浙江省生物信息学学会
Bioinformatics Society of Zhejiang Province

浙苏两省生物信息学学术研讨会

会议手册

2018年3月30 - 4月1日

浙江 杭州

联合主办：浙江省生物信息学学会
江苏省生物信息学专业委员会

一、会议日程

时间		**内 容**	地点
30日15:00-17:30 31日：8:00-9:00		**报到**	唐人儒亿酒店大堂 浙江大学生命科学学院报告厅门厅
上午	8:45-9:00	**开幕式** 1、浙江省生物信息学会理事长 陈铭教授致辞 2、江苏省生物医学工程学会生物信息学专业委员会主任 孙啸教授致辞	地点：生命科学学院245报告厅 主持人:樊龙江副理事长
	9:00-9:30	大会特邀报告：王炜 教授（南京大学） 报告题目： Fluctuations in native states of proteins - critical behaviors	主持人：王俊教授
	9:30-10:00	大会特邀报告：李飞 教授（浙江大学） 报告题目： 昆虫基因组数据库	
	10:00-10:20	**茶 歇**	生命科学学院一楼
	10:20-10:45	大会报告：徐辰武 教授（扬州大学） 报告题目：关联作图和基因组预测的多元统计方法	主持人：宋晓峰教授、郭俊明教授
	10:45-11:10	大会报告：朱峰 教授（浙江大学） 报告题目：Identification of Next Generation Innovative Therapeutic Target from OMICs Data	
	11:10-11:35	大会报告：代琦 教授（浙江理工大学） 报告题目： Bioinformatics Platform for Genomic Island Detection and Analysis	
	11:35-12:00	大会报告：丁彦蕊 教授（江南大学） 报告题目：蛋白质热稳定性的生物信息学研究	
中午	12:00-13:30	**集体合影 & 午餐**	校友中心台阶 食堂二楼东区

下午	13:30-13:55	大会报告：孙啸教授（东南大学） 报告题目：多样本混合测序的编码设计和解码算法	主持人：李飞教授、 潘沈元教授
	13:55-14:20	大会报告：刘鹏渊 教授（浙江大学） 报告题目：肿瘤中 tRNA 来源小片段的基因组图谱	
	14:20-14:45	大会报告：陈兴 教授（中国矿业大学） 报告题目：大数据时代下基于网络算法和机器学习的生物信息学研究	
	14:45-15:10	大会报告：罗洁 副研究员（浙江省农科院） 报告题目：基于因果关系分析的特定表型相关调控网络研究	
	15:10-15:30	**茶 歇**	生命科学学院一楼
	15:30-15:55	大会报告：宋晓峰 教授（南京航天航空大学） 报告题目：环形 RNA 编码蛋白潜能的生物信息学研究	主持人：马飞教授、 刘庆坡教授
	15:55-16:20	大会报告：甘卓慧 副教授（温州医科大学） 报告题目：高通量基因表达数据分析工具的整合、流程自动化和跨物种分析比较	
	16:20-16:45	大会报告： 吴世华 副教授（浙江大学） 报告题目： 天然产物分离及模式识别技术	
	16:45-17:10	简短报告交流讨论： （中国矿业大学、上海培优教育、杭州厚泽生物）	
	17:10-17:20	会议总结：陈铭、孙啸	

4 月 1 日：自由交流讨论、离会

- **会议费用**

个人会员：3 月 20 日前本学会学生会员不收取会务费，其他人员注册费见以下说明（主要用于会议期间餐饮、茶歇等）。与会代表的差旅费、住宿费自理。

缴费时间	学生会员	其他会员	非会员
3 月 20 日前	免费	100	200
3 月 20 日后及现场	100	200	300

企业会员：赞助招展请联系秘书处，提供 1 个展位和 2 名人员免费注册，会议手册 A4 宣传页，会议网站宣传、学术报告。本学会企业会员单位免费拥有 1 个展位（参会人员以学会其他会员身份注册）。

会议预注册网址：

http://www.zjbioinformatics.org/meetings/2018/showMember.php

- **住宿酒店**

请提前一周时间预定，说明参加“浙江大学会议”享受协议价。

酒店名称	预定电话	标准间	单间	含早	备注
唐人儒亿大酒店	0571-28233333	318 元/间	368 元/间	双早	三坝站地铁口边的酒店
紫金港大酒店	0571-88969999	338 元/间	398 元/间	双早	校园内的酒店，离车站远

