0.670, 0.515, 0.578 of Fmax, 0.677, 0.409. 0.507 of AUPR on the PDB cdhit testset for the MFO, BPO, and CCO domains, respectively, outperforming the existing structure-based SOTA method GAT-GO(Fmax 0.633, 0.492, 0.547; AUPR 0.660, 0.381, 0.479). POLAT is also competitive on extensive experiments among sequence-based and multimodal methods and gets the SOTA performance in three of six metrics.

B2701. MMR: A Multi-view Merge Representation Model for Chemical-Disease Relation Extraction

Yi Zhang[1], Baitai Cheng[1], Yang Liu[1], Chi Jiang[1] and Jing Peng[1].

[1]Wuhan University of Technology

Abstracts:
Background: Chemical-Disease relation (CDR) extraction aims to identify thesemantic relations between chemical and disease entities in the unstructuredbiomedical document, which provides a basis for downstream tasks such asclinical medical diagnosis and drug discovery. Compared with general domainrelationship extraction, it requires a more comprehensive representation of theentire document due to the complex characteristics of entities and long textstructures in the biomedical domain.Results: In this paper, we propose a novel Multi-view MergeRepresentation(MMR) model that focuses on entity semantic representation fromthe local-view and entity pair interaction representation from the global-view tocapture more comprehensive textual information. First, in the local-view, we usea pre-trained transformer encoder and prior knowledge embedding to captureentity semantic representations containing contextual and knowledge-enhancedinformation. Then in the global-view, we employ the U-Net network layer and thegraph convolutional network layer to capture the global entity-pairrepresentation, respectively. Finally, we obtain a specific merged representationfor each entity pair to be classified. We evaluated our model on the CDR datasetreleased by the BioCreative-V community and achieved state-of-the-art results. Inaddition, we conducted experiments on the CHR dataset to validate theeffectiveness of our model.Conclusions: This paper integrates the information of entities and entity pairs inChemical-Disease relation extraction through a multi-view merge representationapproach. Experiments verify that this model can be effective for biomedicalrelationship extraction, especially for multi-entity and inter-sentence relationships.

B2804. A Hierarchical Model Based on Semantic Relation Extraction of Small Cell Lung Cancer Patents

Hua-Hui Gao[1], Rong Zhu[1], Jin-Xing Liu[1], Jun-Liang Shang[1] and Ying Guo[2].

[1]School of Computer Science, Qufu Normal University
[2]School of Computer Science and Engineering, Central South

Abstracts:
A patent document file typically stores multiple categories in a hierarchical structure. However, few classification models are available. In order to highlight the characteristics of the patent document, this paper proposes a hierarchical model focusing on the semantic features relations of the patent, which is used to capture the semantic relations among different levels. Firstly, the Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT) is used to implement word, hierarchy, and position embedding. Secondly, the Bidirectional Gating Recurrent Unit (BiGRU) neural network is used to fully learn the context information in the patent document, so as to obtain the entire semantic representation. Thirdly, the Hierarchical Attention-based Memory (HAM) unit is used to model the dependency relationship between different levels overhead to improve the expression ability of semantic information. Finally, the hybrid prediction layer is used to predict the local and global information, so that the model can not only predict the categories of each layer, but also accurately classify all categories in the complete hierarchy. We have conducted a large number of experiments on the small cell lung cancer patent document corpus established. The experimental results demonstrate that the proposed model is superior to the traditional model.

B2850. Multi-scale DCNN with Dynamic Weight and Part Cross-entropy Loss for Skin Lesion Diagnosis

Gaoshuai Wang[1], Linrunjia Liu[2], Fabrice Lauri[1] and Amir Hajjam El Hassani[1].
[1]Utbm [2]Xidian University

Abstracts:
Accurately diagnosing skin lesion disease is a challenging task. Even though present methods often applied the multi-branch model to get more clues, they failed to realize the performance instability of a branch with the heterogeneous disease zone size and the negative effect on generalization caused by the cross-entropy loss. To address these problems, we propose a multi-scale DCNN with dynamic weight and part cross-entropy loss model (MDP-DCNN). The proposed model is composed of three branches with different resolution images, which is prone to gain discriminative context information globally and locally. To alleviate the negative influence of the irrelevant zone, our model locates the essential part in the low-resolution image by the Gradient-weighted Class Activation Mapping (Grad-CAM), then crops the other two input images with defined sizes. Moreover, current works mainly define a fixed weight on each branch or assemble them directly, however, the performance of each branch is different and dynamic. To solve it, the proposed model manipulates the branch weight based on its CAM and input grey image. The cross-entropy loss, focusing on optimizing the targeted label's loss and ignoring the non-targeted labels' influence, could lead to over-fitting. But dealing with all labels will step forward to another over-fitting. Therefore, we propose the part cross-entropy, which optimizes the non-targeted label to decrease the influence on other labels' stability when the prediction is wrong. We conducted our model on the ISIC-2017 and ISIC-2018 datasets. Experimental results demonstrate that MDP-DCNN achieves excellent results in skin lesion classification without external data. Results with several loss functions and fused methods verified the advantages of the part cross-entropy loss and dynamic weight in enhancing the ensemble model's performance.

B2904. MRUNet-3D: a Multi-stride Residual 3D UNet for Lung Nodule Segmentation

Ronald Bbosa[1], Hao Gui[1], Fei Luo[1], Feng Liu[1], Kafui Efio-Akolly[1] and Yi-Ping Phoebe Chen[2].

[1]School of Computer Science, Wuhan University
[2]Department of Computer Science and Information Technology, La Trobe University

Abstracts:
Obtaining an accurate segmentation of the pulmonary nodules in computed tomography (CT) images is a challenging task. This is due to: 1) the heterogeneous nature of the lung nodules. 2) comparable visual characteristics between the nodules and their surroundings. A robust multi-scale feature extraction mechanism that can effectively obtain multi-scale representations at a granular level can improve segmentation accuracy. Being the most commonly used network in lung nodule segmentation, UNet and its variants as well as other image segmentation methods lack this kind of a robust feature extraction mechanism. In this study, we propose a multi-stride residual 3D UNet (MRUNet-3D) to improve the segmentation accuracy of the lung nodules in CT images.It incorporates a multi-slide Res2Net block (MSR) which replaces the simple sequence of convolution layers in each stage of the encoder to effectively extract multi-scale features at a granular level from different receptive fields and resolutions while conserving the strengths of 3D UNet. The proposed method has been extensively evaluated on the publicly available LUNA16 dataset. Experimental results show that it achieves competitive segmentation performance with an average dice similarity coefficient of 83.47% and average surface.
distance of 0.35 mm on the dataset. More notably, our method has shown to be robust to the heterogeneity of lung nodules.It has also proven to show better performance on the segmentation of small lung nodules. Ablation studies have shown that the the proposed MSR and FIA modules are very fundamental in improving the performance of the proposed model.

B3012. SMCC: a Novel Clustering Method for Single- and Multi-Omics Data Based on Co-Regularized Network Fusion

Sha Tian[1], Ying Yang[1], Yushan Qiu[1] and Quan Zou[2]
[1]Shenzhen University
[2]University of Electronic Science and Technology of China

Abstract:
Clustering is a common technique for statistical data analysis and is essential for developing precision medicine. Numerous computational methods have been proposed for integrating multi-omics data to identify cancer subtypes. However, most existing clustering models based on network fusion fail to preserve the consistency of the distribution of the data before and after fusion. Motivated by this observation, we would like to measure and minimize the distribution difference between networks, which may not be in the same space, to improve the performance of data fusion.
We were therefore motivated to develop a flexible clustering model, based on network fusion, that minimizes the distribution difference between the data before and after fusion by co-regularization; the model can be applied to both single- and multi-omics data.We propose a new network fusion model for single- and multi-omics data clustering for identifying cancer or cell subtypes based on co-regularized network fusion (SMCC). SMCC integrates low-rank subspace representation and entropy to fuse networks. In addition, it measures and minimizes the distribution difference between the similarity networks and the fusion network by co-regularization. The model can both reduce the noise interference in the source data and make the statistical characteristics of the fusion result closer to those of the source data. We evaluated the clustering performance of SMCC across 16 real single- and multi-omics dataset. The experimental results demonstrated that SMCC is superior to 17 state-of-the-art clustering methods. Moreover, it is effective for identifying cancer or cell subtypes, thereby promoting the development of precision medicine.

B3039. Tuning Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding

Nour Almadhoun Alserr[1], Gulce Kale[2], Onur Mutlu[1], Oznur Tastan[3] and Erman Ayday[4]

[1]ETH Zurich [2]Bilkent University
[3]Sabanci University [4]Case Western Reserve University

Abstract:
Researchers need a rich trove of genomic datasets that they can leverage to gain a better understanding of the genetic basis of the human genome and identify associations between phenotypes and specific parts of DNA. However, sharing genomic datasets that include sensitive genetic or medical information of individuals can lead to serious privacy-related consequences if data lands in the wrong hands. Restricting access to genomic datasets is one solution, but this greatly reduces their usefulness for research purposes. To allow sharing of genomic datasets while addressing these privacy concerns, several studies propose privacy-preserving mechanisms for data sharing. Differential privacy is one of such mechanisms that formalize rigorous mathematical foundations to provide privacy guarantees while sharing aggregated statistical information about a dataset. Nevertheless, it has been shown that the original privacy guarantees of DP-based solutions degrade when there are dependent tuples in the dataset, which is a common scenario for genomic datasets (due to the existence of family members).
In this work, we introduce a new mechanism to mitigate the vulnerabilities of the inference attacks on differentially private query results from genomic datasets including dependent tuples. We propose a utility-maximizing and privacy-preserving approach for sharing statistics by hiding selective SNPs of the family members as they participate in a genomic dataset. By evaluating our mechanism on a real-world genomic dataset, we empirically demonstrate that our proposed mechanism can achieve up to 40% better privacy than state-of-the-art DP-based solutions, while near-optimally minimizing utility loss.

