会议程序册

CBC2017

THE SECOND CCF BIOINFORMATICS CONFERENCE

第二届中国计算机学会生物信息学会议

2017年10月13-15日 中国·长沙

主办单位

中国计算机学会

承办单位

中南大学

深圳市早知道科技有限公司

会议日程一览表

10 月 13 日会议安排（地点：湘府国际酒店）		
10：00-21：00	注册、报到	湘府国际酒店一楼大厅
18：00-19：00	晚餐	1 楼百合厅
19：00-22：00	生物信息专委委员会议	3 楼聚缘厅
10 月 14 日会议安排（地点：湘府国际酒店）		
8：30-9：15	开幕式	18 楼国际会议厅
9：15-12：00	主会场报告	18 楼国际会议厅
9：15-10：05	Edwin.Wang（加拿大卡尔加里大学）：From health genomics to intelligent precision health	
10：05-10：20	茶歇	
10：20-11：10	张学工（清华大学）：找不同 ： 从基因芯片到单细胞测序	
11：10-12：00	郭茂祖（北京建筑大学）： 基于基因本体的功能相似网络构建与纯化算法	
12：10-13：30	午餐	1 楼百合厅
13：30-18：15	分会场报告	
分会场一	基因组、计算表观遗传及疾病研究	18 楼国际会议厅
分会场二	基因表达、调控大数据与分子进化	19 楼阳光会议室
分会场三	蛋白质组学	3 楼聚缘厅
分会场四	大数据分析与精准医疗	18 楼培训室
分会场五	Poster	18 楼国际会议厅
18：30-20：00	晚宴	1 楼百合厅
10 月 15 日会议安排（地点：湘府国际酒店）		
8：30-12：00	主会场报告	18 楼国际会议厅
8：30-9：20	刘志勇（中国科学院计算技术研究所）： 生物大分子三维重构与高性能计算	
9：20-10：10	陈洛南(中国科学院上海生命科学研究院）:Diagnosing Un-occurred Diseases with Single Sample by Dynamic Network Biomarkers——Detecting the tipping points of biological processes by big data	
10：10-10：30	茶歇	
10：30-12：00	企业报告	
12：10-13：30	午餐	1 楼百合厅
13：30-18：10	分会场报告	
分会场一	机器学习、自然语言处理在生物信息中的应用	18 楼培训室
分会场二	超级计算与生物医学影像	19 楼阳光会议室
YOCSEF 论坛	人工智能与智慧医疗特别论坛	18 楼国际会议厅
18:10-18:30	闭幕式	18 楼国际会议厅
18：30-19：30	晚餐	1 楼百合厅

组织机构

主办单位：

中国计算机学会

承办单位：

中南大学

协办单位：

深圳市早知道科技有限公司

大会主席：

高琳　　陈翔

程序委员会：

主席：李敏 陈钢　彭绍亮

委员：

蔡宏民　蔡瑞初　陈伯林　陈　钢　邓　磊　邓明华　邓赵红　董启文　杜朴风
高敬阳　高　琳　龚新奇　郭茂祖　何增有　雷秀娟　李国君　李国亮　李　敏
刘　滨　刘　辉　刘　娟　刘　琦　刘卫国　刘学军　毛国君　宁　康　彭绍亮
尚学群　宋丹丹　汪国华　王建新　於东军　袁志勇　张道强　张　法　张世华
张艳菊　章　乐　赵兴明　周丰丰　周水庚　朱大铭　朱　敏　朱山风　邹　权
段　磊　冯好娣　甘杨兰　关佶红　郭　菲　郭杏莉　呼加璐　黄德双　姜　伟
蒋庆华　李　杰　李小波　廖明帜　刘丙强　刘金星　路永钢　马小科　潘林强
彭　玮　彭佳杰　彭小清　苏　冉　万晓华　王　飞　王炳波　王春宇　王　峻
魏乐义　魏彦杰　谢民主　叶　凯　叶明全　余国先　鱼　亮　赵宇海　郑春厚
郑文萍　钟端洋　陈禹保　金　钟　吴红艳　朱　峰　刘　礼

交通和会场位置

会场与住宿位置：

本次会议的会场及住宿均在湘府国际酒店

湖南省长沙市天心区竹塘西路 179 号（中南大学铁道学院旁）

机场、高铁站、火车站到会场的交通路线：

1. 长沙黄花国际机场到湘府国际酒店

黄花国际机场到湘府国际酒店方案
方案一：出租车　总路程：26 公里/33 分钟　参考费用：65 元
黄花国际机场乘车至湘府国际酒店
方案二：机场大巴民航酒店线转 105 路　总路程：33 公里　参考费用:
黄花机场站上车至民航酒店站下车
步行 370 米至长岛路口站（105 路）
105 路公交车至竹塘路口站下车，步行 480 米至湘府国际酒店

2. 长沙高铁站到湘府国际酒店

高铁站到湘府国际酒店方案
方案一：出租车　总路程：9.4 公里/23 分钟　参考费用：20 元
高铁站乘车至湘府国际酒店
方案二：68 路　总路程：10.5 公里　参考费用:2 元
步行 1.1 公里至长托站上车
乘坐 68 路至铁道学院站下车，步行 790 米至湘府国际酒店

3. 长沙火车站到湘府国际酒店

火车到湘府国际酒店方案
方案一：出租车　总路程：9.4 公里/17 分钟　参考费用：20 元
方案二：地铁一号线转地铁二号线　总路程：13.5 公里/46 分钟
步行 350 米至长沙火车站上车乘坐地铁 2 号线 （梅溪湖西方向） 至 五一广场站下车
乘坐地铁 1 号线 （尚双塘方向）至铁道学院站（4 口出）下车，步行至湘府国际酒店
方案二：公交车 7 路　总路程：10 公里/49 分钟 参考费用：2 元
步行 350 米至长沙火车站上车，乘坐 7 路 至 铁道学院站下车
步行 1.1 公里至湘府国际酒店

会议日程安排

<table>
<tr><td colspan="6">10 月 13 日会议安排（地点：湘府国际酒店）</td></tr>
<tr><td colspan="2">时间</td><td colspan="4">会议安排</td></tr>
<tr><td colspan="2">10：00-21：00</td><td colspan="4">注册、报到（湘府国际酒店一楼大厅）</td></tr>
<tr><td colspan="6">晚餐（18：00-19：00）</td></tr>
<tr><td colspan="2">19：00-22：00</td><td colspan="4">生物信息专委委员会议（地点：聚缘厅 3 楼）</td></tr>
<tr><td colspan="6">10 月 14 日会议安排（地点：湘府国际酒店）</td></tr>
<tr><td colspan="2">时间</td><td colspan="2">仪式内容</td><td colspan="2">演讲嘉宾</td></tr>
<tr><td colspan="2">8：30-8：35</td><td colspan="2">大会介绍及与会嘉宾介绍</td><td colspan="2">李敏（中南大学 教授）</td></tr>
<tr><td colspan="2">8：35-8：45</td><td colspan="2">中南大学校领导致欢迎词</td><td colspan="2">黄健陵（中南大学 副校长）</td></tr>
<tr><td colspan="2">8：45-8：55</td><td colspan="2">CCF 生物信息学专委会主任致辞</td><td colspan="2">周水庚（复旦大学 教授）</td></tr>
<tr><td colspan="2">8：55-9：05</td><td colspan="2">大会主席致辞</td><td colspan="2">高琳（西安电子科技大学 教授）</td></tr>
<tr><td colspan="2">9：05-9：15</td><td colspan="2">CCF 专委工委委员致辞</td><td colspan="2">黄罡（北京大学 教授）</td></tr>
<tr><td colspan="6">大会报告（9：15-12：00）（地点：国际会议厅 18 楼）</td></tr>
<tr><td colspan="2">时间</td><td colspan="2">报告</td><td colspan="2">主持人</td></tr>
<tr><td colspan="2">9：15-10：05</td><td colspan="2">From health genomics to intelligent precision health
Edwin.Wang 加拿大卡尔加里大学</td><td colspan="2">高琳（西安电子科技大学 教授）</td></tr>
<tr><td colspan="6">茶歇 10：05-10：20</td></tr>
<tr><td colspan="2">10：20-11：10</td><td colspan="2">找不同：从基因芯片到单细胞测序
张学工 清华大学</td><td colspan="2">李敏（中南大学 教授）</td></tr>
<tr><td colspan="2">11：10-12：00</td><td colspan="2">基于基因本体的功能相似网络构建与纯化算法
郭茂祖 北京建筑大学</td><td colspan="2">彭绍亮（国防科技大学 教授）</td></tr>
<tr><td colspan="6">午餐（12：10-13：30）</td></tr>
<tr><td colspan="6">分会场报告（13：30-18：15）</td></tr>
<tr><td rowspan="4">分会场一</td><td colspan="5">主题：基因组、计算表观遗传及疾病研究（地点：国际会议厅 18 楼）</td></tr>
<tr><td colspan="5">第一阶段主持人： 刘娟 （武汉大学 教授）</td></tr>
<tr><td rowspan="3">特邀报告</td><td>时间</td><td>报告人姓名</td><td>工作单位</td><td>报告题目</td></tr>
<tr><td>13：30-14：00</td><td>张岩</td><td>哈尔滨医科大学</td><td>预测 CpG 岛甲基化表型介导的癌症药物反应</td></tr>
<tr><td>14：00-14：30</td><td>李伟忠</td><td>中山大学</td><td>A bioinformatics platform system for omics data analysis</td></tr>
</table>

	第二阶段主持人：蒋庆华（哈尔滨工业大学 教授）				
	主题报告	时间	论文 ID	报告题目	
		14：30-14：55	92	Extracting fitness relationships and oncogenic patterns among driver genes in cancer	
		14：55-15：20	45	Developing an agent based drug model to investigate the synergistic effect of the drug combinations	
		15：20-15：45	36	Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites	
		15：45-16：10	5	MTMO: An efficient network-centric algorithm for subtree counting and enumeration	
	茶歇 16：10-16：25				
	第三阶段主持人：杜朴风 （天津大学 副教授）				
	主题报告	时间	论文 ID	报告题目	
		16：25-16：50	82	NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition	
		16：50-17：15	2	Identifying DNA N6-methyladenosine sites by using nucleotide chemical properties	
		17：15-17：40	50	The integrative method based on Module-network for identifying driver genes in cancer subtypes	
分会场二	主题：基因表达、调控大数据与分子进化（地点：阳光会议室 19 楼）				
	第一阶段主持人：朱大铭（山东大学 教授）				
	特邀报告	时间	报告人姓名	工作单位	报告题目
		13：30-14：00	林昊	电子科技大学	The identification of prokaryotic promoters using pseudo nucleotide composition
		14：00-14：30	汪国华	哈尔滨工业大学	Identification of Regulatory Regions of Bidirectional Genes in Cervical Cancer
	第二阶段主持人：何增有（大连理工大学 教授）				
	主题报告	时间	论文 ID	报告题目	
		14：30-14：55	3	Reconstructing Cell Cycle Pseudo Time-Series via Single-cell Transcriptome Data(highlight)	
		14：55-15：20	100	miRTRS: A Recommendation Algorithm for Predicting miRNA Targets	
		15：20-15：45	81	Identification of Cancer Subtypes by Integrating Multiple Types of Transcriptomics Data with Deep Learning in Breast Cancer Data	
		15：45-16：10	72	Network-based differential analysis to identify molecular features of tumorigenesis for esophageal cancer	

	茶歇 16：10-16：25				
	第三阶段主持人：刘丙强（山东大学 副教授）				
	主题报告	时间	论文 ID	报告题目	
		16：25-16：50	60	NMFGO: Gene function prediction via nonnegative matrix factorization with Gene Ontology	
		16：50-17：15	34	基因变异间的因果关系发现与验证	
		17：15-17：40	14	A Robust Symmetric Nonnegative Matrix Factorization Framework for Clustering Multi-view Microbiome Data	
		17：40-18：15	63	Reconstructing phylogeny by aligning multiple metabolic pathways using functional module mapping	
分会场三	主题：蛋白质组学（地点：聚缘厅 3 楼）				
	第一阶段主持人：龚新奇（中国人民大学 研究员）				
	特邀报告	时间	报告人姓名	工作单位	报告题目
		13：30-14：00	陈兴	中国矿业大学	Network and machine learning algorithms in Bioinformatics research
		14：00-14：30	潘林强	华中科技大学	生物计算及其在系统生物学中的应用
	第二阶段主持人：彭玮（昆明理工大学 副教授）				
	主题报告	时间	论文 ID	报告题目	
		14：30-14：55	39	TPEA: a novel pathway enrichment analysis approach based on topological structure of pathway(highlight)	
		14：55-15：20	31	PageRank Based Method to Identify Essential Proteins by Integrating Gene Expression Profile, Gene Ontology and Protein Complexes	
		15：20-15：45	56	HIGA ： a Running History Information Guided Genetic Algorithm for Protein–Ligand Docking	
		15：45-16：10	90	The analysis on the integrality and evolutionary origin of ABA signaling pathway from aquatic to terrestrial plants	
	茶歇 16：10-16：25				
	第三阶段主持人：雷秀娟（陕西师范大学 教授）				
	主题报告	时间	论文 ID	报告题目	
		16：25-16：50	83	Protein-Protein Interactions Prediction based on Ensemble Deep Neural Networks	
		16：50-17：15	89	Effectively Detecting Protein Complexes in Weighted Dynamic PPI Networks	
		17：15-17：40	10	基于布尔矩阵分解的蛋白质功能预测框架(The Framework of Protein function prediction based on Boolean Matrix Decomposition)	

		17：40-18：15	15	A seed expansion graph clustering method for protein complexes detection in protein interaction networks

分会场四

主题：大数据分析与精准医疗（地点：培训室 18 楼）

第一阶段主持人：姜伟（南京航空航天大学 教授）

	时间	报告人姓名	工作单位	报告题目
特邀报告	13：30-14：00	邹权	天津大学	基因序列的比对、挖掘和功能分析
	14：00-14：30	谢民主	湖南师范大学	单体型组装模型及算法研究

第二阶段主持人：路永钢（兰州大学 教授）

	时间	论文 ID	报告题目
主题报告	14：30-14：55	62	A Hybrid Algorithm based on Tabu Search and Chemical Reaction Optimization for Feature Selection of Highdimensional Biomedical Data
	14：55-15：20	99	MMM: Classification of Schizophrenia Using Multi-modality Multi-atlas Feature Representation and Multi-kernel Learning
	15：20-15：45	52	Cancer Classification Based on Support Vector Machine Optimized by Particle Swarm Optimization and Artificial Bee Colony
	15：45-16：10	12	Ensemble Classification for Gene Expression Data based on Parallel Clustering