B3401. Multi-scale Self-attention Multiple Instance Learning for Whole Slide Image Classification

Fengyu Tian[1], Changjian Wang[1], Ming Feng[2], Kele Xu[1], Zheng Qin[1] and Minpeng Xu[2]

[1]National University of Defense Technology
[2]Tongji University

Abstract:
The field of histopathology image analysis has made great progress since the deep learning revolution. Despite significant efforts, model performance is still severely limited by the following factors: Whole Slide Image (WSI) classification is limited by computational resources and makes limited use of knowledge at different scales. Multi-instance learning (MIL) has demonstrated its potential to extract powerful features in semi-supervised classification tasks, while the integration of multi-scale information from pathological images still needs to be improved. Existing multiexample learning approaches are usually based on the assumption of independent homogeneous distribution and use high-resolution images for processing, ignoring the correlation between instances and the quality of WSI to imply multi-scale information. In this paper, we design a multi-scale fusion MIL framework, using a dual-pipeline input mechanism for multi-scale local fusion and a global attention-based aggregation method to fuse the correlation information between instances while involving multi-scale images in the decision making simultaneously, so that the final generated bag embedding can more accurately describe the WSI. to verify the effectiveness of the method, we conducted experiments using two public datasets, Camelyon16 and TCGA-NSCLS. The obtained results show that the multi-scale fusion MIL framework can significantly improve the WSI classification performance compared to the state-of-the-art methods.

B3469. iLncDA-RSN: Identification of lncRNA-Disease Associations Based on Reconstructed Similarity Network

Junliang Shang[1], Mingrui Zhang[1], Feng Li[1], Boxin Guan[1], Qianqian Ren[1], Jin-Xing Liu[1] and Sun Yan[1]

[1]Qufu Normal University

Abstract:
Identification of disease-associated long non-coding RNAs (lncRNAs) is helpful to promote the important process of prevention, treatment and recovery from complex diseases. However, similarity networks for lncRNAs and diseases directly mix many different types of networks, which contain various noise-causing inefficient identifying potential lncRNA-disease associations（LDAs）. In this study, we propose a computational model iLncDA-RSN based on the constructed reliable lncRNA similarity network and disease similarity network to predict potential LDAs from different perspectives. To remove noise and reasonably confuse the different types of networks, reliable similarity networks for lncRNAs and diseases are constructed by random walk with restart on the non-standard similarity network. Based on constructed reliable similarity networks, LDAs are more efficiently identified from two different types of perspectively. Experimental results demonstrate that iLncDA-RSN is more suitable for predicting disease-associated lncRNAs.

B3565. ResNet Meets PCBAM and SE for Medical Visual Question Answering

Kai Liu[1], Chunping Liu[1], Ying Li[1] and Yi Ji[1]

[1]School of Computer Science and Technology, Soochow University

Abstract:
Medical image feature representation is essential for the medical visual question answering(VQA) tasks because medical image imaging methods cause noise and blurring, which reduces the image's contrast and visibility. In this paper, we employ visual attention mechanisms to enhance the representation of meaningful features and suppress useless information from medical images. Therefore, we propose a visual feature enhancement module(VFEM), including a squeeze-and-excitation(SE) block and an improved convolutional block attention module(CBAM). For the SE block, different channel feature calibration and enhanced feature representation are achieved by modelling the relationship between feature channels. Due to the serial structure of CBAM, the input of the spatial attention submodule is the feature decorated by channel attention, which limits the network's performance. To solve this problem, we propose parallel CBAM(PCBAM), where both spatial attention and channel attention directly take the original feature map as input to obtain the corresponding weights so that the order between the two submodules is not considered. The overall precision rose from 68.8% to 72.1% on the popular VQA-RAD dataset to demonstrate the effectiveness of our proposed method.

B3811. ACT-Tooth: A Semi-Supervised Tooth Volume Segmentation in CBCT images based on Asymmetric CNN-Transformer Network and Cross Consistency

Weiwei Cui[1], Yifan Zhang[2], Hongkun Wu[2], Huiyu Zhou[3], Liaoyuan Zeng[4], Bung San Chong[1], Qun Jin[5], Qianni Zhang[1] and Yaqi Wang[6].

[1]Queen Mary University of London
[2]Sichuan University
[3]University of Leicester)
[4]University of Electronic Science and Technology of China
[5]Waseda University
[6]Communication University of Zhejiang

Abstracts:
Accurate tooth volume segmentation is a prerequisite for computer-aided dental diagnosis and prognosis. Deep learning-based tooth segmentation methods have achieved satisfying performances but require a large quantity of tooth data with ground truth. To alleviate large annotation requirements, we propose a semi-supervised tooth volume segmentation method ACT-Tooth, which consists of two segmentation branches: 3D MTANet and 3D CTFormer. MTANet encodes local tooth features on the whole volumes while CTFormer explores global tooth features based on 3D sub-patches. This asymmetric design facilitates the extraction of representative tooth features from both the global and local views. To ensure stable learning, cross-consistency regularisation is proposed to calculate the segmentation loss of one branch based on the predicted results of this branch and the pseudo labels of the other branch. The proposed ACT-Tooth method is evaluated on an open 3D tooth dataset CTooth+. Experimental results show that ACT-Tooth achieves exhilarating performances in tooth volume segmentation, demonstrating comparable or even better accuracy than the existing state-of-the-art 3D medical Transformer-based methods. The codes will be made publicly available later at https://github.com/liangjiubujiu/ACT-Tooth.

B3985. Structure-Aware Sparse Transformer-Based Model for Predicting Drug-Target Binding Affinity

Zhengda He[1], Linjie Chen[2], Hao Lv[2], Ruining Zhou[2], Jiaying Xu[2], Jianhua Hu[2], Yadong Chen[2] and Yang Gao[1].

[1]Nanjing University
[2]China Pharmaceutical University

Abstracts:
Using Transformer model to process protein long sequence information for affinity prediction faces several challenges. Transformer requires high computational cost for protein long sequence modeling, and its unfocused attention and lack of protein structure information degrade model performance. In this study, we proposed, for the first time, a sequence-based computational contact map as structure-aware sparse attention for Transformer, and also ProbSparse attention based on protein sequence content.
In addition, our work also focuses on the design of the overall architecture of drug-target affinity prediction models. We also introduce the drug Graph-Transformer and the fine-grained parallel drug-target Co-Attention. The proposed model is interpretable. We evaluated our model on benchmark datasets Davis and KIBA. It outperforms current state-of-the-art deep learning approaches for drug-target binding affinity prediction. Our proposed method does not require the structural information of drug target complexes or protein crystals. It directly computes contact maps based on sequence information to further construct structure-based protein sparse Transformer, a strategy that significantly broadens the application scenarios.

B4041. Language Model based on Deep Learning Network for Biomedical Named Entity Recognition

Guan Hou[1], Yuhao Jian[1], Qingqing Zhao[1], Quan Xiongwen[1] and Han Zhang[1]

[1]Nankai University

Abstract:
Biomedical Named Entity Recognition (BioNER) is one of the most basic tasks in biomedical text mining, which aims to automatically identify and classify biomedical entities in text. Recently, deep learning-based methods have been applied to Biomedical Named Entity Recognition and have shown encouraging results. However, many biological entities are polysemous and ambiguous, which is one of the main obstacles to the task of biomedical named entity recognition. Deep learning methods require large amounts of training data, so the lack of data also affect the performance of model recognition.To solve the problem of polysemous words and insufficient data, for the task of biomedical named entity recognition, we propose a multi-task learning framework fused with language model based on the BiLSTM-CRF architecture. Our model uses a language model to design a differential encoding of the context, which could obtain dynamic word vectors to distinguish words in different datasets. Moreover, we use a multi-task learning method to collectively share the dynamic word vector of different types of entities to improve the recognition performance of each type of entity. Experimental results show that our model reduces the false positives caused by polysemous words through differentiated coding, and improves the performance of each subtask by sharing information between different entity data. Compared with other state-of-the art methods, our model achieved superior results in four typical training sets, and achieved the best results in F1 values.

B4052. DLP: Duplex Link Prediction via Subspace Segmentation for Predicting Drug-MiRNA Associations

Kai Zheng[1], Qichang Zhao[1], Mengyun Yang[2], Xiao Liang[1], Yiwei Liu[1] and Jianxin Wang[1].

[1]Central South University
[2]Shaoyang University

Abstracts:
Drug discovery is a long and costly process, and computational methods have the potential to accelerate the drug discovery process by analysis of bioassays or biomedical literature. There is evidence that miRNAs play an important role in disease development and can be considered as therapeutic targets. Therefore, the development of effective computational models to identify potential drug targets, like miRNAs, is urgently needed. In this study, a method based on subspace segmentation, called duplex link prediction (DLP), is proposed to identify potential miRNAs that can serve as drug targets. Specifically, we first use the network enhancement (NE) algorithm to update the similarity metric between miRNAs. Then, we obtain two matrices by pre-filling the association matrix from both drug and miRNA perspectives using the K nearest neighbors (KNN) method. After that, the DLP-base algorithm is used to predict the potential links in them. Finally, the final predicted association scores are obtained by the weighted average of the two matrices. Experimental results show that the DLP algorithm outperforms existing methods in the task of identifying miRNA as potential drug targets. Moreover, the case study confirms that the proposed method is effective in practical applications.

B4168. CLCAP: Contrastive Learning Improves Antigenicity Prediction for Influenza A Virus Using Convolutional Neural Networks

Biao Ye[1], Rui Yin[1], Changyu Yin[1] and Jiang Bian[1]

[1]University of Florida

Abstract:
Influenza viruses are detected year-round over the world and the viruses will usually circulate during fall and winter, causing the seasonal flu. The growing novel variants of influenza viruses pose a significant concern to public health annually. However, the rapid mutation of the influenza viruses makes it challenging to timely track their evolution. Therefore, a fast, low-cost, and precise method to predict the antigenic variant of influenza viruses could help vaccine development and prevent viral transmission. In this study, we propose a multi-channel convolutional neural network using contrastive learning to predict the antigenicity of influenza A viruses. An integrated dataset containing antigenic data and protein sequences was collected from various public resources and literature. The experimental results on three different influenza subtypes indicate our proposed model outperforms other traditional machine learning classifiers for antigenicity prediction. In addition, it also demonstrates superior performance over several state-of-the-art approaches, with 5.18%, 7.03% and 7.82% increase in accuracy compared to the best results for H1N1, H3N2 and H5N1, respectively. The proposed framework is timely and effective in influenza antigenicity prediction and can be adapted to the study of other viruses.

B4185. LPI-IBWA:Predicting lncRNA-protein Interactions Based on Improved Bi-Random Walk Algorithm

Minzhu Xie[1], Hao Wang[1] and Ruijie Xi[1]

[1]Hunan Normal University

Abstract:
Many studies have shown that long-chain noncoding RNAs (lncRNAs) are involved in a variety of biological processes such as post-transcriptional gene regulation, splicing, and translation by combining with corresponding proteins. Predicting lncRNA-protein interactions is an effective approach to infer the functions of lncRNAs.The paper proposes a new computational model named LPI-IBWA. At first, LPI-IBWA uses similarity kernel fusion (SKF) to integrate various types of biological information to construct lncRNA and protein similarity networks. Then, a bounded matrix completion model and a weighted k-nearest known neighbors algorithm are utilized to update the values for the potential interaction entries in the initial lncRNA-protein interaction matrix. Based on the updated lncRNA-protein interaction matrix, the lncRNA similarity network and the protein similarity network are integrated into a heterogeneous network. Finally, a Bi-Random walk algorithm is used to predict novel latent lncRNA-protein interactions. 5-fold cross-validation experiments on a benchmark dataset show that the AUC and AUPR of LPI-IBWA are 0.920 and 0.736, respectively, which are higher than those of other state-of-the-art methods. Furthermore, the experimental results of case studies on a novel dataset also illustrate that LPI-IBWA could efficiently predict potential lncRNA-protein interactions.

B4386. scMSSL: Multi-scale Attention Semi-Supervised Learning with Deep Generative Models to Automatically Identify Cell Types

Hongyu Duan[1], Feng Li[1], Junliang Shang[1], Daohui Ge[1], Xikui Liu[2] and Yan Li[2].

[1]Qufu Normal University
[2]Shandong University of Science and Technology

Abstracts:
Cells are the basic structural and functional units of living organisms. Different cell types perform different functions and play different roles in the development, normal functioning and development of diseases in the organism. The rapid development of high-throughput sequencing technologies in recent years has made it possible to obtain massive single-cell datasets. Analysis of these datasets yields findings at individual cellular resolution that reflect rich cellular heterogeneity. Single-cell datasets are more responsive to the needs of modern precision medicine than the tissue-based resolution datasets obtained by traditional sequencing technologies. However, single-cell datasets tend to have very large sample sizes and very high dimensional characteristics, and they are also very sparse and highly noisy. These characteristics of single-cell datasets make it difficult for researchers to accurately identify cell types, which directly affects a range of downstream analyses. To address these challenges, we propose scMSSL, a semi-supervised deep generation model based on multi-scale attention. scMSSL creatively introduces a multi-scale attention module based on the semi-supervised depth generation model, which cleverly unifies the self-attentive and convolutional layer neural networks. This module overcomes the limitations of self-attentive networks, which focus too much on global features, and allows the network to take into account both local and global features. scMSSL can thus effectively extract features from single-cell datasets, a feature that has been validated in performance comparisons with benchmark models. All source code involved in the paper can be accessed at https://github.com/FengLi12/scMSSL.

B4408. A New Parallel Spiking Neural Network Simulator Using Sunway TaihuLight

Xuelei Li[1], Zhichao Wang[1], Yi Pan[1], Jintao Meng[1], Shengzhong Feng[1] and Yanjie Wei[1]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Abstract:
Spiking neural network simulation is very important for studying the brain function and validating the hypotheses for neuroscience, and real-time simulation of large spiking neural networks is often a challenge due to the highly random memory access pattern of neurons and the demand for high simulation accuracy. In this paper, we propose a new efficient and parallel spiking neural network simulator based on the SW26010 architecture of TaihuLight supercomputer, named SWsnn, which supports highly accurate brain simulation with very small time step ($\sim$1/16 ms) and randomly distributed delay sizes for the synapses. Compared with NVIDIA GPUs, the unique feature of SW26010 about local dynamic memory (LDM) shared via register level communication (RLC) among all the computing processing elements (CPEs) allows user to control data exchange between CPEs directly. Many core group communication is realized via MPI in SWsnn; to reduce communication cost, SWsnn relies on parallel CPE clusters rather than serial MPEs to control the many core-group communication, using RLC, DMA and MPI. In addition, SWsnn is further optimized using vectorization, memory access hiding techniques. Experimental results show that SWsnn can simulate real-time brain activity for 210 000 neurons using 32 core-groups of TaihuLight, and runs 2.5$\sim$4.0 times faster than state-of-the-art simulator GeNN running on GPUs with similar floating point peak performance.

B4482. DDI-Transform: A DDI-Transform Neural Network for Predicting Drug-Drug Interaction Events

Jiaming Su[1] and Ying Qian[1].

[1]East China Normal University

Abstracts:
Drug-drug interaction (DDI) events prediction is a challenging problem, and accurate prediction of DDI events is critical to patient health and new drug development. Recently, many machine learning-based techniques have been proposed for predicting DDI events. However, most of the existing methods do not integrate well the multidimensional features of drugs and no good mitigation of noise information to get effective feature information. To address these limitations, we propose a DDI-Transform neural network framework for DDI events prediction. In DDI-Transform, we design a drug structure information feature extraction module and a drug bind-protein feature extraction module to obtain drug multidimensional feature information. Then a stack of DDI-Transform layers in the DDI-Transform network module are used for adaptive learning, thus adaptively selecting the effective feature information for prediction. The results show that DDI-Transform can accurately predict DDI events and outperform the state-of-the-art models. And the results on different scale datasets confirm the robustness of the method.

B4510. A Multimodal Data Extraction Pipeline for MIMIC with Parallelization

Yutao Dou[1], Wei Li[2], Albert Y. Zomaya[2] and Shaoliang Peng[3].

[1]Hunan University & University of Sydney
[2]University of Sydney
[3]Hunan University

Abstracts:
Medical big data with artificial intelligence are vital in advancing digital medicine. However, the opaque and non-standardised nature embedded in most medical data extraction is prone to batch effects and has become a significant obstacle to reproducing previous works.
This paper aims to develop an easy-to-use time series multimodal data extraction pipeline, Quick-MIMIC, for standardised data extraction from MIMIC datasets.
Our method can fully integrate different data structures into a time-series table, including structured, semi-structured, and unstructured data.
We also introduce two additional modules to Quick-MIMIC, a pipeline parallelization method and data analysis methods, for reducing the data extraction time and presenting the characteristics of the extracted data intuitively.
The extensive experimental results show that our pipeline can efficiently extract the needed data from the MIMIC dataset and convert it into the correct format for further analytic tasks.

B4580. SemiCTrans: Semi-Supervised Medical Image Segmentation Framework combining CNN and Transformer

Fei Xu[1], Zihan Li[1], Qingqi Hong[1], Qingqiang Wu[1], Qingde Li[2] and Jie Tian[3].

[1]Xiamen university
[2]University of Hull
[3]Institute of Automation, Chinese Academy of Sciences

Abstracts:
In medical image segmentation, both local details and global information are important. CNN has demonstrated excellent detail capture ability in many visual tasks, but due to the limitation of the receptive field, its ability to capture global information is insufficient. Transformer is just the opposite, which can model long-range dependency well, but it is insufficient for capturing local details. Therefore, we construct a dual-branch medical image segmentation model based on CNN and Transformer to take advantages of their complementary strengths. Furthermore, for the scarcity of labeled data in medical image segmentation, we incorporate semi-supervised learning to further improve the accuracy and robustness of the model by leveraging the information of a large amount of unlabeled data. Experiments on two public datasets show that our method can achieve excellent performance in both fully-supervised and semi-supervised configurations. Particularly, with very small amount of labeled data, our semi-supervised model can still achieve remarkable results comparable to those of full supervision.