茶歇 16：10-16：25

第三阶段主持人：邓磊（中南大学 副教授）

	时间	论文 ID	报告题目
主题报告	16：25-16：50	38	Modeling and control of a delayed Hepatitis B virus model with incubation period and combination treatment
	16：50-17：15	76	基于网络约束双聚类的癌症亚型分类
	17：15-17：40	13	Protein solvent accessibility prediction by stacked deep bidirectional recurrent network
	17：40-18：15	51	A Generic Multi-Cellular Biological Simulation Platform based on CUDA

晚宴（18：30-20：00）

优秀论文颁奖、优秀墙报颁奖

分会场五

Poster（地点：国际会议厅 18 楼）

论文 ID	题目
46	A bioinformatics web platform for omics data analysis
22	Identifying Drug-pathway Association Pairs via GL2,1-Integrative Penalized Matrix Decomposition
33	Selecting Near-native Protein Structures from Ab Initio Models Using Ensemble Clustering
41	A fast projection-based algorithm for clustering big data

	61	Deletion Genotype Calling on the Basis of Convolutional Neural Network
	47	Malopred: an online prediction tool for lysine malonylation
	64	Identifying novel human miRNA-disease association based on double layer random walk model
	88	Alignment of Dynamic Protein-Protein Interaction Networks Based on Segment Tree Optimization
	30	A Novel Computational Method for Detecting DNA Methylation Sites with Sequence and Physical Structural Properties
	69	A Survey of Computational Methods of the RNA-Seq Data Analysis and Applications
	9	基于梯度投影算法的复杂网络模块划分方法
	16	GDPTrDB：connecting Genotype, Disease, Phenotype & Treatment
	77	Detecting diagnostic biomarkers of Alzheimer disease by integrating gene expression data in six brain regions

10 月 15 日会议安排（地点：湘府国际酒店）		
大会报告（8：30-12：00）（地点：国际会议厅 18 楼）		
时间	报告	主持人
8：30-9：20	生物大分子三维重构与高性能计算 刘志勇 中国科学院计算技术研究所	周水庚（复旦大学 教授）
9：20-10：10	Diagnosing Un-occurred Diseases with Single Sample by Dynamic Network Biomarkers——Detecting the tipping points of biological processes by big data 陈洛南 中国科学院上海生命科学研究院	王建新（中南大学 教授）
茶歇 10：10-10：30		
时间	报告	主持人
10：30-11：00	小 RNA 精准定量分析 唐冲 深圳华大基因股份有限公司	陈钢（Wegene 公司创始人）
11：00-11：30	GTX 基因大数据解决方案 刘齐军 人和未来生物科技（长沙）有限公司	
11：30-12：00	植物育种全基因组数据管理与挖掘 王冰冰 华智水稻生物技术有限公司	
午餐（12：10-13：30）		

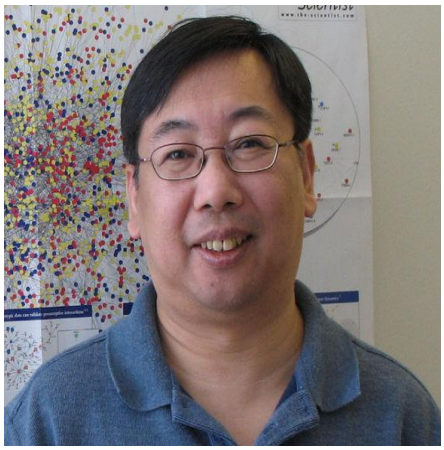
分会场一	主题：机器学习、自然语言处理在生物信息中的应用（地点：培训室 18 楼）				
	第一阶段主持人：关佶红（同济大学 教授）				
	特邀报告	时间	报告人姓名	工作单位	报告题目
		13：30-14：00	刘琦	同济大学	Learning on Pharmaceutical and Gene Editing data

		14：00-14：30	刘滨	哈尔滨工业大学	基于自然语言处理的生物序列分析研究
		14：30-15：00	宁康	华中科技大学	微生物组大数据研究：数据整合挖掘及其领域应用
	第二阶段主持人：朱峰（浙江大学 教授）				
	主题报告	时间	论文 ID	报告题目	
		15：00-15：25	101	Analysis of ribosome stalling and translation elongation dynamics by deep learning(highlight)	
		15：25-15：50	18	基于 Docker 技术架构高移植性生物信息数据软件流虚拟 web 平台	
		15：50-16：15	48	Classification and Feature Selection via Sparse Multi-view Low-Rank Regression	
		16：15-16：40	20	Multi-objective optimization algorithm to discover condition-specific modules in multiple networks	
	茶歇 16：40-16：55				
	第三阶段主持人：刘学军（南京航空航天大学 教授）				
	主题报告	时间	论文 ID	报告题目	
		16：55-17：20	32	Selecting Feature Subset Based on SVM-RFE and Overlapping Ratio	
		17：20-17：45	94	一种基于节点间路径度量的图聚类算法	
		17：45-18：10	66	An Improved Algorithm on Graph Canonization Problem	
分会场二	主题：超级计算与生物医学影像（地点：阳光会议室 19 楼）				
	第一阶段主持人：刘卫国（山东大学 教授）				
	特邀报告	时间	报告人姓名	工作单位	报告题目
		13：30-14：00	张世华	中科院数学与系统科学研究院	Matrix factorization in bioinformatics
		14：00-14：30	蔡宏民	华南理工大学	Identifying”Many-to-Many”Relationships Between Gene-Expression Data and Drug-Response Data via Sparse Binary Matching
		14：30-15：00	赵兴明	复旦大学	Decoding signaling pathways from interactomes with intelligent computational approaches
	第二阶段主持人：袁志勇（武汉大学 教授）				
	主题报告	时间	论文 ID	报告题目	
		15：00-15：25	21	An Interface for Biomedical Big Data Processing on the Tianhe-2 Supercomputer	
		15：25-15：50	37	SIMBA: a single molecule-guided Bayesian localization microscopy for practical live cell super-resolution imaging	

		15：50-16：15	79	癌症组学数据的低维表示
		16：15-16：40	91	Fusion Analysis of Resting State Networks And Its Application to Alzheimer's Disease
	茶歇 16：40-16：55			
	第三阶段主持人：彭佳杰（西北工业大学 副教授）			
	主题报告	时间	论文 ID	报告题目
		16：55-17：20	59	Deep convolutional neural networks-based early automated detection of diabetic retinopathy in fundus image
		17：20-17：45	85	A thickness-based iterative non-uniform Fourier reconstruction algorithm for electron tomogram
		17：45-18：10	96	一种面向大规模序列数据的交互特征并行挖掘算法
闭幕式（18:10-18:30） 主持人：张法（中科院计算所 副研究员）				
晚餐（18：30-19：30）				

主会场特邀报告

特邀报告一



简介：Edwin Wang，现任加拿大卡尔加里大学讲席教授、终身教授，曾任加拿大国家科学院高级研究员和麦吉尔大学教授。具有生物与计算双重教育背景，国际生物信息学知名专家。网络生物学和系统生物学，特别是癌症系统生物学一流学者。美国癌症研究学会（AACR）癌症系统生物学智囊团（Think Tank）的三十名领域内学术领袖之一。生物信息领域顶级期刊 PLoS Computational Biology 的编委。美国国家癌症研究所、美国国立卫生研究院，加拿大国家科学与工程研究委员会、加拿大农业部、加拿大国家创新基金会、加拿大国家卫生研究院基金项目评审专家。主编了癌症系统生物学领域内的第一部专著（2010）。开创了 microRNA/non-coding RNA 基因网络研究领域。有关癌症分子网络模块的开创性研究工作被写进由诺贝尔奖获得者 Hartwell 博士和系统生物学之父 Hood 博士主编的大学《遗传学》教科书（2014 和 2017 年版）. 提出癌症特征分子网络计算框架，将 20 年来传统的癌症特征描述转化为量化网路模型，从而整合癌症组学数据，图像和电子病例，用于建模和发展假说。

题目：From health genomics to intelligent precision health

摘要：Cancer is the leading cause of death and the third largest burden in the healthcare system in the world. Each year, more than 15 million new cancer patients are diagnosed and 7-8 million people die from cancer in the world. Current precision oncology is focusing on cancer treatment, however, with some notable exceptions, improvements in overall survival and morbidity over the past few decades have been modest. Historical data suggest that early detection of cancer is crucial for its ultimate control and prevention. To meet the challenges of the surge in cancer cases in the future, it is envisioned that, besides the promotion of lifestyle changes, improving early diagnosis is the best strategy for reducing the impact of carcinogenesis.

Both genetic and environmental factors (e.g., pollution, lifestyle and so on) interact to induce cancer initiation, progression and metastasis. Therefore, we are aiming to combine the genome sequencing, imaging and electronic medical records of individuals to identify high-risk cancer individuals, ‘healthy lifestyle patterns’ for cancer prevention, and monitor high-risk cancer individuals for cancer early detection. To do so, we have complied a cohort which contains 5 million people whose medical records have been collected. Among them, 0.5 million people’ genomic information has been determined. We are developing new algorithms by applying machine learning and deep learning approaches to the cohort to meet the goals mentioned above.

特邀报告二

简介：张学工，1989 年毕业于清华大学自动化系，1994 年获模式识别与智能系统博士学位。现为清华大学自动化系教授、生命学院和医学院兼职教授，清华信息国家实验室生物信息学研究部主任，清华大学学术委员会委员，国家杰出青年基金获得者、国家九七三计划首席科学家。兼任清华大学数据科学研究院医疗健康大数据研究中心副主任、清华大学合成与系统生物学研究中心执行主任，中国人工智能学会生物信息学与人工生命专业委员会主任、中国生物工程学会计算生物学与生物信息学专委会常务副主任。主要研究方向是机器学习与模式识别、生物和医学数据分析、基因表达和宏基因组分析等。

题目：找不同：从基因芯片到单细胞测序

摘要：对不同样本之间基因表达差异的分析是进行疾病的分子分型、发现生物标志物和研究疾病发生发展分子机理的重要基础，各种观测检验技术的快速发展，对差异分析不断提出新的方法学挑战。比如，近年来最新发展起来的单细胞 RNA 测序能够检测单个细胞中的基因表达，对于研究发育和癌症等过程的基因表达时空异质性具有重要意义。由于单细胞中 RNA 量极低，单细胞测序在 RNA 捕获和反转录扩增阶段与多细胞测序有很多不同的特点，导致单细胞测序数据具有独特的统计特性，很多传统 RNA 测序数据分析方法在处理单细胞数据时会遇到问题。我们系统地研究了现有方法在单细胞 RNA 测序数据差异表达分析方面的性能，从中发现了单细胞测序差异表达分析的特殊性，提出了一种区分不同类型单细胞差异表达的新方法，初步实验展示出很大的应用潜力。本报告将回顾作者在基因差异表达分析方面的一些工作，分享在单细胞 RNA 测序数据分析方法方面的最新研究进展。

特邀报告三

简介：郭茂祖，现任北京建筑大学电气与信息工程学院院长，教授、博士生导师。省杰出青年科学基金获得者（2006年），宝钢优秀教师奖获得者（2015年）。曾留学瑞典、英国、德国、日本。本、硕、博均毕业于计算机专业（1997年哈工大博士），2002年破格评为教授、2003年任博导。研究方向包括：生物信息学、机器学习、人工智能、城市计算等，曾任中国机器学习会议（CCML2015）大会主席。现为国家自然基金委重大研究计划指导专家组成员、中国计算机学会（CCF）生物信息学专业组副主任、CCF人工智能和模式识别专委会常委、中国人工智能学会机器学习专委会常委。主持完成自然基金重点项目等20多项，发表论文200余篇，已培养毕业博士16名、硕士60余名。

题目：基于基因本体的功能相似网络构建与纯化算法

摘要：基于基因本体比较基因之间的功能相似度，对基因功能分析和预测等问题具有重要意义。报告介绍基因功能相似度计算及其加速算法、人类基因功能相似网络纯化和疾病基因预测等计算问题，包括：（1）提出基于加权继承语义的基因功能相似度计算方法，能够更加准确地度量基因之间的功能相似度；（2）提出基于哈希结构的基因功能相似度计算方法，具有速度优势；（3）基因功能相似网络作为一个全连接网络往往存在“噪声”，提出基于参考网络纯化基因功能相似网络的方法，纯化后的网络符合生物分子网络的特征；（4）给出基于基因相似网络和数据融合的人类疾病基因挖掘方法。

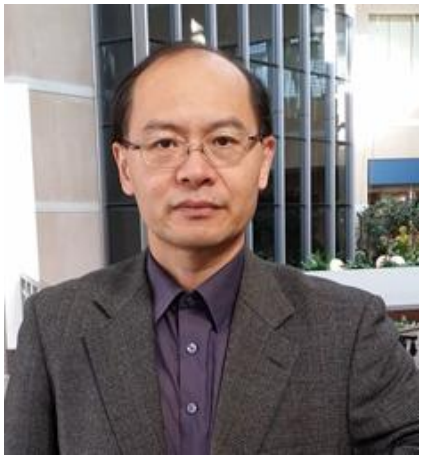
特邀报告四

简介：刘志勇，博士，研究员。曾任国家自然科学基金委员会信息科学部常务副主任，现任中国科学院计算技术研究所前瞻研究中心主任研究员；CCF 会士。参加及主持过计算机控制系统、计算机算法、体系结构、并行处理等领域的多项研究；获得全国科学大会奖；在国内外杂志（如 JACM, SJSC, IEEE Trans., JSAC, JCST, 计算机学报，等）及会议（如 ACM ICS, IEEE FOCS, INFOCOM, ICDCS, 等）上发表学术论文 200 余篇；研究成果获得学术界引用和/或国内外多家机构的实际应用；发表学术性综述、管理和科学政策等方面的论文多篇；做过多项学术服务工作，包括：中国计算机学会理事、人工智能学会理事等；国际和国内会议的主席、程序或组织委员会主席或委员、学术刊物编委及常务副主编，研究项目的专家委员会委员；现在是国家自然科学基金委员会国际合作专家咨询委员会专家，国家 973 计划信息领域咨询专家组长，等。刘志勇的研究兴趣包括计算机算法与体系结构、互联网络、并行与分布式处理系统、生物信息学等。

题目：生物大分子三维重构与高性能计算

摘要：冷冻电镜图像生物大分子三维重构已成为确定生物大分子三维结构的重要前沿技术。然而冷冻电镜所产生的数据规模、特性及生物学家对重构的三维模型的高精度要求，给当前计算科学提出了严峻的挑战。本报告将概要介绍冷冻电镜生物大分子三维重构的基本原理和主要过程，从计算科学的角度阐述冷冻电镜三维重构所面临的几个主要科学和技术问题和近期的一些研究进展，并从超级计算的角度分析大规模、大尺度、高精度冷冻电镜数据处理与三维重构面临的重要问题和研究方向。

特邀报告五



简介：Luonan Chen received BS degree in the Electrical Engineering, from Huazhong University of Science and Technology, and the M.E. and Ph.D. degrees in the electrical engineering, from Tohoku University, Sendai, Japan, in 1988 and 1991, respectively. From 1997, he was an associate professor of the Osaka Sangyo University, Osaka, Japan, and then a full Professor. Since 2010, he has been a professor and executive director at Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. He was the founding director of Institute of Systems Biology, Shanghai University, and is also research professor at the University of Tokyo since 2010. He was elected as the founding president of Computational Systems Biology Society of OR China, and Chair of Technical Committee of Systems Biology at IEEE SMC Society. He serves as editor or editorial board member for major systems biology related journals. In recent years, he published over 280 SCI journal papers and two monographs (books) in the area of systems biology.