B4770. Predicting Drug Target Interaction Based on BERT Model and Subsequence Embedding

Zhihui Yang[1], Juan Liu[1], Feng Yang[1], Xiaolei Zhang[1], Qiang Zhang[1], Xuekai Zhu[1] and Peng Jiang[1]

[1]WuHan University

Abstract:
Exploring the relationship between proteins and drugs plays a significant role in discovering new synthetic drugs. The Drug-Target Interaction (DTI) prediction is a fundamental task in the relationship between proteins and drugs. Unlike encoding proteins by amino acids, we use amino acid subsequence to encode proteins, which simulates the biological process of DTI better. For this research purpose, we proposed a novel deep learning framework based on Bidirectional Encoder Representation from Transformers (BERT), which integrates high-frequency subsequence embedding and transfer learning methods to complete the DTI prediction task. As the first key module, subsequence embedding allows to explore the functional interaction units from drug and protein sequences and then contribute to finding DTI modules. As the second key module, transfer learning promotes the model learn the common DTI features from protein and drug sequences in a large dataset. Overall, the BERT-based model can learn two kinds features through the multi-head self-attention mechanism: internal features of sequence and interaction features of both proteins and drugs, respectively. Compared with other methods, BERT-based methods enable more DTI-related features to be discovered from general features of proteins and drugs through transfer learning. We conducted extensive experiments for the DTI prediction task on three different benchmark datasets. The experimental results show that the model achieves an average prediction metrics higher than most baseline methods. In order to verify the importance of transfer learning, we conducted an ablation study on datasets, and the results show the superiority of transfer learning. In addition, we test the scalability of the model on the dataset in unseen drugs and proteins, and the results of the experiments show that it is acceptable in scalability.

B4906. PIN: a Penalized Integrative Deep Neural Network for Variable Selection among Multiple Omics Datasets

Yang Li[1], Xiaonan Ren[2], Haochen Yu[2], Tao Sun[1] and Shuangge Ma[3].

[1]Center for Applied Statistics, School of Statistics,Renmin University of China
[2]School of Statistics, Renmin University of China
[3]Department of Biostatistics, Yale University

Abstracts:
Deep learning has been increasingly popular in omics data analysis. Recent works incorporating variable selection into deep learning have greatly enhanced the model's interpretability. However, because deep learning desires a large sample size, the existing methods may result in uncertain findings when the dataset has a small sample size, commonly seen in omics data analysis. With the explosion and availability of omics data from multiple populations/studies, the existing methods naively pool them into one dataset to enhance the sample size while ignoring that variable structures can differ across datasets, which might lead to inaccurate variable selection results. We propose a penalized integrative deep neural network (PIN) to simultaneously select important variables from multiple datasets. PIN directly aggregates multiple datasets as input and considers both homogeneity and heterogeneity situations among multiple datasets in an integrative analysis framework. Results from extensive simulation studies and applications of PIN to gene expression datasets from elders with different cognitive statuses or ovarian cancer patients at different stages demonstrate that PIN outperforms existing methods with considerably improved performance among multiple datasets.

B4979. Deep Imputation Bi-Stochastic Graph Regularized Matrix Factorization for Single-cell RNA-Sequencing Data Clustering

Wei Lan[1], Jianwei Chen[1], Mingyang Liu[1], Ruiqing Zheng[2], Jin Liu[2] and Yi Pan[3]

[1]Guangxi University
[2]Central South University
[3]Shenzhen Institute of Advanced Technology

Abstract:
The application of fruitful achievement of single-cell RNA-sequencing (scRNA-seq) technology has generated huge amount of gene transcriptome data. It has provided a whole new perspective to analyze the transcriptome at single-cell level. Cluster analysis of scRNA-seq is an efficient approach to reveal unknown heterogeneity and functional diversity of cell populations, which could further assist researchers to explore pathogenesis and biomarkers of diseases. In this paper, we propose a new cluster method (DSINMF) based on deep matrix factorization to detect cell type in the scRNA-seq data. In our method, the feature selection is used to reduce redundant features. Then, the imputation method is utilized to impute dropout events. Further, the dimension reduction is utilized to reduce the impact of noise. Finally, the deep matrix factorization with bi-stochastic graph regularization is employed to cluster scRNA-seq data. To evaluate the performance of DSINMF, eight datasets are used as test sets in the experiment. The experimental results show DSINMF outperformances than other state-of-the-art methods in clustering performance.

B5169. AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes

Jeremie S. Kim[1], Can Firtina[1], Meryem Banu Cavlak[1], Damla Senol Cali[2], Nastaran Hajinazar[1], Mohammed Alser[1], Can Alkan[3] and Onur Mutlu[1]

[1]ETH Zurich
[2]Carnegie Mellon University
[3]Bilkent University

Abstract:
AirLift is the first read remapping tool that enables users to quickly and comprehensively map a read set, that had been previously mapped to one reference genome, to another similar reference. Users can then quickly run downstream analysis of read sets for each latest reference release. Compared to the state-of-the-art method for remapping reads (i.e., full mapping), AirLift reduces the overall execution time to remap read sets between two reference genome versions by up to 27.4×. We validate our remapping results with GATK and find that AirLift provides high accuracy in identifying ground truth SNP/INDEL variants.

Code Availability. AirLift source code and readme describing how to reproduce our results are available at https://github.com/CMU-SAFARI/AirLift.

B5515. Fully Automated Annotation of Mitochondrial Genomes Using a Cluster-Based Approach with De-Bruijn Graphs

Lisa Fiedler[1], Martin Middendorf[1] and Matthias Bernt[2]

[1]Leipzig University
[2]Helmholtz Centre for Environmental Research

Abstract:
The recent development of new sequencing techniques leads to strongly increasing amounts of available mitochondrial sequence data. This generates a need for highly efficient automatic annotation methods. Automatic annotation methods are typically based on databases that contain knowledge on already annotated (and often pre-curated) mitogenomes of different species. However, the existing annotation methods have several shortcomings: (i) they do not scale well in the size of the database, (ii) do not allow for a fast (and easy) update of the database, and/or (iii) can be applied only to a relatively small taxonomic subset of all species.
In this work, a new automatic annotation method, DeGeCI, is presented that does not have any of the shortcomings (i), (ii), and (iii). DeGeCI uses a reference database where the annotated mitogenomes are represented as a de-Bruijn graph. The annotation process for a new user-supplied mitogenome starts off by mapping its sequence to the database graph. The graph is searched for trails of high sequence similarity whenever sequence segments cannot be matched exactly. Heuristic quality measures are used to limit the number of trails that need to be scanned. Finally, gene predictions are generated with a clustering approach that uses the joint annotations of a database subgraph.
DeGeCI proves to generate gene predictions of high conformity with expert annotations for a large set of mitogenomes for which expert-curated annotations are known. In a comparative evaluation with MITOS2, a state-of-the-art annotation tool for mitochondrial genomes, DeGeCI shows better database scalability while still matching MITOS2 in terms of result quality and offering a fully automated means to update the underlying database. Moreover, unlike MITOS2, DeGeCI can be run in parallel on several processors to make use of modern multi-processor systems.

B5527. RLFDR: A Graph Representation Learning Framework for Drug Repositioning over Heterogeneous Information Networks

Bo-Wei Zhao[1], Xiao-Rui Su[1], Meng-Long Zhang[1], Zhu-Hong You[2], Peng-Wei Hu[3] and Lun Hu[3]

[1]The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences
[2]School of Computer Science, Northwestern Polytechnical University
[3]The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

Abstract:
Drug repositioning is a promising development strategy to discover potential drug candidates for diseases. However, existing computational models lack the ability to appropriately handle multiple relationships available in biological heterogeneous information networks (HINs). To this end, a novel representation learning framework, called RLFDR, is proposed for improved performance of drug repositioning. More specifically, RLFDR first constructs a function similarity graph and several semantic relation subgraphs to capture the characteristics of drug-disease associations (DDAs) from the functional and semantic perspectives. Then, the representations of drugs and diseases are learned by using different graph representation learning strategies. Finally, RLFDR adopts a Random Forest classifier to discover novel DDAs by concatenating the embeddings of drugs and diseases. Experimental results demonstrate that RLFDR achieves superior performance over state-of-the-art DDA prediction models on three benchmark datasets. Furthermore, our case studies indicate that the simultaneous consideration of function similarity and semantic relationships of drugs and diseases allows RLFDR to precisely predict DDAs in a more comprehensive manner.

B5528. A Feature Extraction Framework for Discovering Pan-Cancer Driver Genes Based on Multi-Omics Data

Xiaomeng Xue[1], Feng Li[1], Junliang Shang[1], Lingyun Dai[1], Daohui Ge[1] and Qianqian Ren[2].

[1]Qufu Normal University
[2]Shandong Normal University

Abstracts:
Identification of tumor driver genes facilitates accurate diagnosis and treatment of cancer, but how to identify driver genes from the large number of genes plays an important role in precision oncology research. And the signaling and regulation of genes and the interaction of protein complexes are equally crucial. In this work, we construct a feature extraction framework for discovering pan-cancer driver genes based on multi-omics data. We combine multi-omics pan-cancer data (mutations, gene expression, copy number variants and DNA methylation) with protein-protein interaction (PPI) networks, and use a network propagation algorithm to mine functional information among nodes in the PPI network, especially genes with weak node information. Then, the distribution features of the pan-cancer data are extracted from the obtained gene functional features. Meanwhile, the TOPSIS features of the pan-cancer data are extracted using the ideal solution method for the pan-cancer data under the gene functional features. We further extract the SetExpan features of the pan-cancer data using the gene functional features, which is a method to rank pan-cancer data according to their average inverse ranking. Finally, we use the lightGBM classification algorithm to predict genes. The experimental results show that our method generally outperforms existing methods in terms of area under the check precision-recall curve (AUPRC) and has better performance compared to other methods on different PPI networks. Our study can predict more potential cancer genes for biological scientists.

B5582. GenMPI: Cluster Scalable Variant Calling for Short/Long Reads Sequencing Data

Tanveer Ahmad[1] and Zaid Ars[1].
[1]Tu Delft

Abstracts:
Rapid technological advancements in sequencing technologies allow producing cost effective and high volume sequencing data. Processing this data for real-time clinical diagnosis is potentially time-consuming if done on a single computing node. This work presents a complete variant calling workflow, implemented using the Message Passing Interface (MPI) to leverage the benefits of high bandwidth interconnects. This solution (GenMPI) is portable and flexible, meaning it can be deployed to any private or public cluster/cloud infrastructure. Any alignment or variant calling application can be used with minimal adaptation. To achieve high performance, compressed input data can be streamed in parallel to alignment applications while uncompressed data can use internal file seek functionality to eliminate the bottleneck of streaming input data from a single node. Alignment output can be directly stored in multiple chromosome-specific SAM files or a single SAM file. After alignment, a distributed queue using MPI RMA (Remote Memory Access) atomic operations is created for sorting, indexing, marking of duplicates (if necessary) and variant calling applications. We ensure the accuracy of variants as compared to the original single node methods. We also show that for 300x coverage data, alignment scales almost linearly up to 64 nodes (8192 CPU cores). Overall, this work outperforms existing big data based workflows by a factor of two and is almost 20\% faster than other MPI-based implementations for alignment without any extra memory overheads. Sorting, indexing, duplicate removal and variant calling is also scalable up to 8 nodes cluster. For pair-end short-reads (Illumina) data, we integrated the BWA-MEM aligner and three variant callers (GATK HaplotypeCaller, DeepVariant and Octopus), while for long-reads data, we integrated the Minimap2 aligner and three different variant callers (DeepVariant, DeepVariant with WhatsHap for phasing (PacBio) and Clair3 (ONT)). All codes and scripts are available at: https://github.com/abs-tudelft/gen-mpi

B5970. Synthesis Cost-Optimal Targeted Mutant Protein Libraries

Dimitris Papamichail[1], Madeline Febinger[1], Shm Almeda[1] and Georgios Papamichail[2]

[1]The College of New Jersey
[2]New York College

Abstract:
Protein variant libraries produced by site-directed mutagenesis are a useful tool utilized by protein engineers to explore variants with potentially improved properties, such as activity and stability. These libraries are commonly built by selecting residue positions and alternative beneficial mutations for each position. All possible combinations are then constructed and screened, by incorporating degenerate codons at mutation sites. These degenerate codons often encode additional unwanted amino acids or even STOP codons. Our study aims to take advantage of annealing based recombination of oligonucleotides during synthesis and utilize multiple degenerate codons per mutation site to produce targeted protein libraries devoid of unwanted variants. Toward this goal we created an algorithm to calculate the minimum number of degenerate codons necessary to specify any given amino acid set, and a dynamic programming method that uses this algorithm to optimally partition a DNA target sequence with degeneracies into overlapping oligonucleotides, such that the total cost of synthesis of the target mutant protein library is minimized. Computational experiments show that, for a modest increase in DNA synthesis costs, beneficial variant yields in produced mutant libraries are increased by orders of magnitude, an effect particularly pronounced in large combinatorial libraries.

B6529. Automatic Coarse-to-Refinement Kidney Segmentation in Ultrasound Images

Tao Peng[1], Yidong Gu[2] and Jing Cai[3].

[1]Soochow University
[2]Suzhou Municipal Hospital
[3]Hong Kong Polytechnic University

Abstracts:
Because of missing or ambiguous boundaries, developing accurate segmentation methods for ultrasound images of the kidney is a significant challenge. In this study, we developed a coarse-to-refinement method with four novel aspects. First, we used the characteristics of a principal curve (PC) to fine-tune the shape of the curve automatically, and used the learning ability of a neural network to de-crease the model error. Second, a deep fusion learning network was used for the coarse segmentation step, in which a parallel architecture was used to improve the deep-learning performance. Third, to solve the issue that standard PC-based methods cannot determine the number of vertices automatically, we proposed an automatic searching polygon tracking method for the first time, using a mean shift clustering-based method to replace the projection and vertex extension step of standard PC-based methods. Fourth, an explainable mathematical map function of the kidney contour was developed, as denoted by the output of the neural net-work (i.e., optimized vertices), which matched well with the ground truth con-tour. Several experiments were performed to evaluate the performance of our method.

B6539. Knowledge Enhanced Attention Aggregation Network for Medicine Recommendation

Jiedong Wei[1], Yijia Zhang[1], Xingwang Li[1], Mingyu Lu[1] and Hongfei Lin[2].

[1]Dalian Maritime University
[2]Dalian University of Technology

Abstracts:
Recently, the close combination of deep learning and the medical field has achieved great success. An important aspect is to recommend medicine for the patient. Patients often have repeated medical information during their clinical records, and this information has a significant impact on the patient's condition. Most existing methods model longitudinal patient information, ignoring the impact of individual diagnoses and procedures on the patient's health, which may lead to insufficient patient representation, thus limiting the accuracy of medicine recommendations. Therefore, we propose a medicine recommendation model KEAN, which is based on an attention aggregation network and enhanced graph convolution. More specifically, KEAN can aggregate the individual diagnoses and procedures in patient visit sequences to capture significant features that affect patients' diseases. In addition, KEAN learns medicine knowledge from complex medicine combinations, reduces drug-drug interactions (DDIs), and recommends medicines beneficial to patients' health. The experimental results on the MIMIC-III dataset show that our model is superior to the existing advanced methods.

B6643. JLONMFSC: Clustering scRNA-seq Data Based on Joint Learning of Non-Negative Matrix Factorization and Subspace Clustering

Wei Lan[1], Mingyang Liu[1], Jianwei Chen[1], Jin Ye[1], Ruiqing Zheng[2], Xiaoshu Zhu[3] and Wei Peng[4]

[1]Guangxi University
[2]Central South University
[3]Yulin Normal University
[4]Kunming University of Science and Technology

Abstract:
The development of single cell RNA sequencing (scRNA-seq) has provided new perspectives to study biological problems at the single cell level. One of the key issues in scRNA-seq data analysis is to divide cells into several clusters for discovering the heterogeneity and diversity of cells. However, the existing scRNA-seq data are high-dimensional, sparse and noisy, which challenges the existing single-cell clustering methods. In this study, we propose a joint learning framwork (JLONMFSC) for clustering scRNA-seq data. In our method, the dimension of original data is reduced to minimize the effect of noise. In addition, the graph regularized matrix factorization is used to learn the local features. Further, the Low-Rank Representation (LRR) subspace clustering is utilized to learn the global features. Finally, the joint learning of local features and global features is performed to obtain the results of clustering. We compare the proposed algorithm with other eight state-of-the-art algorithms for clustering performance on six datasets, and the experimental results demonstrate that the JLONMFSC achieves optimal performance in all datasets.

B6730. HKFGCN: A Novel Multiple Kernel Fusion Framework on Graph Convolutional Network to Predict Microbe-Drug Associations

Ziyu Wu[1], Shasha Li[2], Lingyun Luo[1] and Pingjian Ding[1]

[1]University of South China
[2]The University of Hong Kong

Abstract:
Accumulating clinical studies have shown that the microbes in the human body closely interact with the human host and participate in regulating the energy efficiency of drugs. Identifying the link between microbes and drugs can facilitate the development of drug discovery and reuse, so microbes have become a new target for antimicrobial drug development. However, most of the links between microbes and drugs were discovered by biological experiments, which are time-consuming, expensive, and sometimes risky. Therefore, it is necessary to leverage computational ways to predict microbe-drug associations to aid biological experiments. In this study, we propose a new method, called HKFGCN (Heterogeneous information Kernel Fusion Graph Convolution Network), to predict the microbe-drug associations. Instead of mixing different topological information together, HKFGCN extracts different features of topological information separately, and further extracts Gaussian kernel features after extracting features. HKFGCN consists of three main steps. Firstly, we constructed two similarity networks and a microbe drug association network based on numerous biological data, including drug similarity and microbe similarity. Secondly, we used two kinds of encoders to extract features from the two networks. Based on the extracted features, we further obtained Gaussian kernel features from each layer of drug and microbe, and fused them. Finally, we reconstructed the bipartite microbe-drug graph based on the learned representations. Experiments show that the HKFGCN model performs well on the three datasets through the cross-validation method. In addition, case studies were conducted on HIV and the results were confirmed by existing literature.

B7071. Prediction of CRISPR/Cas9 Editing Repair Outcomes Using Blended Machine Learning and Distributed Hyperparameter Optimization

Quang Hien Kha[1], Thi Oanh Tran[1], Phung-Anh Nguyen[1] and Nguyen Quoc Khanh Le[1].

[1]Taipei Medical University

Abstracts:
CRISPR/Cas9 is a technology that can be used to edit genes within organisms. This editing process has various applications, including fundamental biological research, developing biotechnological products, and treating diseases. Understanding repair outcomes after Cas9-induced DNA cleavage is still limited, especially in primary human cells. We used the repair outcomes data at 1,656 on-target genomics sites in primary human T cells to train a machine learning model to accurately predict the length, probability, and diversity of nucleotide insertions and deletions. The developed model will facilitate the design of SpCas9 guide RNAs in therapeutically important primary human cells. Machine learning algorithms, including eXtreme Gradient Boosting (XGBoost), Convolutional Neural Network (CNN), and Light Gradient Boosting Machine (LightGBM), were stacked together with hyper-parameter tuning, which achieved the RMSE of 0.1012 and R2 of 0.5573. We also noticed that Bayesian optimization could bring more advanced results than sequential grid search techniques regarding the hyperparameter tuning process. Furthermore, our model outperformed the previous CRISPR/Cas9-related outcome repair prediction benchmark. We believe our framework can help avoid unwanted pathogenic mutations with promising performance.