题目：Diagnosing Un-occurred Diseases with Single Sample by Dynamic Network Biomarkers -Detecting the tipping points of biological processes by big data

摘要：Considerable evidence suggests that during the progression of complex diseases, the deteriorations are not necessarily smooth but are abrupt, and may cause a critical transition from one state to another at a tipping point. Here, we develop a model-free method to detect early-warning signals of such critical transitions (or un-occurred diseases), even with only a small number of samples. Specifically, we theoretically derive an index based on a dynamical network biomarker (DNB) that serves as a general early-warning signal indicating an imminent sudden deterioration before the critical transition occurs. Based on theoretical analyses, we show that predicting a sudden transition from small samples is achievable provided that there are a large number of measurements for each sample, e.g., high-throughput data. We employ gene expression data of three diseases to demonstrate the effectiveness of our method. The relevance of DNBs with the diseases was also validated by related experimental data (e.g., liver cancer, lung injury, influenza, type-2 diabetes) and functional analysis. DNB can also be used for the analysis of nonlinear biological processes, e.g., cell differentiation process.

分会场特邀报告

特邀报告一

简介：张岩教授，哈尔滨医科大学生物信息科学与技术学院。任多个国际期刊的学术编辑及审稿人，国家自然科学基金的通讯评审及二审专家。多年来一直从事生物信息学和计算表观遗传学领域的教学和科研工作。所带领的计算表观遗传学课题组致力于表观遗传领域的高通量数据挖掘的算法和软件开发，可视化网络平台及数据库构建。科研成果在《Nucleic Acids Research》，《Development》、《Database》、《Briefings in Bioinformatics》、《Science Reports》等有国际影响的杂志上发表了多篇论文。研究方向涉及到肿瘤生物学、发育生物学、比较基因组学等领域的基因组、表观基因组、转录组和代谢组学的数据分析。

题目：预测 CpG 岛甲基化表型介导的癌症药物反应

摘要：近年来对癌症机制的理解取得了一定的进展，但癌症的药物治疗面临着敏感性差、容易产生耐药和开发周期长、耗资大等问题，从分子层面挖掘抗癌药物敏感性因素和耐药机制非常有意义。我们以 CpG 岛甲基化表型（CpG island methylator phenotype, CIMP）作为媒介，通过整合癌症细胞系和 TCGA 组织肿瘤样本的 DNA 甲基化谱来预测药物反应。首先，对 966 个癌症细胞系进行拷贝数变异和差异甲基化分析，识别出 32 个与癌症相关的标记基因，其中包括 12 个频发拷贝数扩增基因和 20 个频发拷贝数缺失基因；然后基于标记基因集的拷贝数变异水平对细胞系进行拷贝数分型。对于 136 个拷贝数扩增和 142 个拷贝数缺失细胞系，分别基于标记基因集中的频发拷贝数扩增基因和频发拷贝数缺失基因的甲基化水平进行 CIMP 分类，拷贝数扩增细胞系中包括 65 个 CIMP-L 和 71 个 CIMP-H，拷贝数缺失细胞系包括 50 个 CIMP-L 和 92 个 CIMP-H。接着我们基于线性回归和方差分析方法在拷贝数扩增和拷贝数缺失细胞系中分别识别出 9 种和 22 种与 CIMP 显著相关的药物。并利用二次规划算法和偏最小二乘算法评估 TCGA 肿瘤样本的癌症细胞系组成，构建组织样本-细胞系-药物三者之间的网络，预测部分肿瘤样本对抗肿瘤药物的反应。本研究中得到的结果将对癌症患者的药物治疗提供指导。

特邀报告二

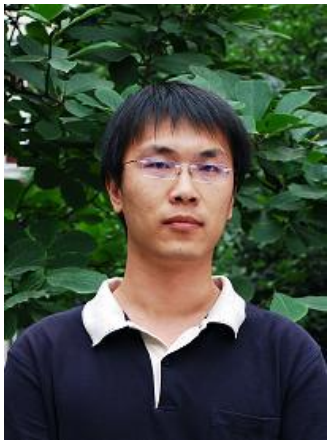
简介：赵兴明，复旦大学类脑智能科学与技术研究院，教授。目前主要从事模式识别与生物信息学交叉领域的研究，上海市青年科技启明星（2010）和上海市浦江人才计划（2013）入选者。担任 IEEE Senior Member、IEEE SMC Technical Committee on Systems Biology 委员、IAPR TC20 委员、中国运筹学会计算系统生物学分会常务理事、中国细胞学会功能基因组信息学与系统生物学分会理事、上海市生物信息学学会理事、上海市计算机学会生物信息学专委会副主任、中国人工智能学会生物信息学与人工生命专业委员会委员、中国计算机学会生物信息学专业组委员、中国计算机学会计算机应用专委会委员。同时担任 IEEE/ACM TCBB 和 Neurocomputing 等

国际期刊的客座编辑和编委。已完成 863 和国家自然科学基金重大研究计划在内的多项科研课题。在 Nucleic Acids Research、PLoS Computational Biology 和 Bioinformatics 等国际著名期刊发表 SCI 论文 70 余篇。

题目：Decoding signaling pathways from interactomes with intelligent computational approaches

摘要：Signaling pathways play key roles in biological systems, and adherent signaling pathways may lead to diseases. Unfortunately, our knowledge about signaling pathways is far from complete. In this talk, I'll present the computational approaches that unveil the topological structures of signaling pathways in an accurate way. Specifically, I'll present a hybrid intelligent approach for detecting directed signaling pathways from a large interactome. I'll showcase some applications of our computational approaches on yeast MAPK and human EGFR/ErbB signaling pathways.

特邀报告三



简介：林昊，博士，研究员，电子科技大学信息生物学中心成员。长期从事生物大分子数据信息挖掘、识别相关的生物信息学与系统生物学研究。在真核 DNA 复制与转录的计算表观遗传机制、蛋白质结构与功能的生物信息挖掘、基因调控网络的构建等方面进行了深入的研究。在 Nucleic Acids Res、Bioinformatics、Mol Ther-Nucl Acids、J Proteomics 等期刊上发表 SCI 检索学术论文 80 余篇，被 100 余种 SCI 期刊累计引用超过 3000 篇次。先后有 22 篇论文被 ESI 收录。建立生物信息在线服务网站 40 个，为来自美国、欧洲等 50 余个国家、地区的研究提供服务。担任 Scientific Reports 期刊编辑，多个 SCI 期刊特刊编辑，为 40 余种国外 SCI 期刊审稿。受到国家自然基金、四川省基金和中央高校基本业务费等项目资助。获得电子科技大学校百人计划、学术新人奖、唐立新奖教金、四川省科技进步三等奖和河北省科技进步三等奖。

题目：The identification of prokaryotic promoters using pseudo nucleotide composition

摘要：Promoters are modular DNA structures containing complex regulatory elements required for gene transcription initiation. In prokaryotes, the sigma (σ) factor of RNA holoenzyme plays key roles in recognizing and binding to the promoter sequences during gene transcription. Thus, the types of prokaryotic promoters are defined according to the types of σ factor. Two families of σ factors, namely σ70 family and σ54 family, determine promoter specificity. The σ70 promoters commonly contain three basic regulatory elements: Pribnow box (or called TATA box) with consensus TATAAT around -10 bp upstream of transcription start site (TSS), -35 box with consensus TTGACA around -35 bp upstream of TSS and initiator (Inr) around TSS. The σ54 factor can recognize the unique regulatory elements with the consensus sequence TGC[AT][TA] around -12 bp and [CT]TGGCA[CT][GA] around -24 bp upstream of the TSS. Although the biochemical experimental approaches can provide the details for prokaryotic promoters, the wet-experimental technique is time-consuming and expensive. With the avalanche of biological sequences generated in the postgenomic era, it is highly desirable to develop computational

methods to identify prokaryotic promoters in prokaryotic genomes.

Our study was devoted to enhance the prediction power and quality in identifying the prokaryotic promoters. To provide an up-to-date, interactive and extensible database for σ54 promoter, a free and easy accessed database called Pro54DB (http://lin.uestc.edu.cn/database/pro54db) was built to collect information of σ54 promoter. In the current version, it has stored 210 experimental-confirmed σ54 promoters with 297 regulated genes in 43 species manually extracted from 133 publications, which is helpful for researchers in fields of bioinformatics and molecular biology. By building objective and strict benchmark datasets, two predictors respectively called "iPro54-PseKNC" and "iPro70-PseZNC" were developed. In the predictors, the samples of DNA sequences were formulated by a novel feature vector called 'pseudo k-tuple nucleotide composition', which was further optimized by the incremental feature selection procedure. The performances of two predictors were examined by the cross-validation tests. For the convenience of the vast majority of experimental scientists, two web-servers for the iPro54-PseKNC predictor and iPro70-PseZNC predictor were established and can be freely accessible at http://lin.uestc.edu.cn/server/iPro54-PseKNC and http://lin.uestc.edu.cn/server/iPro70-PseZNC.

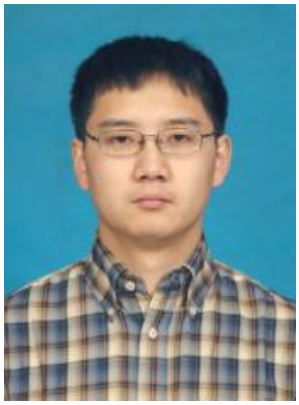
特邀报告四

简介：谢民主，男，博士，湖南师范大学物理与信息科学学院教授，博士研究生导师，电子信息科学与技术系主任，湖南师范大学物联网重点实验室主任，主要研究单体型组装和全基因关联分析算法。2003 年毕业于中南大学信息科学与工程学院，获计算机应用技术工学硕士学位。2008 年毕业于中南大学信息科学与工程学院，获计算机应用技术工学博士学位。中南大学和美国加州大学河滨分校博士后。目前从事计算机算法设计和生物信息学的研究。以第一作者或通信作者身份在 ISMB 2008、Bioinformatics、BMC Bioinformatics、Aglorithmica、计算机学报、软件学报等生物信息学和计算机算法领域的国际知名会议和杂志上作为第一作者发表论文 20 余篇，其中 SCI 收录 9 篇。主持国家自然科学基金面上项目 3 项，湖南省自然科学基金 1 项，2009 年获中国博士后科研基金一等资助，获 2013 年教育部高等学校科学研究优秀成果奖自然科学奖二等奖（排名第三）。

题目：单体型组装模型及算法研究

摘要：目前人类的基因组参考序列是来自不同群体的多个个体的基因组表决序列，而人类是二倍体生物，其基因组由两套染色体组成，获得每套染色体的 DNA 序列是理解人类遗传变异，复杂遗传疾病致病基因的有效手段。单体型组装是指在人类个体 DNA 测序片段数据的基础上，通过计算优化算法获得其两套染色体的 DNA 序列，即两套单体型。本报告将对单体型组装各种优化模型及优化算法进行概述，对其应用及发展进行展望。

特邀报告五



简介：汪国华，哈尔滨工业大学计算机科学与技术学院教授、博士生导师。2009 年获得哈尔滨工业大学计算机科学与技术博士学位。2006 年至 2008 年美国印地安那大学-普渡大学访问学者，2014 至 2016 年在美国约翰霍普金斯大学从事博士后工作。2013 年度当选教育部“新世纪优秀人才支持计划”，2011 年获得中国计算机学会“CCF 优秀博士学位论文奖提名”奖，主持多项国家自然科学基金、国家 863 项目等。目前是中国计算机学会生物信息专委会委员，人工智能学会生物信息学与人工生命专委会委员。在 Nat Rev Genet、Nat Protoc、Nucleic Acids Res、Bioinformatics 等国内外重要生物信息学期刊发表多篇 30 余篇论文。主要从事生物信息学、机器学习、人工智能研究。目前主要方向为：

（1）生物大数据分析与管理

（2）基于基因组测序大数据的基因结构挖掘算法，药物靶点识别算法

（3）基于高通量数据的 DNA 甲基化调控机制研究

题目：Identification of Regulatory Regions of Bidirectional Genes in Cervical Cancer

摘要：Bidirectional promoters are shared promoter sequences between divergent gene pair (genes proximal to each other on opposite strands), and can regulate the genes in both directions. In the human genome, >10% of protein-coding genes are arranged head-to-head on opposite strands, with transcription start sites that are separated by <1,000 base pairs. Many transcription factor binding sites occur in the bidirectional promoters that influence the expression of 2 opposite genes. Recently, RNA polymerase II (RPol II) ChIP-seq data are used to identify the promoters of coding genes and non-coding RNAs. However, a bidirectional promoter with RPol II ChIP-Seq data has not been found. In some bidirectional promoter regions, the RPol II forms a bi-peak shape, which indicates that 2 promoters are located in the bidirectional region. We have developed a computational approach to identify the regulatory regions of all divergent gene pairs using genome-wide RPol II binding patterns derived from ChIP-seq data, based upon the assumption that the distribution of RPol II binding patterns around the bidirectional promoters are accumulated by RPol II binding of 2 promoters. In HeLa S3 cells, 249 promoter pairs and 1094 single promoters were identified, of which 76 promoters cover only positive genes, 86 promoters cover only negative genes, and 932 promoters cover 2 genes. Gene expression levels and STAT1 binding sites for different promoter categories were therefore examined. The regulatory region of bidirectional promoter identification based upon RPol II binding patterns provides important temporal and spatial measurements regarding the initiation of transcription. From gene expression and transcription factor binding site analysis, the promoters in bidirectional regions may regulate the closest gene, and STAT1 is involved in primary promoter.