B7076. Efficient Sequencing Data Compression and FPGA Acceleration Based on a Two-Step Framework

Shifu Chen[1], Zhouyang Wang[1], Wenjian Qin[2], Yaru Chen[1], Jing Zhang[1], Heera Nand[3], Jishuai Zhang[3], Xiaoming Liang[3] and Mingyan Xu[1]

[1]HaploX Biotechnology
[2]Shenzhen Institutes of Advanced Technology
[3]Xilinx Inc.

Abstract:
With the increasing throughput of modern sequencing instruments, the cost of storing and transmitting sequencing data has also increased dramatically. Although many tools have been developed for compressing sequencing data, there is still a need to develop a compressor with a higher compression ratio. In this paper, we present a two-step framework for compressing sequencing data. The first step is to repack original data into a binary stream, while the second step is to compress the stream with a LZMA encoder. We present a strategy to encode the original file into a LZMA highly compressible stream. We also present an FPGA-accelerated implementation of LZMA to speedup the second step. As a demonstration, we present repaq as a lossless non-reference compressor of FASTQ format files. We introduce multi-file redundancy elimination, which is very useful for compressing paired-end sequencing data. According to our test results, this tool has a much higher compression ratio than other FASTQ compressors. For some deep sequencing data, the compression ratio of repaq can be higher than 25, almost 4 times of Gzip. The framework presented in this paper can also be applied to develop new tools for compressing other sequencing data. This tool is available at: https://github.com/OpenGene/repaq

B7675. DTKGIN: Predicting Drug-Target Interactions Based on Knowledge Graph and Intent Graph

Yi Luo[1], Qichang Zhao[1], Guihua Duan[1] and Jianxin Wang[1].

[1]Central South University

Abstracts:
Predicting drug-target interactions(DTIs) plays a crucial role in drug discovery and drug development. Considering the high cost and risk of biological experiments, developing computational approaches to explore the interactions between drugs and targets can effectively reduce the time and cost of drug development. Recently, many methods have made significant progress in predicting DTIs. However, existing approaches still suffer from the high sparsity of DTI datasets and the cold start problem. In this paper, we develop a new model to predict drug-target interactions via a knowledge graph and intent graph named DTKGIN. Our method can effectively capture biological environment information for targets and drugs by mining their associated relations in the knowledge graph and considering drug-target interactions at a fine-grained level in the intent graph. DTKGIN learns the representation of drugs and targets from the knowledge graph and the intent graph. Then the probability of interactions between drugs and targets is obtained through the inner product of the representation of drugs and targets. Experimental results show that our proposed method outperforms other state-of-the-art methods in 10-fold cross-validation, especially in cold-start experimental settings. Furthermore, the case studies demonstrate the effectiveness of DTKGIN in predicting potential drug-target interactions. The code is available on GitHub: https://github.com/Royluoyi123/DTKGIN.

B7880. A Novel Meta Sparse Learning Method for Brain Imaging Genetics without Individual-level Data

Duo Xi[1], Dingnan Cui[1], Minjianan Zhang[1], Zhang Jin[1], Muheng Shang[1], Lei Guo[1], Junwei Han[1] and Lei Du[1]

[1]Northwestern Polytechnical University

Abstract:
Imaging genetics is an emerging neuroscience topic that aims to identify risk genetic variations based on neuroimaging measurements. In this field, sparse learning (SL) methods have achieved great success. Generally, SL runs on individual-level imaging and genetic data, and thus can not work when the individual-level data is unable to access. We propose a meta SL (metaSL) method, which takes advantage of the summary statistics from genome wide association study (GWAS) on neuroimaging quantitative traits (QTs). metaSL endows the conventional SL methods with the capability of being insensible of the original individual-level imaging and genetic data. On the one hand, metaSL directly runs on the widely available summary statistics. On the other hand, metaSL possesses the merits of conventional SL methods such as modeling and feature selection. We first evaluate our approach on three real imaging genetic data sets generated from the Alzheimer's Disease Neuroimaging Initiative (ADNI) where the individual-level data is available. The results are very encouraging in the sense of being comparable to conventional SL on modeling error and risk loci identification. We additionally evaluate our method on two independent GWAS data sets, one comes from white matter microstructures and the other is from whole brain imaging phenotypes, where the original data was inaccessible. The results show that metaSL can not only replicate GWAS' significant loci but also identify interesting structures within SNPs that were missed by GAWS. These results suggest that metaSL possesses equivalent modeling and feature selection capability to conventional SL while freeing the constraints of requiring the original individual-level imaging and genetic data.

B8006. A New Network-Based Multi-Omics Pathway Analysis Method

Li Zhou[1], Jie Li[1], Dechen Xu[1], Dong Wang[1], Qiaoming Liu[1], Jiahuan Jin[1] and Yadong Wang[1]

[1]Harbin institute of technology

Abstract:
Background: Currently, several multi-omics pathway analysis methods have been proposed to identify significant pathways between two phenotypic. However, these methods either involve fewer data types or do not design a better scheme to integrate multi-omics data. To overcome these shortcomings, we proposed a new network-based multi-omics pathway analysis method named EMS (Expression, Methylation, Somatic).

Results: In the proposed method, Principal Component Analysis, Sparse Canonical Correlation Analysis, and random walk with restart algorithm are utilized to integrate protein-protein interaction network, gene expression, DNA methylation, and somatic mutation profiles. We tested the proposed method with lung, breast, and liver cancer datasets. We gained similar results under different cancer datasets. Firstly, judged on the existing biological knowledge and scholarly research, the top 15 pathways had a close relationship with corresponding cancer. Secondly, cancer driver genes were significant enrichment in the top 30 pathways. Finally, compared with other pathway analysis methods, the proposed method could pick out the pathways which have significant biological meaning.

Conclusion: Results imply that the proposed method overcomes the shortcomings of the existing pathway analysis method and can identify more accurately and reliably significant pathways from multiple perspectives.

B8098. GFDet: Multi-level Feature Fusion Network for Caries Detection Using Dental Endoscope Images

Nan Gao[1], Yukai Li[1], Peng Chen[1], Jijun Tang[2] and Tianshuang Liu[3]

[1]Zhejiang University of Technology
[2]University of South Carolina Columbia
[3]Hangzhou Stomatology Hospital

Abstract:
Early dental caries detection by endoscope can prevent complications such as pulpitis and apical infection. However, the need for sufficient saliency of caries features and the significant differences in caries size are tremendous challenges to identifying caries automatically. To address these problems, we propose the GFDet model, which incorporates a feature selection pyramid network (FSPN) and an adaptive assignment-balanced mechanism(AABM). FSPN performs upsampling with the semantic information of adjacent feature layers to mitigate the semantic information loss due to sharp channel reduction and collect global contextual features for discriminative feature enhancement. In addition, a new label assignment mechanism is proposed that enables the model to select more high-quality samples as positive samples, which can address the problem of small objects that are easily ignored. Meanwhile, we have built an endoscopic dataset for caries detection, consisting of 1318 images labeled by five dentists. For experiments on the collected dataset, the F1-score of our model is 75.6%, which out-performances the state-of-the-art models by 7.1%.

B8100. Re-examine Statistical Relationships among Dietary Fats, Risk Factors, and Cardiovascular Disease Risks based on Two Crucial Datasets

Jiarui Ou[1] and Le Zhang[1].

[1]Sichuan University

Abstracts:
The cardiovascular disease is the major cause of death in many regions, and several of its risk factors might be linked to diets. To improve understanding and public health of this topic, we look at the recent Minnesota Coronary Experiment (MCE) analysis that used paired t-test and Cox model. However, these parametric methods might suffer from problems of small sample size, right-censored bias, and lack of long-term evidence. To overcome these three challenges respectively, we utilize a nonparametric permutation test, a resampling-based rank test, and extra Framingham Heart Study (FHS) data with an A/B test. We show that, firstly, the causality between unsaturated fat diets and reduction in the cholesterol is certain; secondly, the link of reducing cholesterol leading to an increased CVD hazards is not robustly established in the diet group; lastly, the A/B test result suggests a more complicated relationship that abnormal blood pressure ranges by diets might affect the associative link between the cholesterol level and heart disease risks. This study not only helps us derive more reasonable results with the MCE dataset, but also reveals possible complex relationships behind diets, risk factors, and heart diseases with the long-term FHS data.