特邀报告六

简介：潘林强，“华中学者”特聘岗教授，湖北省运筹学学会理事长，中国电子学会图论与系统优化专业委员会副理事长，中国计算机学会高级会员，中国电子学会高级会员，中国人工智能学会高级会员。主要从事计算机科学和生物信息处理的研究和教学工作。先后主持 6 项国家自然科学基金面上项目、1 项国家自然科学基金重点项目、1 项国家自然科学基金重大国际合作项目、3 项教育部博士点基金项目。2005 年入选教育部新世纪优秀人才计划，2007 年“非传统高性能计算中的生物计算理论”获湖北省自然科学一等奖，2014 年“基于生物机理的计算模型和算法”获教育部自然科学一等奖。

题目：生物计算及其在系统生物学中的应用

摘要：细胞是生物体最基本的结构和功能单元，蕴含了大自然千万年进化所沉淀的智能，其中 DNA 分子具有存储容量大、存储密度高、自组装等特点，是信息处理值得探索的理想载体之一。该报告将介绍生物计算的研究背景、理论和实验的研究进展，及其在系统生物学中的应用，如靶向载药、测序、高分辨率荧光成像等。

特邀报告七

简介：邹权，2009 年于哈尔滨工业大学计算机学院获得博士学位。随后到厦门大学计算机系工作，任助理教授、副教授，2015 年调入天津大学计算机学院，任研究员，博士生导师。主要研究方向为生物信息学。目前，以第一作者或通讯作者发表且被 SCI 检索的论文 40 余篇。google scholar 显示引用超过 3000 次，其中代表作发表在 Briefings in Bioinformatics、Bioinformatics、PLoS Computational Biology 等知名学术期刊上。近几年，担任 SCI 期刊 Current Bioinformatics 副主编，担任 Bioinformatics 等多个杂志的审稿人。

题目：基因序列的比对、挖掘和功能分析

摘要：基因序列的比对、挖掘和功能分析是计算生物学中最基础的科学问题。本次报告围绕这三个方面介绍报告人近年来的部分成果。在序列比对方面，介绍一种基于后缀树和并行加速的多序列比对和进化树构建算法；在基因挖掘方面，介绍本人近年来在 microRNA 识别方面的部分工作，主要贡献在于探索高质量反例数据和设计全新的集成分类器；在基因功能分析方面，分别介绍基因-疾病关联分析和基因-作物产量关联分析等工作。

特邀报告八

简介：李伟忠，中山大学百人计划 2016 年国外引进高层次人才，中山大学中山医学院和中山大学精准医学科学中心教授、博士生导师，现任广东生物信息学会副理事长，国际期刊 Precision Clinic Medicine 编委，中国人工智能学会生物信息学与人工生命专委会委员；曾任欧洲生物信息研究所（EMBL-European Bioinformatics Institute，英国剑桥）生物信息学家/高级软件工程师（2009-2016），国际分子生物智能系统（ISMB）会员、国际核酸联合会（INSDC）欧洲部成员，国际蛋白质联盟（UniProt）成员。回国后主要从事面向精准医学大数据的生物信息学研究，包括生物医学数据的整合和注释、精准高效检索、大型软件工作流系统以及相关应用大平台开发。主持国家重点研发计划子课题“精准医学大数据的整合与注释”（500 万，2016-2020 年）。

论文成果：已在《核酸研究》（Nucleic Acids Research，影响因子 10.162)、《分子系统生物》（Molecular Systems Biology，影响因子 9.75)、《美国科学院院刊》（PNAS，影响因子 9.661)、《生物信息》（Bioinformatics，影响因子 7.307） 等国际权威期刊发表论文 20 余篇，其中专著两部，总影响因子近 150，总他引 SCI 超 4000、Google Scholar 近 9500。担任 Nucleic Acids Research (影响因子 10.162)、Bioinformatics（影响因子 7.303）、 Database (Oxford)、欧洲计算生物学大会（ISCB）等国际权威期刊和大会评审。

国际贡献：设计和实施 EMBL-EBI 以高性能计算为基础的核心生物大数据分析应用大平台，服务超过 180 个国家和地区，被同行视为最闪亮的生物信息平台之一；与世界四大专利局合作，建立了全球最完整的生物序列专利数据库群；设计开发的蛋白序列精确迭代检索引擎 PSI-Search，领跑国际生物序列迭代检索的研究；深度参与国际重大生物信息项目，如国际蛋白数据库 UniProt、国际核酸数据库 ENA/GenBank、基因组数据库 Ensembl Genomes、多序列比对工具 Clustal Omega、大分子功能注释工具 InterProScan，为世界范围的生物医学大数据建设和共享作出了积极贡献，同时为在中国建设大型生物信息项目积累了宝贵经验。

题目：A bioinformatics platform system for omics data analysis

摘要：We establish a bioinformatics platform system (http://lilab.sysu.edu.cn/Tools/) for omics data analysis at Sun Yat-sen University, which currently offers applications of sequence similarity search (SSS), genomic data workflows (GDW), deep learning analysis (DLA), multiple sequence alignment (MSA) and protein functional analysis (PFA). Biological sequence data including UniProt, ENA/GenBank and 1000Genomes are searchable through our PSISearch2， BLAST and FASTA applications; gene variations and protein binding sites can be functionally analysed by the deep learning applications DeepSea and DeepBind; Clustal Omega and InterProScan5 are integrated for multiple sequence alignment and protein functional analysis. Ongoing implementations include latest genomic analysis workflows for whole exome sequencing, whole genome sequencing and RNA-seq data, and the ultrafast TB-scale genomic data search tools - SBTblast & ReverseSearch. Analysis jobs can be run through webform interfaces and web service APIs. Example web service client programs in common computing languages such as Java, Perl & Python are provided for user to run high-throughput analyses systematically and consume our computing resource remotely. To facilitate software and data sharing, the platform system can be

packed by Docker container and migrated to other high-performance computing clusters.

特邀报告九

简介：刘滨，哈尔滨工业大学深圳研究生院教授、博士生导师。于 2010 年 10 月在哈尔滨工业大学深圳研究生院获得博士学位，2010 年 12 月至 2012 年 1 月在美国俄亥俄州立大学从事博士后研究工作，2012 年 1 月至今在哈尔滨工业大学深圳研究生院担任助理教授、副教授和教授。刘滨长期从事生物信息学研究工作，致力于基于序列的生物分子结构和功能识别研究。围绕该目标，系统研究了生物序列语言模型，并以此为基础提出基于自然语言处理技术的生物序列模式识别方。在 Bioinformatics、Nucleic Acids Research、Briefings in Bioinformatics 等领域权威期刊发表 SCI 论文 50 余篇，其中 2 篇论文入选“中国百篇最具影响国际学术论文”。提出的方法被多个国际研究机构作为核心特征提取算法用于设计预测模型。获得广东省自然科学杰出青年基金、“广东省特支计划”科技创新青年拔尖人才、深圳市青年科技奖、哈工大科研工作优秀个人称号、深圳市地方级领军人才和深圳市海外高层次人才（B 类）。

题目：基于自然语言处理的生物序列分析研究

摘要： 生物序列到其结构和功能的映射关系与语言中词到词义的映射关系类似。在语言学中，由字词通过语法组成有意义的句子。在生物学中，氨基酸或核酸组成具有特定结构和功能的生物大分子序列。因此，可以把这些氨基酸或核酸视为有意义的词，把他们的排列规律看成语言学中的语法，把生物结构和功能看成语义。基于两者的相似性，通过借鉴自然语言处理中基于词和语法来分析句子语义的方法，可以为解决生物序列分析领域中一些重要问题提供新的理论和技术。本次报告将介绍生物序列语言模型和基于自然语言处理技术的 DNA，RNA 和蛋白质序列分析的研究进展和挑战。

特邀报告十

简介：宁康，华中科技大学生命科学与技术学院教授，博士生导师，生物信息与系统生物学系系主任，湖北省楚天学者特聘教授。在生物信息学领域从事科研工作 10 余年，发表高水平学术论文 50 余篇，并是 10 余项发明专利和软件著作权的拥有者或主要作者。其中 2010 年回国以来，已在 Bioinformatics、PLoS Genetics、Plant Cell、Scientific Reports 等高水平学术期刊发表学术论文 50 余篇，其中以第一作者或通讯作者发表 SCI 论文 30 余篇，文章总引用超过 1500 次（Google Scholar）。获得软件著作权 6 项，申请国家发明专利 7 项。目前主持国家自然科学基金、科技部 863 课题等若干项。Scientific Reports、Genomics, Proteomics & Bioinformatics 等国际期刊编委。担任英国生物技术与生物科学研究理事会（UK-BBSRC）等基金评委。

题目：微生物组大数据研究：数据整合挖掘及其领域应用

摘要：微生物组通常包含几十到数千种不同的微生物，这些物种相互协作来适应环境的变化来完成生命的进化；同时它们的生命活动也对其所处的微环境产生了长期而深刻的影响。随着人类对于微生物了解的深入，微生物群落基础研究及其在健康和环境等领域的应用研究日益重要。

微生物组大数据的研究，是微生物组学领域中重要的研究内容。通过对微生物组学大数据的深入挖掘，将有助于深入的理解微生物群落在进化和生态上的重要规律，挖掘其应用价值。本报告介绍了微生物组学大数据研究的发展现状和发展趋势，特别是微生物组大数据的挖掘工具和应用，以及现阶段微生物组学的研究进展、应用、优势和瓶颈。尤其会详细介绍若干基于微生物组大数据的健康和环境领域应用。最后将会展望微生物组学大数据研究的巨大价值和潜力。

特邀报告十一

简介：蔡宏民，华南理工大学计算机学科学与技术学院教授，博士生导师，2014 广东省优秀青年教师， 2016 年科技部重点领域创新团队成员，全国系统生物学专业委员会委员，生物信息学与人工生命专业委员会委员，CCF 生物信息学专业组委员会委员。2012 年 9 月至今在华南理工大学任教，2016 年 9 月破格晋升博士生导师，同年破格晋升教授。研究兴趣包括医学图像分析与理解和多源生物数据信息分析。作为访问研究人员在哈佛大学 Center for Bioinformatics 实验室、宾夕法尼亚大学（UPenn）Section for Biomedical Analysis 实验室从事生物医学图像方面的研究。受邀访问香港浸会大学、日本京都大学等从事生物信息方面的合作研究。在国际顶级杂志及一流会议上发表论文 40 多篇。主持或完成国家自然科学基金三项，省部级项目十多项，累计获资助经费 500+万元。

题 目： Identifying ``Many-to-Many'' Relationships Between Gene-Expression Data and Drug-Response Data via Sparse Binary Matching

摘要：High-throughput technologies produce large amounts of data, such as levels of gene expression or drug responses. Identification of gene-drug interaction patterns among the expressing data is important to determine disease mechanisms and for drug discovery. However, accurately discriminating of the interaction patterns remains challenging due to the noisy measurements widely seen in high-throughput technologies and intrinsically ``many-to-many'' interactions within both genes and drug components. We have developed a binary matching model (NBM) to improve decoding of gene-expression and drug-response datasets through incorporation of pre-existing knowledge by network-based regularization. We present here the numerical derivation of the NBM algorithm, and show that it compares favorably with two popular methods for analysis of synthetic data. Further empirical analysis by NBM in heterogeneous data sources of drug responses and gene expression in 641 cell lines identified 16 clusters of drugs with similar gene-interaction patterns, demonstrating the effectiveness of NBM to reveal gene-drug patterns.

特邀报告十二

简介：陈兴，中国矿业大学信息与控制工程学院教授，博士生导师，中国矿业大学生物信息研究所所长，中国矿业大学首批越崎学者，中国工业与应用数学学会数学生命科学专业委员会秘书长，辽宁省生物大分子计算模拟与信息处理工程技术研究中心专家委员会副主任，江苏省生物信息学专业委员会委员。担任 37 家国际主流杂志的副主编、编委、首席特约编委和审稿人，特别是担任中科院二区杂志 BMC Systems Biology(影响因子 2.303）杂志副主编、中科院二区杂志 Scientific Reports (影响因子 4.259)编委、SCI 杂志 Current Protein & Peptide Science (影响因子 2.576）杂志编委以及中科院二区杂志 Frontiers in Microbiology（影响因子 4.076）、中科院二区杂志 Current Medicinal Chemistry（影响因子 3.249）、JCR 二区杂志 Current Topics in Medicinal Chemistry（影响因子 2.864）等五家 SCI 杂志首席特约编委。从事生物信息学和系统生物学领域的相关研究，并取得一系列重要进展。至今在中科院一区期刊 Nucleic Acids Research (影响因子 10.162) 、Bioinformatics (影响因子 7.307)、Plos Computational Biology(影响因子 4.542)、Briefings in Bioinformatics (影响因子 5.134)等国际期刊发表论文 65 篇（SCI 论文 60 篇，EI 检索 3 篇，影响因子累计约 255），其中第一作者 31 篇, 通讯作者 45 篇。以一作或者通讯发表中科院一区论文 18 篇，以一作或者通讯发表中科院二区以上论文 31 篇，以一作或者通讯发表 JCR 一区论文 33 篇。发表论文被 Nature Reviews Genetics、Nature Chemistry、Nature Reviews Endocrinology 等国际著名杂志引用共计 1349 次，5 篇论文入选 ESI 高被引论文，其中被影响因子 5 以上杂志他引 200 余次，H-因子为 20，单篇最高引用次数为 300，参编专著 4 部，曾获教育部高等学校科学研究优秀成果奖自然科学奖二等奖、中国矿业大学越崎学者、国际网络博弈论大会最佳论文奖、第七届图论与组合算法国际研讨会"青年论文奖" 二等奖、第四届世界华人数学家大会新世界数学奖等荣誉，主持或以骨干身份参与国家自然科学基金重大研究计划培育项目、重点基金、面上项目、青年基金、中国矿业大学越崎学者人才引进项目、中国矿业大学学科前沿科学研究专项面上项目等 14 项重要项目。担任七家国际生物信息学会议的程序委员会成员，担任北京青少年科技俱乐部活动委员会学术指导导师和《2012－2013 运筹学学科发展报告》编写组成员。

题目：Network and machine learning algorithms in Bioinformatics research

摘要：In this talk, the following effective computational models developed by Chen Group for the Bioinformatics research will be introduced: 1) PBMDA (PLOS Computational Biology, 2017, 13(3): e1005455, cited 11 times): Path-Based MiRNA-Disease Association (PBMDA) prediction model was proposed by constructing a heterogeneous graph consisting of three interlinked sub-graphs and further adopting depth-first search algorithm to infer potential miRNA-disease associations. 2) LRLSLDA (Bioinformatics, 2013,29(20):2617-2624, cited 62 times): We proposed the assumption that similar diseases tend to be associated with functionally similar lncRNAs and further developed the method of Laplacian Regularized Least Squares for LncRNA-Disease Association (LRLSLDA) in the semi-supervised learning framework. 3) KATZHMDA (Bioinformatics, 2017, 33(5):733-739): We constructed a microbe-human disease association network and further developed a novel computational model of KATZ measure for

Human Microbe–Disease Association prediction (KATZHMDA) based on the assumption that functionally similar microbes tend to have similar interaction and non-interaction patterns with noninfectious diseases, and vice versa. To our knowledge, KATZHMDA is the first tool for microbe–disease association prediction. 4) NRWRH (Molecular BioSystems,2012,8(7):1970-1978, cited 152 times): The method of Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) is developed to predict potential drug–target interactions on a large scale under the hypothesis that similar drugs often target similar target proteins and the framework of Random Walk. NRWRH makes full use of the tool of the network for data integration to predict drug–target associations. 5) NLLSS (PLOS Computational Biology, 2016,12(7): e1004975, cited 19 times): We proposed similar nature of drug combinations: principal drugs which obtain synergistic effect with similar adjuvant drugs are often similar and vice versa, and further developed a novel algorithm termed Network-based Laplacian regularized Least Square Synergistic drug combination prediction (NLLSS) to predict potential synergistic drug combinations by integrating different kinds of information such as known synergistic drug combinations, drug-target interactions, and drug chemical structures.

特邀报告十三

简介：张世华，现任中国科学院数学与系统科学研究院副研究员、中国科学院随机复杂结构与数据科学重点实验室副主任、中国科学院大学岗位教授。主要从事运筹学、模式识别与生物信息学交叉研究。目前担任 BMC Genomics，Frontiers in Genetics，Scientific Reports，Current Bioinformatics 等杂志的编委以及 IEEE/ACM TCBB 的客座编委。曾荣获中国青年科技奖、国家自然科学基金“优秀青年”基金、中组部万人计划“青年拔尖人才”计划。

题目：Matrix factorization in bioinformatics

摘要：Nonnegative matrix factorization (NMF) is a powerful technique for dimension reduction and pattern recognition. In this talk, I will survey the recent applications of NMF and its variants joint NMF in bioinformatics. Next, I will introduce some new algorithmic exploration about joint NMF and its application in RNA-protein binding prediction. Lastly, I will present a flexible NMF framework CSMF to combine data dimension reduction and differential analysis into one paradigm to simultaneously reveal common and specific patterns from data generated under interrelated biological scenarios. We demonstrate the effectiveness of CSMF with four biological applications. Extensive analysis yields novel insights into hidden combinatorial patterns embedded in these interrelated multi-modal data. Results demonstrate that CSMF is a powerful tool to uncover common and specific patterns with significant biological implications from data of interrelated biological scenarios

特邀报告十四

简介：刘琦，同济大学生物信息学系教授，博士生导师，同济大学青年“百人计划”。IEEE 会员，ACM 会员，中国计算机协会 CCF 会员（生物信息学专业组委员），中国人工智能学会会员 (生物信息学与人工生命专委会委员)，上海市启明星人才、浦江人才。致力于计算机和生命科学的智能计算和机器学习的交叉研究，目前关注于基于人工智能和机器学习方法进行药物信息学，肿瘤免疫治疗以及基因编辑的小 RNA 设计等方向的研究工作。在 Nature 旗下刊物 Clinical Pharmacology & Therapeutics, Nature Communications, CELL 旗下刊物 Trends in Biotechnology, Molecular Therapy - Nucleic Acids 以及主流的生物信息学期刊如 Bioinformatics, Briefings in Bioinformatics, NAR，及机器学习领域国际顶级期刊及会议如 TKDE, SDM, ICDM 等发表系列论文，开发了相应的生物数据挖掘平台 20 余项，和 AZ, Roche 等国际制药公司开展广泛合作。目前主持及参与了科技部重大研发计划，863 计划，国家自然科学基金面上，青年项目以及上海市级重点基金项目等。担任国家重点研发计划精准医学项目以及生物安全-遗传资源库建设项目评审专家等。

题目：Learning on Pharmaceutical and Gene Editing data

摘要：In this talk, we will briefly introduce sevearl cases recently developed in our research group to use artificial intelligence and machine learning models for the analysis of pharmaceutical data and genome editing data, hopefully to provide useful clues for future precision medicine study.

会议论文报告摘要

Extracting fitness relationships and oncogenic patterns among driver genes in cancer(92)

Xindong Zhang, Lin Gao

摘要: Driver mutation provides fitness advantage to cancer cells, while the accumulation of driver mutations increases the fitness of cancer cells and accelerates cancer progression. The work seeks to extract patterns that driver genes accumulated ("fitness relationships") in tumorigenesis. We introduce a network-based method to extract fitness relationships among driver genes by modeling the network properties of the "fitness" of cancer cells, and colon adenocarcinoma (COAD) and skin cutaneous malignant melanoma (SKCM) are studied as cases. Consistent results derived from different biological networks suggest the reliability of identified fitness relationships. Also co-occurrence analysis and pathway analysis indicate significant consistence of fitness relationships with signaling transduction. In addition, a subset of driver genes of high indegree (called "fitness core") is recognized for each case. Further analyses indicate functional importance in carcinogenesis and potential therapeutic opportunities in medicinal intervention of fitness core. Fitness relationships from non-core genes to core genes reveal potential oncogenetic patterns among driver genes in the carcinogenesis.Fitness relationships characterize functional continuity among driver genes in carcinogenesis, and suggest new insights in understanding oncogenic mechanisms of cancers, as well as guiding information for medicinal intervention.

Developing an agent based drug model to investigate the synergistic effect of the drug combinations（45）

Hongjie Gao, Zuojing Yin, Zhiwei Cao and Le Zhang

摘要:

Motivation: The growth and survival of cancer cells are greatly affected by their surrounding microenvironment. To understand the regulation under the influence of anti-cancer drugs and the synergistic effects. We develop a multiscale agent based model to investigate the synergistic effect of the drug combinations.

Result: The model can not only describe multicellular system as well as the interactions between the microenvironment and cells in detail, but also can predict the synergistic effect of the drug combinations after its key parameters and predictive power are trained and validated by the experimental data.

Availability: This research develops such a multiscale novel algorithm that can use limited experimental data to build up a predictive model for the synergistic effect of the drug combination prediction.

Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites(36)

Wei Leyi, Ran Su, Pengwei Xing, Bing Wang, Xiuting Li and Quan Zou

摘要: N6-methyladenosine (m6A), as one of the most well-studied RNA modifications, has been found to be involved with a wide range of biological processes. Recently, diverse computational methods have been developed for automated identification of m6A sites within RNAs. To accurately identify m6A sites, one of the major challenges is to extract informative features to describe characteristics of m6A sites. However, existing feature representation methods are usually hand-crafted based, and cannot capture discriminatory information of m6A sites. In this paper, we develop a m6A site predictor, named DeepM6APred. In this predictor, we propose to use a deep learning based feature descriptor with deep belief network (DBN) to extract high-level latent features. By integrating the deep features with traditional handcrafted features, we train a classification model based on support vector machine to improve the prediction ability of m6A sites. Experimental results on a benchmark dataset show that our proposed method outperforms the state-of-the-art predictors, at least 2% higher in terms of Matthew's correlation coefficient (MCC). Moreover, a webserver that implements the DeepM6APred is established, which is currently available at the website: http://server.malab.cn/DeepM6APred. It is expected to be a useful tool to assist biologists to reveal the functional mechanisms of m6A sites.

MTMO: An efficient network-centric algorithm for subtree counting and enumeration(5)

Guanghui Li, Jiawei Luo, Zheng Xiao and Cheng Liang

摘要： The frequency of small subtrees in biological, social, and other types of networks could shed light into the structure, function, and evolution of such networks. However, counting all possible subtrees of a prescribed size can be computationally expensive because of their potentially large number even in small, sparse networks. Moreover, most of the existing algorithms for subtree counting belong to the subtree-centric approaches, which search for a specific single subtree type at a time, potentially taking more time by searching again on the same network. In this paper, we propose a network-centric algorithm (MTMO) to efficiently count k-size subtrees. Our algorithm is based on the enumeration of all connected sets of k-1 edges, incorporates a labeled rooted tree data structure in the enumeration process to reduce the number of isomorphism tests required and uses an array-based indexing scheme to simplify the subtree counting method. The experiments on three representative undirected complex networks show that our algorithm is roughly an order of magnitude faster than existing subtree-centric approaches and base network-centric algorithm which does not use rooted tree, allowing for counting larger

subtrees in larger networks than previously possible. Moreover, we show major differences between unicellular and multicellular organisms.

NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition(82)

Juanjuan Kang, Yewei Fang, Pengcheng Yao, Qiang Tang and Jian Huang

摘要: Neuropeptides (NPs) are short secreted peptides produced mainly in the nervous system and digestive system. It activates signaling cascades to control a wide range of biological functions, such as metabolism, sensation and behavior. NPs are typically produced from a larger NP precursor (NPP) which includes a signal peptide sequence, one or more NP sequences and other sequences. With the drastic growth of unknown protein sequences generated in the post-genomic age, it is highly desired to develop computational methods for rapidly and effectively identifying NPP. In this article, we developed a predictor for NPPs based on optimized sequence composition of single amino acid, dipeptide and tripeptide. Evaluated with independent datasets, the predictor showed good performance that achieved an accuracy of 88.65% with AUC of 0.95. The corresponding web server was developed, which is freely available at http://i.uestc.edu.cn/neuropeptide/neuropp/home.html. It can help relevant researchers to screen candidate NP precursor and shorten experimental cycle and cost.

Identifying DNA N6-methyladenosine sites by using nucleotide chemical properties(2)

Pengmian Feng, Hui Ding, Wei Chen and Hao Lin

摘要: DNA N6-methyladenine (6mA) is associated with a wide range of biological processes. Accurate identification of 6mA site will be helpful for revealing its biological functions. Although experimental methods are available for identifying 6mA site, they are still labor-intensive and expensive. Therefore, in the present study, a support vector machine based-method was proposed to identify 6mA site in Mus musculus. It was found that the proposed method obtained an accuracy of 96.73% in the jackknife test. The predictive performances of the proposed method for identifying 6mA site in other genomes are also promising. These results indicate that the proposed method will become a useful tool for identifying 6mA sites.

The integrative method based on Module-network for identifying driver genes in cancer subtypes(50)

Xinguo Lu, Xing Li, Xin Qian and Qiumai Miao

摘要: With advances in next-generation sequencing(NGS) technologies, large number of multiple types of high-throughput genomics data are available. A highly challenge for cancer genomics is to identify the driver genes from the passenger genes by analyzing and integrating big and noisy genomics data. Breast cancer is a heterogeneous disease including five subtypes. The identification of subtype-specific biomarkers is critical to guide the diagnosis, assessment of prognosis and treatment of breast cancer. We have developed an integrated analysis frame based on gene expression profiling data and copy number aberrations data to identify breast cancer subtype-specific biomarkers. In this frame, we employed statistical machine-learning method to select genesets and utilizing an integrative method based on module network analysis to identify potential driver genes. Then,we obtained the final list of candidate biomarkers through the comparision of the results of subtypes. Lastly, to validate the final result of gene module analysis, we used classification methods to classify samples of selected driver genes and analyzed the biological significance of driver genes.The study result indicated the driven computational method can identify the potentially and biologically relevant genes of breast cancer subtypes. Moreover, this method can be used to comparing two condition with highly heterogeneous datasets in a wide range of diseases

Reconstructing Cell Cycle Pseudo Time-Series via Single-cell Transcriptome Data(highlights)(3)

Ting Chen

摘要：Single-cell RNA sequencing (scRNA-seq), which permits transcriptional profiling of individual cells, has been applied to heterogeneous groups of cells to study growth and development of tissues and tumors. Resolving cell cycle transcriptional dynamics and assessing cell cycle status for such groups of cells are essential to gain a thorough understanding of the above processes, but may not be adequately achieved by commonly used approaches. Here we developed a computational method to recover cell cycle along time (reCAT) for unsynchronized single-cell transcriptome data. First, reCAT recovers a high-resolution pseudo cell cycle time-series by solving a traveling salesman model. Next, reCAT applies a hidden Markov model (HMM) to segment the time-series into cell cycle stages. reCAT was independently tested for accuracy and reliability using several datasets. New candidate cell cycle regulated genes were identified via the recovered time-series. We found that cell cycle genes cluster into two major waves of expression, which correspond to the two well-known checkpoints, G1 and G2. We also quantitatively estimated distributions of cell cycle stages among distinct tissue and tumor cell groups. Moreover, we leveraged reCAT to exhibit methylation variation along the recovered cell cycle. Thus, reCAT demonstrates the potential to elucidate diverse aspects of cell cycle profiles during general cellular processes with single-cell resolution.

miRTRS: A Recommendation Algorithm for Predicting miRNA Targets(100)

Hui Jiang, Jianxin Wang, Wei Lan, Fang-Xiang Wu and Yi Pan

摘要:microRNAs (miRNAs) are small and important non-coding RNAs that regulate gene expression in transcriptional and post-transcriptional level. Predicting miRNA targets is an important problem in biological research. It is expensive and time-consuming to identify miRNA targets by using biological experiments. Many computational methods have been proposed to predict miRNA targets.However, these methods suffer from the high false positive rate because of the complex relationships between miRNAs and their targets. In this study, we develop a novel method, named miRTRS, for predicting miRNA targets based on a recommendation algorithm. miRTRS can predict targets for a new miRNA with miRNA sequence similarity, as well as new targets for a miRNA with gene function similarity. Furthermore, comparing to supervised machine learning methods, miRTRS does not need to select negative samples. We use ten-fold cross validation to evaluate the performance of our method. The experimental results have shown that our method miRTRS outperforms other miRNA target prediction methods in terms of AUC and other evaluation metrics.