B8120. Tensor Improve Equivariant Graph Neural Network for Molecular Dynamics Prediction

Chi Jiang[1], Yi Zhang[1], Yang Liu[1] and Jing Peng[1].
[1]School of Computer Science and Artificial Intelligence, Wuhan University of Technology

Abstracts:
Background：Molecular dynamics(MD) simulations are essential for molecular structure optimization, drug-drug interactions, and other fields of drug discovery by simulating the motion of microscopic particles to calculate their macroscopic properties (e.g., energy). The main problems of the existing work are as follows: (1) Failure to fully consider the chemical bonding constraints between atoms,(2) Group equivariance can help achieve robust and accurate predictions of MD under arbitrary reference transformations and should be incorporated into the model design,(3) Tensor information such as relative position, velocity, and torsion angle can be used to enhance the prediction of molecular dynamics. And the existing methods are mainly limited to the scalar domain.Results：In this paper, we propose a new model—tensor improve equivariant graph neural network for molecular dynamics prediction(TEGNN):(1) The model materialization of chemical bond constraints between atoms into geometric constraints. The molecule's forward kinematic information (position and velocity) is represented by generalized coordinates. In this way, the interatomic chemical bonding constraints are implicitly and naturally encoded in the forward kinematics,(2) The equivariant information transfer is allowed in TEGNN, which significantly improves the accuracy and computational efficiency of the final prediction,(3) TEGNN introduces equivariant locally complete frames into scalar-only equivariant graph neural networks, thus allowing the projection of tensor information of a given order onto the frame. On the N-body dataset of simulated molecular systems consisting of particles, sticks, and hinges, as well as on the real dataset MD17 for molecular dynamics prediction, multiple experiments support that TEGNN is ahead of the current state-of-the-art GNN in terms of prediction accuracy, constraint satisfaction, and data efficiency.Conclusions：We extend the current state-of-the-art equivariant neural network model. The proposed TEGNN accommodates more tensor information and considers the chemical bonding constraints between atoms during motion, ultimately improving the performance of predicting the kinematic states of molecules.

B8524. Planning Biosynthetic Pathways of Target Molecules Based on Metabolic Reaction Prediction and AND-OR Tree Search

Xiaolei Zhang[1], Juan Liu[1], Feng Yang[1], Qiang Zhang[1], Zhihui Yang[1] and Hayat Ali Shah[1]

[1]Institute of Artificial Intelligence, School of Computer Science, Wuhan University

Abstract:
Bioretrosynthesis problem is to predict synthetic routes using substrates for given natural products (NPs). However, the huge number of metabolic reactions leads to a combinatorial explosion of searching space, which is high time-consuming and costly. Here, we propose a framework called BioRetro to predict bioretrosynthesis pathways using a one-step bioretrosynthesis network, termed HybridMLP combined with AND-OR tree heuristic search. The HybridMLP predicts precursors that will produce the target NPs, while the AND-OR tree generates the iterative multi-step biosynthetic pathways. The one-step bioretrosynthesis prediction experiments are conducted on MetaNetX dataset by using HybridMLP, which achieves 46.48\%, 74.60\%, 81.57\%, 85.77\%, 89.64\% in terms of the top-1, top-5, top-10, top-20, top-50 accuracies. The great performance demonstrates the effectiveness of HybridMLP in one-step bioretrosynthesis. Besides, the evaluation of two benchmark datasets reveals that BioRetro can significantly improve the speed and success rate in predicting biosynthesis pathways. In addition, the BioRetro is further shown to find the synthetic pathway of compounds, such as ginsenoside F1 with the same substrates as reported but different enzymes, which may be the novel potential enzyme to have better catalytic performance.

B8532. scLRSSC: a Low Rank and Sparse Subspace Clustering Framework for scATAC-seq

Xi Shen[1], Yongxin He[1] and Ruiqing Zheng[1].

[1]Central South University

Abstracts:
Single-cell ATAC-seq reveals the accessibility of chromatin at the level of a single cell, which can be used to study the epigenetic heterogeneity among cells. However, due to the high dimensionality and extreme sparseness of scATAC-seq data, it is difficult for computational methods to analyze and mine the valuable biological information based on scATAC-seq data. In order to overcome these limitations, we propose scLRSSC, a low rank and sparse subspace clustering method for analyzing cell heterogeneity based on scATAC-seq. scLRSSC applies the measurement TPM from RNA-seq to normalize the original reads counts matrix and combines low-rank and sparse structures to construct the similarities between cells. Then, it uses ADMM algorithm to solve the optimization problem. The obtained similarity matrix can be integrated with spectral clustering algorithm and t-SNE visualization for downstream analysis. Compared with other clustering methods on scATAC-seq, scLRSSC achieves more accurate and robust results on multiple datasets.

B8595. MMAR-net: a Multi-stride and Multi-resolution Affine Registration network for CT images

Fu Zhou[1], Fei Luo[1], Ruoshan Kong[1], Yi Ping Phoebe Chen[2] and Feng Liu[3]

[1]School of Computer Science, Wuhan University
[2]Department of Computer Science and Information Technology, La Trobe University
[3]School of Computer Science, Wuhan University

Abstract:
The evolution of lung lesions can be assessed by examining multiple CT screenings, which requires to align two CT images accurately. In this study, we propose a Multi-stride and Multi-resolution Affine Registration network, called MMAR-net, for 3D affine registration of medical images, which works in an unsupervised way by optimizing the similarity loss. In order to extract more extensive image features, we use a multi-stride module to replace the conventional convolution module. Furthermore, we make use of the image features at multiple scales by dot product between two feature vectors, which could enhance the robustness of image representation. We conduct comprehensive comparison experiments between our proposed model and the existing affine registration methods on two publicly available datasets, DIR-Lab and Learn2Reg (2020), which are both relevant to lung CT image registration. Quantitative and qualitative comparison results demonstrate that our model outperforms existing single-step affine registration networks. Our method improves the key metric of dice similarity coefficient on DIR-Lab and Learn2Reg to 90.57 % and 95.51 %.

B8847. TargetCall: Eliminating the Wasted Computation in Basecalling via Pre-Basecalling Filtering

Meryem Banu Cavlak[1], Gagandeep Singh[1], Mohammed Alser[1], Can Firtina[1], Joël Lindegger[1], Mohammad Sadrosadati[1], Nika Mansouri Ghiasi[1], Can Alkan[2] and Onur Mutlu[1].

[1]Eth Zurich
[2]Bilkent University

Abstracts:
Basecalling is an essential step in nanopore sequencing analysis where the raw signals of nanopore sequencers are converted into nucleotide sequences, i.e., reads. State-of- the-art basecallers employ complex deep learning models to achieve high basecalling accuracy. This makes basecalling computationally-inefficient and memory-hungry; bottlenecking the entire genome analysis pipeline. However, for many applications, the majority of reads do no match the reference genome of interest (i.e., target reference) and thus are discarded in later steps in the genomics pipeline, wasting the basecalling computation.To overcome this issue, we propose TargetCall, the first fast and widely-applicable pre-basecalling filter to eliminate the wasted computation in basecalling. TargetCall's key idea is to discard reads that will not match the target reference (i.e., off-target reads) prior to basecalling. TargetCall consists of two main components: (1) LightCall, a lightweight neural network basecaller that produces noisy reads; and (2) Similarity Check, which labels each of these noisy reads as on-target or off-target by matching them to the target reference. TargetCall filters out all off-target reads before basecalling; and the highly-accurate but slow basecalling is performed only on the raw signals whose noisy reads are labeled as on-target.Our thorough experimental evaluations using both real and simulated data show that TargetCall 1) improves the end-to-end basecalling performance of the state-of-the-art basecaller by 3.31× while maintaining high (98.88%) sensitivity in keeping on- target reads, 2) maintains high accuracy in downstream analysis, 3) precisely filters out up to 94.71% of off-target reads, and 4) achieves better performance, sensitivity, and generality compared to prior works. We freely open-source TargetCall to aid future research in pre-basecalling filtering at https://github.com/CMU-SAFARI/TargetCall.

B9348. KD_ConvNeXt: Knowledge Distillation-Based Images Classification of Lung Tumor Surgical Specimen Sections

Zhaoliang Zheng[1], Henian Yao[2], Chengchuang Lin[1], Kaixin Huang[1], Luoxuan Chen[1], Ziling Shao[3], Haiyu Zhou[2] and Gansen Zhao[1]

[1]South China Normal University, Key Lab on Cloud Security and Assessment technology of Guangzhou
[2]The First School of Clinical Medicine, Guangdong Medical University, Zhanjiang
[3]Jinan University-University of Birmingham Joint Institute at Jinan University, Guangdong

Abstract:
Background: In clinical practice, accurately identifying the specific subtypes of lung cancer is an essential task in diagnosing and treating lung lesions. The pathological classification of lung tumors mainly relies on CT images and paraffin pathology results for analysis. However, the analysis of CT images is only suitable for early screening and diagnosis of lung cancer, and paraffin pathology results usually take a week to be available.
Method: This paper aims to collect natural images of surgical specimen sections of lung tumors to construct a clinical dataset, and to try to use deep learning techniques to study and explore the problems of classifying specific subtypes of lung tumors. We proposes a teacher-student network architecture called KD_ConvNeXt, hoping to assist the clinical application, which is based on the knowledge distillation mechanism for specific subtype classification of lung tumor surgical specimen section images. Our approach enables the student

network (ConvNeXt) to extract knowledge from the intermediate feature layers of the teacher network (Swin Transformer) to improve ConvNeXt's feature extraction and fitting ability. Also, Swin Transformer provides soft labels containing information about the distribution of images in different categories, making the model focused more on the information carried by categories with smaller sample sizes while training.The identification of specific subtypes of lung tumors allows rapid assessment of the surgical approach, thus supporting the surgeon in arranging the next step of the surgical process and reducing the risk of surgery.
Results and Conclusions: We have designed many experiments on a clinical lung tumor image dataset, and the KD_ConvNeXt achieved the best classification accuracy of 85.17% and F1-score of 0.7568 compared with other advanced image classification methods. The quantitative evaluation results demonstrated the effectiveness and advancedness of the proposed method for the image classification problems of lung tumor surgical specimen sections, which can be a future reference to assist physicians in deciding subsequent surgical steps and treatment strategies.