Identification of Cancer Subtypes by Integrating Multiple Types of Transcriptomics Data with Deep Learning in Breast Cancer Data (81)

Yang Guo and Xuequn Shang

摘要: Many cancers are composed of multiple subtypes in terms of distinct pathogeneses and clinical therapeutics. Identification of cancer subtypes is vital to advance the precision of disease diagnosis and therapy. In recent years, the advance of the high-throughput sequencing techniques produced huge and multiple types of genomics data, and provided good opportunities to comprehensively understand the mechanism of cancer progress. Over the past few years, numerous approaches have been proposed to integrate multiple types of genomics data, such as gene expression, DNA methylation and miRNA, etc., to investigate cancer subtypes. However, on the one hand, few of them particularly considered the intrinsic correlations among genetic elements in each type of genomics data; on the other hand, to the best of our knowledge, none of them considered the alternative splicing regulations in data integrations. Nevertheless, it has been demonstrated that many cancers occurrences are related to abnormal alternative splicing regulations in recent years. In this paper, we proposed a hierarchical deep learning framework, HI-SSAE, to integrate gene expression and transcriptome-wide alternative splicing profiles data to identify cancer subtypes. We first adopted stacked sparse autoencoder (SSAE) neural network to learn high-level representations of each type of data set, respectively. Then, a sparse autoencoder learning layer was designed to integrate all of high-level representations in each data set to learn more complex top representation features. Finally, based on learned data presentations, we simply used k-means algorithm to cluster patient samples into different cancer subtypes. Comprehensive experiments based on TCGA breast cancer data set demonstrated our integration approach outperformed traditional cancer subtype clustering methods, and it provided a useful tool to integrate multiple types of genomics data to identify cancer subtypes.

Network-based differential analysis to identify molecular features of tumorigenesis for esophageal cancer(72)

Suxia Jiang, Qi Zhang, Yansen Su and Linqiang Pan

摘要：Esophageal cancer has a poor prognosis and high mortality rate in the world. The diagnosis and treatment of esophageal cancer are hindered by the limited knowledge about the pathogenesis mechanisms of esophageal cancer. In this work, we proposed a method to select candidate biomarkers of esophageal cancer based on the topological difference analysis between the gene-gene interaction networks for esophageal cancer and normal. We established the gene-gene

interaction networks for esophageal cancer and normal based on the correlation of genes. For each gene, we calculated and compared five centrality measures, which could reflect the topological property of a network. According to five centrality measures, the genes with large differences between the two networks were regarded as candidate biomarkers for esophageal cancer. Total of 21 candidate biomarkers were identified for esophageal cancer, and seven of them have been confirmed to be biomarkers of esophageal cancer by previous researches. In addition, six genes (RBPMS2, PDK4, IGK, SBSN, IFIT3, and HSPB6) were likely to be the biomarkers of tumorigenesis for esophageal cancer due to the fact that the biological processes they participated in are closely related with esophageal cancer. As the limitations of our existing knowledge, the potential relationships between the rest eight genes and esophageal cancer are still not clear. Statistical analysis indicates that effectiveness of all the detected biomarkers of esophageal cancer. The current method could be extended to other complex diseases for detecting the molecular features of pathopoiesis and targets for targeted therapy.

NMFGO: Gene function prediction via nonnegative matrix factorization with Gene Ontology(60)

Guoxian Yu, Keyao Wang, Guangyuan Fu and Jun Wang

摘要: Gene Ontology (GO) is a controlled vocabulary of terms that describe molecule function, biological roles and cellular locations of gene products (i.e., proteins and RNAs), it hierarchically organizes more than 43,000 GO terms via the direct acyclic graph. A gene is generally annotated with several of these GO terms. Therefore, accurately predicting the association between genes and massive terms is a difficult challenge. To combat with this challenge, we propose an matrix factorization based approach called NMFGO. NMFGO stores the available GO annotations of genes in a gene-term association matrix and adopts an ontological structure based taxonomic similarity measure to capture the GO hierarchy. Next, it decomposes the association matrix into two low-rank matrices via nonnegative matrix factorization regularized with the GO hierarchy. After that, it employs a semantic similarity based k nearest neighbor classifier in the low-rank matrices approximated subspace to predict gene functions. Empirical study on three model species (S. cerevisiae, H. sapiens and A. thaliana) shows that NMFGO achieves significantly better prediction performance than other related comparing methods and is robust to the input parameters.

基因变异间的因果关系发现与验证(34)

Cai Ruichu, Zhen Qiqi and Hao Zhifeng

摘 要:基因的变异间的相关性是全基因组关联分析等领域中的难点。当前基因变异间关系的研究主要基于基因在染色体上的相对位置展开，本文从另一个角度研究发现位于同一信号通路中的基因变异之间具有较强的因果性。具体来说，首先，我们基于单基因核苷酸多态数据

（Single Nucleotide Polymorphisms data, SNPs 数据）对基因状态进行编码，得到离散型的基因变异状态数据；然后，基于大量的基因变异数据构建基因间的因果贝叶斯网模型；最后，将基因变异数据上发现的因果贝叶斯网络与在真实信号通路上进行了验证。在WTCCC(Wellcome Trust Case Control Consortium)数据集上的实验结果表明，相互调控的基因之间的变异具有较强的因果关系。同时，实验也发现了一批具有较强因果关系的基因变异，可能对相关研究具有一定的启发意义。

A Robust Symmetric Nonnegative Matrix Factorization Framework for Clustering Multi-view Microbiome Data(14)

Yuanyuan Ma, Xiaohua Hu, Tingting He, Xingpeng Jiang and Dajun Xiao

摘要: Integration of multi-view datasets which are comprised of heterogeneous sources or different representations is challenging to understand the subtle and complex relationship among data samples. Data integration methods are needed to combine efficiently the complementary information of multiple data types to construct a comprehensive view of underlying data. Here, we propose a fast and robust framework (RSNMF) based on symmetric nonnegative matrix factorization (SNMF) and similarity network fusion (SNF) for clustering human microbiome data including functional, metabolic and phylogenetic profiles from Human Microbiome Plan. In contrast to many existing methods that typically utilize all the information provided by each view to create a consensus representation, the robust symmetric nonnegative matrix factorization (RSNMF) combines the strength of SNMF and the advantage of SNF to form a robust clustering indicator matrix which avoids suffering from much noise. We conduct experiments on one synthetic and two real datasets and the results show that the proposed RSNMF has better performance over the baseline and the state-of-art methods, which demonstrates the potential application of RSNMF for microbiome data analysis.

Reconstructing phylogeny by aligning multiple metabolic pathways using functional module mapping(63)

Yiran Huang, Cheng Zhong, Hai Xiang Lin, Jianyi Wang and Yuzhong Peng

摘要: Comparing metabolic pathways provides a systematic way of understanding evolutionary and phylogenetic relationships in systems biology. Although a number of phylogenetic methods have been developed, few efforts have been made to provide a unified phylogenetic framework which sufficiently reflects the metabolic features of organisms. Here we propose a phylogenetic framework which can characterize the metabolic features of organisms by aligning multiple metabolic pathways using functional module mapping. First, our method transforms the alignment of multiple metabolic pathways into constructing the union graph of pathways, and then the proposed method builds mappings between functional modules of pathways in union graph, and finally it infers phylogenetic relationships among organisms based on the module mappings.

Experimental results show that the use of functional module mapping enables us to correctly categorize organisms into main categories with specific metabolic characteristics. Traditional genome-based phylogenetic methods can reconstruct phylogenetic relationships, whereas our method can afford in-depth metabolic analysis for phylogenetic reconstruction, which is not sufficiently reflected by genome-based phylogenetic reconstruction. For example, our analysis reveals that the metabolic structure of archaea is different from other species. The results also demonstrate that our method can reconstruct better phylogenies in comparison to existing classification methods using metabolic pathway data. With the increasing mass of metabolic annotations, our method would be a useful complement to the traditional phylogenetic methods and can assist the metabolic analysis of phyletic reconstruction.

TPEA: a novel pathway enrichment analysis approach based on topological structure of pathway(highlight)（39）

Qian Yang and Wei Jiang

摘要: Pathway enrichment analysis has been widely used to identify cancer risk pathways, and contributes to elucidating the mechanism of tumorigenesis. However, most of the existing approaches use the outdated pathway information and neglect the complex gene interactions in pathway. Here, we firstly reviewed the existing widely used pathway enrichment analysis approaches briefly, and then we proposed a novel Topology-based Pathway Enrichment Analysis (TPEA) method, which integrated topological properties and global upstream/downstream positions of genes in pathways. We compared TPEA with four widely used pathway enrichment analysis tools, including DAVID, GSEA, CePa and SPIA, through analyzing six gene expression profiles of three tumor types (colorectal cancer, thyroid cancer and endometrial cancer). As a result, we identified several well-known cancer risk pathways that could not be obtained by the existing tools, and the results of TPEA were more stable than that of the other tools in analyzing different datasets of the same cancer. Ultimately, we developed an R package to implement TPEA, which could online update KEGG pathway information and was available at the Comprehensive R Archive Network (CRAN): https://cran.r-project.org/web/packages/TPEA/.

PageRank Based Method to Identify Essential Proteins by Integrating Gene Expression Profile, Gene Ontology and Protein Complexes（31）

Xiujuan Lei, Xiaoqin Yang and Fangxiang Wu

摘要: Essential proteins are regarded as the crucial components of organisms, and thus identifying essential proteins is a hot and significant topic in biomedical research. A great deal of

computational methods based on network topology have been proposed to characterize protein essentiality. However, the prediction accuracy still needs to be improved because of the false interactions of PPI data. In this paper, a novel approach is proposed to identify essential proteins from the PPI network by applying the Page Rank model as well as Integrating gene Expression profiles, gene Ontology and protein Complexes information (named PR_EOC). Distinguished from other approaches, to detect essential proteins, PR_EOC filters the PPI network by deleting these unreliable interactions firstly, and besides network topology and biological information data, we calculate the degrees of proteins in complexes and estimate the protein's importance in the whole PPI network by calculating Betweeness Centrality (BC). Moreover, it takes the neighbors' characteristics into account by adopting PageRank algorithm.The computational experiments show that our approach PR_EOC is superior to the other eight state-of-the-art methods (DC, SC, IC, LAC, NC, SoECC, WDC, PeC) for predicting essential proteins in PPI network.

HIGA : a Running History Information Guided Genetic Algorithm for Protein–Ligand Docking(56)

Boxin Guan, Changsheng Zhang and Yuhai Zhao

摘要: Though Lamarckian genetic algorithm has demonstrated excellent performance in terms of protein-ligand docking problems, it can not memorize the evaluated solutions that it has accessed, rendering it effort-consuming to discover some promising solutions. This paper illustrates a novel and robust optimization algorithm (HIGA) based on Lamarckian genetic algorithm (LGA) for solving the flexible protein-ligand docking problems with an aim to overcome the above-mentioned drawback. A running history information guided model is applied in the method, which makes it more efficient to find the lowest energy of protein-ligand docking. We evaluate the performance in the aspects of lowest energy and highest accuracy of HIGA in comparison with GA, LGA, SODOCK, and ABC, the results of which indicate that HIGA outperforms other search algorithms.

The analysis on the integrality and evolutionary origin of ABA signaling pathway from aquatic to terrestrial plants(90)

Zhiyong Pei, Manjiao Liu, Senbiao Fang and Yubao Chen

摘要： As one of the most important kinds of hormones, abscisic acid (ABA) regulates crucial physiologically developmental processes and water stress responses. The insight of ABA signaling pathway would be of great significance to understand the physiological mechanism of plant response in drought situation and helpful in agricultural application. The components factors involved in this complicated signaling pathway have been identified by genetic analyses. While the evolutionary research and the origins of the complete pathway from aquatic living to land

remained unclear. In this work, the integrality of ABA signal conduction pathway in ten plants species，including aquatic algae and terrestrial plants， were whole genome wide searched. The phylogenic analysis of the ABA signaling gene sequences have been performed and the molecular evolution analysis were also carried on to examine the position selection pressure the gene undertaken during evolutionary periods . The whole genome blast search result shows that the numbers of the key protein family involved in the ABA signaling pathway correlated with the speciation evolution from aquatic to terrestrial plants. The earlier complete ABA signaling pathway found in Physcomitrella patens which is one of the lower plants live on land. The family expansion of ABA pathway related genes and the positive selection of the key protein might have contributed to the terrestrial living environment adaptation for higher plants on land.

Protein-Protein Interactions Prediction based on Ensemble Deep Neural Networks(83)

Long Zhang, Guoxian Yu, Dawen Xia and Jun Wang

摘要: Protein-protein interactions (PPIs) are of vital importance to most biological processes. Plenty of PPIs have been identified by wet-lab experiments in the past decades, but there are still abundant uncovered PPIs. Furthermore, wet-lab experiments are expensive and limited by the adopted experimental protocols. Although various computational models have been proposed to automatically predict PPIs, and provided reliable interactions for experimental verification, the problem is still far from being solved. Novel and competent models are still anticipated. In this study, a neural network based approach called EnsDNN (Ensemble Deep Neural Networks) is proposed to predict PPIs based on different representations of amino acids. Particularly, EnsDNN separately uses auto covariance descriptors, local descriptors and multi-scale continuous and discontinuous local descriptors, to describe and explore the pattern of interactions between sequentially distant and spatially close amino acid residues, and then trains deep neural networks (DNNs) with different configurations based on each descriptor. Next, EnsDNN integrates these DNNs into an ensemble predictor to leverage complimentary information of these descriptors and of DNNs, and then to predict potential PPIs. EnsDNN achieves superior performance with accuracy as 94.07 %, sensitivity as 94.07 %, and precision as 94.07 % on predicting PPIs of Saccharomyces cerevisiae. Results on other five independent PPI datasets also demonstrate that EnsDNN gets better prediction performance than other related comparing methods.

Effectively Detecting Protein Complexes in Weighted Dynamic PPI Networks (89)

Yunjia Shi, Heng Yao, Shuigeng Zhou and Jihong Guan

摘要: The identification of protein complexes is significant to understand the mechanisms of cellular processes. Up to now, many methods have been developed to identify protein complexes in static PPI networks. However, static PPI networks cannot accurately describe the behaviors of proteins in the different stages of life cycle of a cell. In this paper, we integrate gene expression data and GO terms into high-throughput PPI data to construct weighted dynamic PPI networks, on which we propose a new method to detect protein complexes. Specifically, we first calculate protein active probability and protein functional similarity to construct weighted dynamic PPI networks, then define a high-order topological overlap measure of similarity to extract protein complexes based on the core-attachment model. In our experiments, four PPI datasets are used to detect protein complexes. Experimental results indicate that our method is superior to the existing methods in overall.

基于布尔矩阵分解的蛋白质功能预测框架(The Framework of Protein function prediction based on Boolean Matrix Decomposition)(10)

Liu Lin, Tang Mingjing, Tang Lin and Zhou Wei

摘要: 蛋白质功能预测问题本质上是一个多标签分类问题，但庞大的功能标签数量使得各种多标签分类器在蛋白质功能预测中的应用面临巨大挑战。本文针对蛋白质功能标签数量庞大且标签关联性较高的特点，提出了一种基于布尔矩阵分解的蛋白质功能预测框架(PFP-BMD)。同时，针对目前布尔矩阵分解算法中精确分解和列利用条件难以同时满足的问题，提出一种基于标签簇的精确布尔矩阵分解算法，该算法可通过标签关联矩阵实现标签的层次扩展聚簇，并通过相关推论证明了该算法可实现最优的精确布尔矩阵分解。实验结果表明，本文提出的布尔矩阵分解算法在计算复杂度上具有较大优势，且应用了该算法的蛋白质功能预测框架可有效提升蛋白质功能预测的准确率，本文的研究为各种多标签分类器在蛋白质功能预测中的高效应用奠定了基础。

A seed expansion graph clustering method for protein complexes detection in protein interaction networks(15)

Jie Wang, Wenping Zheng and Jiye Liang

摘要: Most proteins perform their biological functions while interacting as complexes. The detection of protein complexes is an important task not only for understanding the relationship between functions and structures of biological network, but also for predicting the function of unknown proteins. In the paper, we present a new measure of a node to reflect its representability to a cluster in a protein interaction (PPI) network by integrating the topological information of its neighborhood, which reflects representability of the node in a larger local

neighborhood using an iterative procedure. Based on the measure, we propose a seed-expansion graph clustering algorithm (SEGC) for protein complexes detection. A roulette wheel strategy is used in the selection of the seed to enhance the diversity of clustering. We also define closeness NC(u,C) between a candidate node u and a cluster C, where both the density of C and the connection density between u and C. During the expansion process of SEGC, the candidate node with highest closeness is added to the cluster in consider. In SEGC, one node might be assigned to multiple clusters. We compare the F-measure and accuracy of the proposed SEGC algorithm with some representative algorithms as CFinder, DPClus, IPCA, Core and SR-MCL on Saccharomyces Cerevisiae protein interaction networks. The experimental results show that our SEGC outperforms other algorithms under full coverage.

A Hybrid Algorithm based on Tabu Search and Chemical Reaction Optimization for Feature Selection of Highdimensional Biomedical Data（62）

Chaokun Yan, Jingjing Ma, Junwei Luo and Huimin Luo

摘要: In the last few years, there has been a rapid development in various bioinformatics technologies, which has led to the accumulation of a large amount of biomedical data. The biomedical data can be analyzed to enhance assessment of at-risk patients and improve the diagnosis, treatment and prevention of disease. However, these datasets usually have a large number of features which contains much irrelevant or redundant information. Feature selection is a solution that involves finding the optimal subset, which is known to be an NP problem due to the large search space. For the issue, the paper present a new feature selection approach based on improved chemical reaction optimization algorithm (CRO) and KNN. Tabu search is integrated with CRO framework to enhance it's the capacity of local search. KNN is adopted to evaluate the quality of selected candidate subset. The experimental results on nine standard medical datasets show our approach significantly improves the efficiency compared with the other state-of-the-art approaches.

MMM: Classification of Schizophrenia Using Multi-modality Multi-atlas Feature Representation and Multi-kernel Learning（99）

Jin Liu, Jianxin Wang, Xiang Wang, Fang-Xiang Wu and Yi Pan

摘要: Schizophrenia (SCZ) is a complex neuropsychiatric disorder that seriously affects the daily life of patients. Therefore, the accurate diagnosis of SCZ including its subtypes (e.g., deficit SCZ (DSCZ) and nondeficit SCZ (NDSCZ)) is essential for patient care. Several T1-weighted magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI) markers (e.g., cortical thickness

(CT), mean diffusivity (MD)) for SCZ have been identified by using some existing brain atlases, and have been used successfully to discriminate SCZ patients from healthy controls (HCs). Currently, these markers have mainly been used separately. Thus, the full potential of T1-weighted MRI images and DTI images for SCZ diagnosis might not yet have been used comprehensively. Furthermore, the extraction of these markers based on single brain atlas might not yet be able to use the full potential of these images. Therefore, in this study, we propose a multi-modality multi-atlas feature representation-based multi-kernel learning method (MMM) to perform SCZ classification. Firstly, we extract 8 feature sets from T1-weighted MRI images and DTI images via 4 existing brain atlases and 4 markers. Then, a three-step feature selection method is proposed to select the most discriminative features of each feature set for SCZ classification. Finally, a multiple feature sets-based multi-kernel SVM learning method (MFMK-SVM) is proposed to combine multiple feature sets for SCZ classification. Experimental results show that our proposed method achieves an accuracy of 89.88% and an AUC of 0.9473 for SCZ/HC classification, an accuracy of 92.75% and an AUC of 0.9601 for DSCZ/HC classification, an accuracy of 88.05.79% and an AUC of 0.9286 for NDSCZ/HC classification, and an accuracy of 74.32% and an AUC of 0.7418 for DSCZ/NDSCZ classification, respectively. Experimental results demonstrate that our proposed classification method is efficient and promising for clinical applications for the diagnosis of SCZ, especially DSCZ and NDSCZ.

Cancer Classification Based on Support Vector Machine Optimized by Particle Swarm Optimization and Artificial Bee Colony（52）

Lingyun Gao, Mingquan Ye and Changrong Wu

摘要：Intelligent optimization algorithms have advantages in dealing with complex nonlinear problems accompanied by good flexibility and adaptability. In this paper, the FCBF (Fast Correlation-Based Feature selection) method is used to filter the irrelevant and redundant features in the data. Then we perform the cancer classification based on SVM (Support Vector Machine) optimized by PSO (Particle Swarm Optimization) and ABC (Artificial Bee Colony) approaches, which is represented as PA-SVM. The proposed method PA-SVM is applied to nine cancer datasets, including five datasets of outcome prediction and an ovarian cancer dataset of protein data. By comparison with other classification methods, the results demonstrate the effectiveness and the robustness of the proposed method PA-SVM in handling various types of data for cancer classification.

Ensemble Classification for Gene Expression Data based on Parallel Clustering（12）

Jun Meng, Ding-Ling Jiang, Jing Zhang and Yu-Shi Luan

摘要：Analysis of large-scale gene expression data is a research hotspot in the field of bioinformatics, which can be used to diagnose the disease of human and animal, and to study the abnormal phenomenon in plant growth process. This paper proposes a biological knowledge integration method based on parallel clustering to select gene subsets effectively. Gene ontology is utilized to obtain the biological function similarity, and combine it with gene expression data. Parallelized affinity propagation algorithm is used to cluster fusion data since it can not only obtain more biologically meaningful subsets, but also avoid the loss of some potential value in genes from simple gene primary selection. Based on clustering result, neighborhood rough set is used to select representative genes which are used to train classifier for each cluster. Random hill climbing search algorithm is applied for classifier selection for an appropriate group of classifiers for ensemble classification. Experimental results on plant stress response datasets demonstrate that the proposed method can select genes with stronger classification ability.

Modeling and control of a delayed Hepatitis B virus model with incubation period and combination treatment（38）

Deshun Sun and Fei Liu

摘要：In this paper, a Hepatitis B virus (HBV) model with an incubation period, and delayed state and control variables is firstly proposed; furthermore the combination treatment is adopted in order to have a longer-lasting effect than mono-therapy. The equilibrium points and basic reproduction number are calculated, and then the local stability is analyzed. We then present optimal control strategies based on Pontryagin's minimum principle with an objective function that is not only to reduce the levels of exposed cells, infected cells and free viruses nearly to zero at the end of therapy, but also to minimize the drug side-effect and the cost of treatment. What's more, we develop a numerical simulation algorithm for solving our HBV model based on the combination of forward and backward difference approximations. The state dynamics of uninfected cells, exposed cells, infected cells, free viruses, CTL and ALT are simulated with or without optimal control, which show that HBV is reduced nearly to zero based on the time-varying optimal control strategies whereas the disease would break out without control. At last, by the simulations, we prove that strategy A is the best among the three kinds of strategies.

基于网络约束双聚类的癌症亚型分类（76）

Xing Wang, Jun Wang, Guoxian Yu and Maozu Guo

摘要：利用双聚类算法在大规模基因表达数据上进行聚类分析可以发现不同的癌症亚型，结合基因网络数据可以提高癌症亚型分类的准确度. 已有整合网络的双聚类算法通常仅基于基因的度加权选择基因，易受网络中噪声互作的干扰和缺失互作的误导. 为此，本文提出了一种基于基因网络正则化的双聚类算法(Network Regularized Bi-Clustering algorithm，

NetRBC)．NetRBC 首先通过最小化聚类簇上的平方残差分别求取癌症基因表达数据矩阵上的行/列簇指示矩阵；其次利用基因网络和行簇指示矩阵构建图正则项；最后将此正则化项结合到基于平方残差的非负矩阵分解中，约束行簇和列簇矩阵的协同分解，以期提高癌症亚型分类的精度. 在多个癌症基因表达数据上的实验结果表明 NetRBC 比已有相关方法能够更准确的区分癌症亚型.

Protein solvent accessibility prediction by stacked deep bidirectional recurrent network（13）

Buzhong Zhang and Qiang Lyu

摘要：The prediction of residue solvent accessibility(RSA) can provide more information for analyzing protein structures and functions. Many computing methods have been proposed to predict it for better performance of prediction is more useful for protein study. In this work, we present a deep learning method to predict the solvent accessibility which is based on stacked deep bidirectional recurrent network(SDBRNN) applied to the sequence profiles. By this method, the continuous relative solvent accessible area prediction as well as two-state discrete prediction can be obtained.The Bidirectional Long Short-term Memory(BLSTM) network, a typical bidirectional recurrent network is first adopted to predict RSA. In order to obtain more long-ranged sequence information, merging operator is proposed when bidirectional information from hidden nodes is merged for outputting.Three types of merging operator are used in our improved model, SDBRNN. The trained database is constructed from 7361 proteins extracted form PISCES culling server with the cutoff 25% sequences identities.Three public benchmark data sets, CB502, Manesh215 and CASP10 are used for testing model. The experimental results show that our deep learning method achieves significant improvement on the prediction quality. MAE, PCC is 8.8% and 74.9% on CB502, 8.3% and 78.1% on Manesh215 which are all better than existing method on relative solvent accessibility. and the two state prediction of our method is also obtained the best performance except the accuracy of 50% threshold prediction

A Generic Multi-Cellular Biological Simulation Platform based on CUDA(51)

Siyu Yang, Jinzhi Lei and You Song

摘要: Simulating multi-scale dynamics of complex living systems is the major challenge in the researches of computational system biology. In this work, we propose a CUDA-based generic multi-cellular biological simulation platform. The platform supports user implementing cell-centered hybrid models, which is a popular modeling structure in computational system biology. Based on GPU and NVIDIA CUDA framework, simulations implemented using the platform enjoy high-performance offered by parallel computing. We propose to 摘要 complex

data structures and CUDA-related algorithm implementations into object-oriented programming interfaces, which is defined as common biological concepts like components and behaviors. Researchers are able to declare their computational model in the form of declaration in code, which not only makes it looks intuitive but provides high-level of flexibility in model scaling. The Custom Functions mechanism integrates user's model implementation naturally with parallel programming paradigm. And several methods are proposed to optimize memory management during biological simulation. We have tested the platform with a model from real research project, which presents comparable code complexity, better code readability as well as scalability, and 2-22 times simulation acceleration compared with corresponding sequential implementation.

Analysis of ribosome stalling and translation elongation dynamics by deep learning(highlight) (101)

Jianyang Zeng

摘要: Ribosome stalling is manifested by the local accumulation of ribosomes at specific codon positions of mRNAs. Here, we present ROSE, a deep learning framework to analyze high-throughput ribosome profiling data and estimate the probability of a ribosome stalling event occurring at each genomic location. Extensive cross validation tests demonstrated that ROSE possessed higher prediction accuracy than conventional prediction models, with an increase in AUROC by up to 18.4%. In addition, genome-wide statistical analyses showed that ROSE predictions can be well correlated with diverse putative regulatory factors of ribosome stalling. Moreover, the genome-wide ribosome stalling landscapes of both human and yeast computed by ROSE recovered the functional interplays between ribosome stalling and cotranslational events in protein biogenesis, including protein targeting by the signal recognition particles (SRPs) and protein secondary structure formation. Overall, our study provides a novel method to complement the ribosome profiling techniques and further decipher the complex regulatory mechanisms underlying translation elongation dynamics encoded in mRNA sequence.

基于 Docker 技术架构高移植性生物信息数据软件流虚拟 web 平台 (18)

Minglei Yang and Weizhong Li

摘要：Docker 应用容器引擎可实现打包生物信息数据流应用程序以及依赖包到一个可移植的容器中，然后部署到任何主流的 Linux 机器上。本实验室利用 Docker 技术结合 make 搭建面向 RNA-Seq、全基因组重测序、Pacbio 三代全长转录组测序等生物信息分析软件工作流程的 Docker 容器。产出的大型工作流可以实现 RNA-Seq 表达差异分析及 GO、KEGG 等相关注释分析，同时能实现对 fusion genes，lncRNAs 的鉴定分析。另外，可以实现全基因组 variation 的分析注释以及实现 Pacbio 三代全长转录组的数据分析。再次，运用 Docker

技术沙箱机制和 django 框架搭建基于二，三代测序技术的高移植性、高操作性、高可视化的 web 应用平台。未来将继续在此基础上开发免疫组库，HLA 分型，甲基化等生物信息学分析流程。

Classification and Feature Selection via Sparse Multi-view Low-Rank Regression (48)

Yao Lu, Ying-Lian Gao, Jin-Xing Liu, Xiang-Zhen Kong, Jun-Liang Shang and Chun-Hou Zheng

摘要: Multi-view classification and feature selection have received considerable attention in recent years. In many real classification problems, the data in each view may have noise. The low-rank regression model has been proved and applied to capture underlying classes correlation patterns, such that the classification results can be enhanced. In order to make it sparse, in this paper, we propose a novel method for sparse multi-view low-rank regression (SMLLR) and sparse multi-view full-rank regression (SMFRR). The method based on sparse theory is to make the matrix decomposition produce sparse results by adding the penalty factors in the matrix transformation process. The SMLLR model is constructed by imposing sparse low-rank constraints on the objective function. The L2-norm and L2,1-norm constraints are added to loss and regularization functions, respectively. The model is convenient for feature selection because the L2,1-norm regularization can penalize each row of the matrix and enforce sparsity among the rows. This method has a flavor of low-rank and sparse decomposition. Experimental results on gene expression datasets show that the proposed method has superior performance over several state-of-the-art methods for multi-view classification and feature selection.