B9980. MLRR-ATV: A Robust Manifold Nonnegative Low-Rank Representation with Adaptive Total-Variation Regularization for scRNA-seq Data Clustering

Gaofei Wang[1], Juan Wang[1], Shasha Yuan[1], Chunhou Zheng[1] and Jinxing Liu[1]

[1]School of Computer Science, Qufu Normal University Rizhao, Shandong

Abstract:
Since genomics was proposed, the exploration of genes has been the focus of research. The emergence of single-cell RNA sequencing (scRNA-seq) technology makes it possible to explore gene expression at the single-cell level. Due to the limitations of sequencing technology, the data contains a lot of noise. At the same time, it also has the characteristics of high-dimensional and sparse. Clustering is a common method of analyzing scRNA-seq data. Due to the characteristics of scRNA-seq data, the traditional clustering methods are facing great challenges and cannot achieve particularly ideal results. This paper proposes a novel single-cell clustering method called Robust Manifold Nonnegative Low-Rank Representation with Adaptive Total-Variation Regularization (MLRR-ATV). The Adaptive Total-Variation (ATV) regularization is introduced into Low-Rank Representation (LRR) model to reduce the influence of noise through gradient learning. Then, the linear and nonlinear manifold structures in the data are learned through Euclidean distance and cosine similarity, and more valuable information is retained. Because the model is non-convex, we use the Alternating Direction Method of Multipliers (ADMM) to optimize the model. We tested the performance of the MLRR-ATV model on eight real scRNA-seq data sets and selected nine state-of-the-art methods as comparison methods. The experimental results show that the performance of the MLRR-ATV model is better than the other nine methods.

Posters

Tittle	Authors
Heterogeneous Knowledge Graph Reasoning Dialogue System for Symptom Prediction and Disease Matching	Mingjiang Tang, Yanlong Qiu, Zhichang Zhang, Ziqin Zhang and Ruirui Han.
ML-PRDF：A Syndrome Differentiation Model of TCM based on PCC-MLRF and Multi-label Deep Forest	Lj Gong.
An automated pipeline to extract the Drosophila modular transcription regulators and targets	Tzu-Hsien Yang, Sheng-Hang Wu, Fang-Yuan Zhang, Hsiu-Chun Tsai, Ya-Chiao Yang, Yan-Yuan Tseng and Wei-Sheng Wu.
DEE-Net: A Dual-encoder Enhanced Network based on Transformer and CNN for Biomedical Image Classification	Zhiqiang Li, Xiaogen Zhou and Tong Tong.
Comparative Research on Mechanism of Fang Long Yi Zheng Qi Fang and Lianhua Qingwen Fang in Treating COVID-19	Xianfang Wang.
An efficient convolutional neural network-based diagnosis system for citrus fruit diseases	Zhangcai Huang, Xiaoxiao Jiang, Shaodong Huang and Su Yang.
Automatic Hemiplegia Gait Assessment for Stroke Rehabilitation by an Attention-based Lightweight CNN	Chengju Zhou, Daqin Feng, Lewei He, Nianming Ban, Shuxi Wang and Jiahui Pan.
A Feature-fusion based Deep Learning Approach for Method Type IV Secretory Effector Proteins	Qi Le, Baoqi Huang, Bing Jia and Runze Yang.
HR-MolBERT: A Molecule Representation Model Enhanced by Hydrogen and Radius based on Morgan Fingerprint	Shuyan Fang, Yu Liu, Along Hou, Huanhuan Qin and Song Liu.
Trilinear Distillation Learning and Question Feauture Capturing for Medical Visual Question Answering	Shaopei Long, Yong Li, Xiaobo Qian, Kun Zeng, Fu Lee Wang and Tianyong Hao.
An unsupervised deep learning framework	Guo Mao, Zhengbin Pang, Xiangdong Pei, Ke Zuo and Jie Liu.

Literature-Related Biomedical Knowledge Graph Based on Distant Supervision Relation Extraction	Rui Hua, Zixin Shu, Dengying Yan, Kuo Yang, Xinyan Wang, Chuang Cheng, Qiang Zhu and Xuezhong Zhou.
A Clinical Trial Termination Prediction Model based on Denoising Autoencoder and Deep Survival Regression (DAE-DSR)	Huamei Qi, Wenhui Yang, Wenqin Zou and Yuxuan Hu.
Transferring labels from scRNA-seq to scATAC-seq data with neighborhood contrastive regularization	Xuhua Yan, Ruiqing Zheng and Min Li
BioBERT-Enhanced Bi-directional Long Short-term Memory Model with Attention and gating Mechanisms for Protein-Protein Interactions Article classification	Zhan Tang, Xiaochuang Yao, Yuantian Xia, Xupeng Kou, Hongcheng Xue and Lin Li.
Bioinformatics analysis of the human microRNA-disease associations based on macrophage polarization and immune response during cholestatic liver injury	Elise Slevin, Ying Wan, Jennifer M. Salinas, Wenjuan Xu1, Sugeily Ramos Lorenzo, Bingru Zhou, and Fanyin Meng
DGCA: Distribution-Guided Context-Aware Two-stage Framework for Cerebral microbleeds Segmentation in MRI	Tianxiang Xia, Rong Zhang, Lijun Guo and Zhongding Fang.
A-RFP: An adaptive residue flexibility prediction method improves protein-ligand docking based on homologous proteins	Chuqi Lei, Senbiao Fang, Yaohang Li, Fei Guo, and Min Li

Transportation Manual

报到地点： 长沙福盛源大酒店(中南大学店)
Registration Site： Fushengyuan Hotel (Central South University)

会　场： 中南大学信息楼
Conference Venue： Information Building, Central South University

住宿： 长沙福盛源大酒店(中南大学店)
Hotel： Fushengyuan Hotel (Central South University)

Please print this NOTE and show the note to the taxi driver if you take a taxi at Changsha. The driver will take you to the destination:

- **Please take me to the Fushengyuan Hotel (Central South University) at the intersection of Fengshun Road and Huanghe Road, Yuelu District, Changsha, China.**
  (Chinese: 请送我到湖南省长沙市岳麓区丰顺路与黄鹤路交汇处的长沙福盛源大酒店)
- **Please take me to the Changsha Huanghua International Airport.**
  (Chinese: 请送我到长沙黄花国际机场)

国内的参会者，可以参考以下交通信息，也可以使用高德等导航软件获取帮助。

终点 / 起点	长沙福盛源大酒店(中南大学店)	中南大学新校区信息楼
长沙黄花国际机场	地铁 6 号线(黄花机场 T1T2 地铁站 3 号口☮朝阳村)☮地铁 3 号线（朝阳村☮阳光地铁站 3 号口）	地铁 6 号线(黄花机场 T1T2 地铁站 3 号口☮朝阳村)☮地铁 3 号线(朝阳村☮中南大学地铁站 2 号口）
	打车预计需要 71 元	打车预计需要 74 元
长沙南站（高铁站）	地铁 4 号线(长沙火车南站地铁站☮阜埠河地铁站)☮地铁 3 号线（阜埠河地铁站☮阳光地铁站 3 号口）	地铁 4 号线(长沙火车南站地铁站☮阜埠河地铁站)☮地铁 3 号线(阜埠河地铁站☮中南大学地铁站 2 号口）
	打车预计需要 34 元	打车预计需要 36 元
长沙火车站	地铁 3 号线(长沙火车站地铁站☮阳光地铁站 3 号口）	地铁 3 号线(长沙火车站地铁站☮中南大学地铁站 2 号口）
	打车预计需要 31 元	打车预计需要 27 元

终点 起点	长沙黄花国际机场	长沙南站（高铁站）	长沙火车站
长沙福盛源大酒店（中南大学店）	地铁 3 号线(阳光地铁站 3 号口✇朝阳村)✇地铁 6 号线（朝阳村✇黄花机场 T1T2 地铁站）	地铁 3 号线(阳光地铁站 3 号口✇阜埠河地铁站)✇地铁 4 号线（阜埠河地铁站✇长沙火车南站地铁站）	地铁 3 号线(阳光地铁站 3 号口✇长沙火车站地铁站）
	打车预计需要 71 元	打车预计需要 34 元	打车预计需要 31 元
中南大学新校区信息楼	地铁 3 号线(中南大学地铁站 2 号口 3 号口✇朝阳村)✇地铁 6 号线（朝阳村✇黄花机场 T1T2 地铁站）	地铁 3 号线(中南大学地铁站 2 号口✇阜埠河地铁站)✇地铁 4 号线（阜埠河地铁站✇长沙火车南站地铁站）	地铁 3 号线(中南大学地铁站 2 号口✇长沙火车站地铁站）
	打车预计需要 74 元	打车预计需要 36 元	打车预计需要 27 元

Hotel Accommodations

Fushengyuan Hotel (Central South University)

Intersection of Fengshun Road and Huanghe Road (North end of Sunshine One hundred International New City), Yuelu District, Changsha, Hunan, 410012, China

长沙福盛源大酒店(中南大学店)，中国，湖南，长沙，岳麓区，丰顺路与黄鹤路交汇处（阳光一百国际新城北端）

Accommodation Information

We are glad to help you book the hotel in advance. If necessary, you can choose the type of room you need and the duration of your stay at the registration screen. The room details are as follows.

Room	Reference Price(per room/night)
Premium Single Room	CNY 368
Deluxe king room	CNY 328
Deluxe twin room	CNY 288

Sponsors