Selecting Feature Subset Based on SVM-RFE and Overlapping Ratio (32)

Xiaohui Lin, Chao Li, Yanhui Zhang, Meng Fan and Hai Wei

摘要: Defining informative features from complex and high dimensional biological data is of great importance in disease study, drug development, etc. Support vector machine - recursive feature elimination (SVM-RFE) is a very popular data analysis technique and has shown its power in many fields. It ranks the features according to the recursive deletion sequence based on SVM. This paper studies the criteria to determine how many top ranked features should be selected according to the recursive feature elimination procedure of SVM-RFE. The classification accuracy rate and overlapping ratio of the samples on the current subspace are adopted to measure the discriminative ability of the feature subset. Meanwhile to measure the weights of the features more accurately, the samples which lie in a heavy overlapped area are temporally blocked in the iteration of SVM-RFE. The experiments on the eight public biological data sets show that combining the accuracy rate and overlapped degree of the samples could select a more powerful

feature subset than using classification accuracy rate alone. And temporally screening the heavily overlapped samples in a iteration of SVM could weigh the feature discriminative ability more accurately. Combining the two techniques together could define a more informative feature subset.

一种基于节点间路径度量的图聚类算法 (94)

Wenping Zheng, Chenhao Che and Gui Yang

摘要： 图聚类算法可以用于发现社会网络中的社区结构、蛋白质互作用网络的功能模块等,是当前复杂网络研究的热点之一。合理度量网络中节点的相似性是设计有效图聚类算法的核心问题. 针对此问题,本文提出了一种基于两点间短路径的节点相似性度量方法,并在此基础上给出了一种面向复杂网络的图聚类算法(A Graph Clustering Algorithm Based on Paths between Nodes in Complex Networks, PGC).首先基于通过一对节点对间存在的长度不超过 3 的路径数给出了一种节点相似性度量的方法;其次,选择与网络中所有节点相似性较高且与已有种子节点相似性较低的节点作为种子节点;将其余节点分配至与该节点相似性最高的种子节点所在模块以得到初始模块划分结果;最后,在初始模块划分基础上,以文献[15]提出的基于互补熵的簇质量评价函数为目标函数更新每个模块的种子节点,并迭代更新模块划分结果,得到最终簇发现结果.通过与 FMM 算法, LPA 算法, BGLL 算法, ISCD+算法,MCODE 算法等在 4 个带标签网络数据集、GAVIN2006 蛋白质互作用网络数据集及人工网络数据上进行分析比较结果表明,算法具有良好的性能.

An Improved Algorithm on Graph Canonization Problem (66)

Jing Li, Jialu Hu and Xuequn Shang

摘要：Graph canonization is a fundamental problem both in theoretical and practical computer science. However, it is still an open problem to study in graph theory. In this paper, we propose a new graph canonization algorithm based on resolving sets, also known as distinguishing sets in some literatures. In theory, we prove the existence of a tighter complexity bound of graph canonization problem, $o(\exp(\sqrt{n \log_2 n + 4 \log n}))$, on strongly regular graphs with $\mu = \lambda + 1$ using a statistical model. Furthermore, a fast and effective computational tool, sgip, was developed in the distribution of SeqAn library. To test the performance of sgip and a notable package nauty, both of them were performed on a same graph benchmark databases. The results show that sgip outperforms nauty package on many graph cases, e.g. multi-dimension meshes. The source code of sgip is freely accessible in https://github.com/seqan/seqan/tree/master/apps/sgip and the binary code in http://packages.seqan.de/sgip/.

An Interface for Biomedical Big Data Processing on the Tianhe-2 Supercomputer（21）

Xi Yang, Chengkun Wu, Kai Lu, Lin Fang, Yong Zhang, Shengkang Li, Guixin Guo and Yunfei Du

摘要：Big data, cloud computing and HPC are at the verge of convergence. Cloud computing is already playing an active part in big data processing with the help of big data frameworks like Hadoop and Spark. Recently, the upsurge of high performance computing in China provides extra possibilities and capacity to address the big data challenge. In this paper, we proposed Orion, a big data interface on the Tianhe-2 supercomputer, to enable big data applications to run on Tianhe-2 via a single command or a shell script. Orion supports multiple users and each user can launch multiple tasks. It minimizes the effort needed to initiate big data applications on the Tianhe-2 supercomputer via automated configuration. Orion follows the “allocate-when-needed” paradigm and it avoids idle occupation of computational resources. We tested the utility and performance of Orion using a big genomic dataset and achieved a satisfactory performance on Tianhe-2 with very few modifications to existing applications that were implemented in Hadoop/Spark. In summary, Orion provides a practical and economical interface for big data processing on Tianhe-2.

SIMBA: a single molecule-guided Bayesian localization microscopy for practical live cell super-resolution imaging（37）

Fan Xu and Fa Zhang

摘要：Practical live-cell super-resolution (SR) techniques are long-desired in many routine biological labs to image biomolecule dynamics. However, the current methods either require sophisticated optical setups and deep experts or have difficulties to achieve high spatial and temporal resolutions simultaneously. Here we present a powerful single molecule guided Bayesian localization microscopy (SIMBA), which uses simple off-the-shelf total internal reflection fluorescence (TIRF) equipment to produce an appropriate 50 nm SR image with 0.5-2 s total acquisition time in living cells. The SIMBA calculates a series of whole-cell live structures for 50 time points on a desktop computer. The reconstruction results show reliable structures with practical resolution comparable with PALM results. With good compatibility to TIRFM, PALM/STORM and light-sheet microscopy equipped in many labs, SIMBA should be useful in a wide variety of live-cell SR imaging applications.

癌症组学数据的低维表示（79）

Jin Gu

摘要：癌症组学数据是典型的高维数据，即使在经过预处理后，特征数量通常也在数千或数万的量级。虽然特征众多，但生物分子并不是独立的，它们之间存在大量相互作用，形成网络结构，并自组织为若干生物过程。研究表明，癌症的生物学与临床特性主要由少数起驱动作用的生物过程决定，找到组学数据的低维特征表示、并建立其与癌症生物过程的对应关系，是揭示癌症生物学机制与调控规律的重要途径。我们提出了多个组学数据低维表示的方法，并将其应用于癌症高维组学数据的可视化、癌症分型与聚类分析：1）找到多层次组学数据共享的低维子空间，对发现起主要调控作用的生物过程具有重要意义。我们提出了基于低秩近似的降维方法 LRAcluster，该模型借鉴了 iCluster+用广义线性回归来处理不同类型组学数据的办法，但将目标函数中的因子分析的概率似然改为基于核范数的低秩约束。2）我们提出了基于深度变分自编码器（deep variational AutoEncoder）的多层次组学特征表示与低维可视化算法 VASC，该方法可对组学特征进行多层次自动表示，在多个测试数据集上表现出良好的性能，并有助于发现组学特征之间的结构信息。

Deep convolutional neural networks-based early automated detection of diabetic retinopathy in fundus image（59）

Kele Xu, Dawei Feng and Haibo Mi

摘要：The automatic detection of the diabetic retinopathy is of importance, as it is the main cause of irreversible vision lost in the working-age population in the developed world. Early detection of diabetic retinopathy occurrence can be greatly helpful for the clinical treatment, although several different feature extraction approaches have been proposed, the classification task for retinal images is still tedious even for the trained clinicians. Recently, deep convolutional neural networks have manifested superior performance in image classification compared to previous handcrafted feature-based image classification methods. Thus, in this paper, we explore the use of deep convolutional neural network methodology for the automatic classification of diabetic retinopathy for color fundus image, and we obtained an accuracy of 94.5% on our dataset, outperforming the results obtained by using classical approaches.

A thickness-based iterative non-uniform Fourier reconstruction algorithm for electron tomogram(85)

Lun Li, Yu Chen and Fa Zhang

摘要：Electron tomography (ET) is a widely applicable method for obtaining three-dimensional information by their projections. A major challenge for reconstruction is the limited sample angles (the "missing wedge" problem). Iterative Non-uniform Fast Fourier Transform (NUFFT) reconstruction (INFR) is capable of retrieving meaningful information in some region of the missing wedge in Fourier space for biological dataset and generating high resolution reconstruction. However, the huge computational demand, the enormous amount of memory consumption have become the major problems for the application of INFR. In this work, we developed a thickness-based iterative non-uniform Fourier reconstruction algorithm called region-of-interest INFR by restricting the thickness of the reconstructed volume in INFR. Accordingly, we modified the NUFFT and adjoint NUFFT in INFR and parallel them on graphics processing unit (GPU) to generate a GPU program region-of-interest INFR-GPU. Experiment results show that, by using region-of-thickness strategy, we can reduce memory consumption, speed up the running rate and improve reconstruction resolution at the same time.

一种面向大规模序列数据的交互特征并行挖掘算法（96）

Yuhai Zhao

摘要：序列是一种重要的数据类型，在诸多应用领域广泛存在。基于序列的特征选择具有广阔的现实应用场景。交互特征是指一组整体具有显著强于单独个体与目标相关性的特征集合。从大规模序列中挖掘交互特征面临着位点的“组合爆炸”问题，计算挑战性极大。针对该问题，以生物领域高通量测序数据为背景，提出了一种新的基于并行处理和演化计算的高阶交互特征挖掘算法。位点数是制约互作挖掘效率的根本因素。本研究摈弃了现有方法基于序列分块的并行策略，采用基于位点分块的并行思想，具有天然的效率优势。进一步，提出了极大等位公共子序列（Maximal Allelic Common Subsequence，简称 MACS）的概念并设计了基于 MACS 的特征区域划分策略。该策略能将交互特征的查找范围缩小至许多“碎片”空间，并保证不同“碎片”间不存在交互特征，避免计算耦合引起的高额通讯代价。利用基于置换搜索的并行蚁群算法，执行交互特征选择。大量真实数据集和合成数据集上的实验结果，证实本文提出的 PACOIFS 算法在有效性和效率上优于同类其他算法。

Multi-objective optimization algorithm to discover condition-specific modules in multiple networks (20)

Xiaoke Ma and Penggang Sun

摘要：Many biological networks have been accumulated with various conditions due to the advances in high-throughput technologies. And, discovering the condition-specific modules in multiple networks has a great merit in understanding the underlying molecular mechanisms of cells. The available algorithms construct a specific network for each condition based on the multiple networks and discover modules in the constructed network, which are criticized by the low accuracy because they ignore the connection among the multiple networks. To attack this issue, we define the condition-specific module as a group of genes whose connectivity is strong in the corresponding condition and weak in others, which provides a better way to characterize specific modules because all the networks are taken into consideration. A stringent mathematical model is presented, where the condition-specific module discovery problem is formulated as a multi-objective optimization problem. Then, a multi-objective genetic algorithm for condition-specific modules in multiple networks (aka, MOGA-CSM) is developed to discover specific modules. By using the artificial networks, we demonstrate that
the MOGA-CSM algorithm outperforms state-of-the-art methods in terms of accuracy. Furthermore, the MOGA-CSM discover stage-specific modules in breast cancer networks based on TCGA data, and these modules serves as biomarkers to predict stages of breast cancer. The proposed model and algorithm provide an effective way to analyze multiple networks.

Fusion Analysis of Resting State Networks And Its Application to Alzheimer's Disease(91)

Shengbing Pei, Jihong Guan and Shuigeng Zhou

摘要： Functional networks are extracted from resting state functional magnetic resonance imaging data to explore biomarkers for distinguishing brain disorder in disease diagnosis. Previous works have primarily focused on using a single resting state network (RSN) with various techniques. Here, we apply fusion analysis of RSNs to capturing biomarkers, which can combine the complementary information among the RSNs. Experiments are carried out on three groups of subjects, i.e., cognition normal, early mild cognitive impairment and Alzheimer's disease (AD), which correspond to the three progressing stages of Alzheimer's disease, and each contains 18 subjects. First, we apply group independent component analysis (ICA) to extracting default mode network (DMN) and dorsal attention network (DAN) for each subject group, then by taking the common DMN and DAN as templates for each group, we employ individual ICA to extract DMN and DAN for each subject, and finally we fuse these DMNs and DANs to explore biomarkers. Results show that (1) the templates generated by group ICA can help extract the RSN for each subject by individual ICA effectively; (2) RSNs extracted by fusion analysis can obtain more informative biomarkers than without fusion analysis; (3) The most different regions of DMN and DAN in the early and late stages of AD are different. For DMN, medial prefrontal cortex affected decreases in late stage while posterior cingulate affected increases; And for DAN, intraparietal sulcus affected decreases in late stage; 4) Extracting DMN and DAN for each subject via back reconstruction of group ICA seems not valid.

会议 Poster 列表

A bioinformatics web platform for omics data analysis (46)
Weizhong Li

Identifying Drug-pathway Association Pairs via GL2,1-Integrative Penalized Matrix Decomposition (22)
Dong-Qin Wang, Ying-Lian Gao, Jin-Xing Liu and Chun-Hou Zheng

Selecting Near-native Protein Structures from Ab Initio Models Using Ensemble Clustering (33)
Li Li, Yong gang Lu and Huanqian Yan

A fast projection-based algorithm for clustering big data (41)
Yun Wu, Zhiquan He, Hao Lin, Yufei Zheng, Jingfen Zhang and Dong Xu

Deletion Genotype Calling on the Basis of Convolutional Neural Network (61)
Jing Wang, Jingyang Gao and Jiayin Wang

Malopred: an online prediction tool for lysine malonylation (47)
Lina Wang, Zhiyou Zhou and Jianding Qiu

Identifying novel human miRNA-disease association based on double layer random walk model (64)
Chaokun Yan, Li Ma, Junwei Luo and Huimin Luo

Alignment of Dynamic Protein-Protein Interaction Networks Based on Segment Tree Optimization (88)
Jiaye Zhu, Yinglong Song, Jihong Guan and Shuigeng Zhou

A Novel Computational Method for Detecting DNA Methylation Sites with Sequence and Physical Structural Properties (30)
Fei Guo and Gaofeng Pan

A Survey of Computational Methods of the RNA-Seq Data Analysis and Applications (69)
Yang Guo

基于梯度投影算法的复杂网络模块划分方法 (9)
Wenwen Min and Juan Liu

GDPTrDB：connecting Genotype, Disease, Phenotype & Treatment (16)

Wenliang Zhang and Weizhong Li

Detecting diagnostic biomarkers of Alzheimer disease by integrating gene expression data in six brain regions (77)
Lihua Wang and Zhi-Ping Liu

赞助单位

深圳市早知道科技有限公司

深圳华大基因股份有限公司

人和未来生物科技（长沙）有限公司

华智水稻生物技术有限公司

英文合作期刊

- Neurocomputing
- MOLECULES
- Int. J. DataMing and Bioinformatics
- TSINGHUA SCIENCE AND TECHNOLOGY
- Interdisciplinary Sciences: Computational Life Sciences
- Quantitative Biology

中文合作期刊

- 《中国科学（信息学版）》
- 《计算机学报》
- 《计算机研究与发展》
- 《计算机科学》