- **会务组联系方式**

黄老师：13666652591 huangfl@zju.edu.cn；

沈老师：13666606799 xxshen@zju.edu.cn

- **路线**

自驾：导航浙江大学紫金港校区

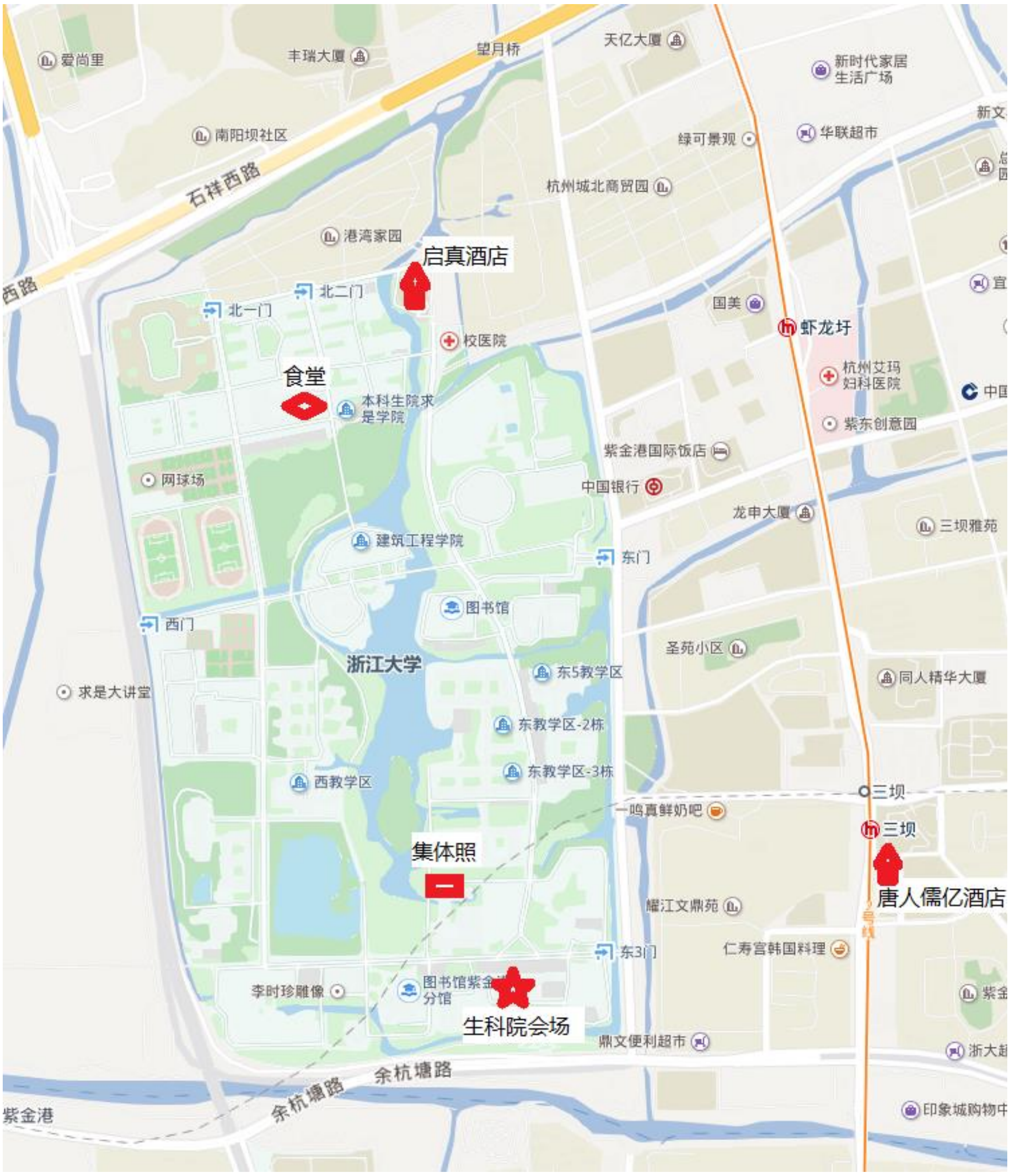
地铁站：地铁 2 号线三坝站步行至校区约 10 分钟；

高铁站：地铁 1 号线至凤起路站换乘 2 号线至三坝站，步行约 10 分钟到达；

公交车：10 路、74 路、89 路、97 路，227 路、505 路浙大紫金港校区站；

机场：机场大巴至武林门，换乘地铁 2 号线至三坝站步行约 10 分钟到达。

位置图


启真酒店
食堂
集体照
生科院会场
唐人儒亿酒店
浙江大学
爱尚里
丰瑞大厦
望月桥
天亿大厦
新时代家居生活广场
南阳坝社区
石祥西路
绿可景观
华联超市
杭州城北商贸园
港湾家园
北二门
北一门
校医院
国美
虾龙圩
杭州艾玛妇科医院
紫东创意园
本科生院求是学院
紫金港国际饭店
网球场
中国银行
龙申大厦
三坝雅苑
建筑工程学院
东门
图书馆
西门
圣苑小区
东5教学区
同人精华大厦
求是大讲堂
东教学区-2栋
东教学区-3栋
西教学区
三坝
一鸣真鲜奶吧
耀江文鼎苑
东3门
仁寿宫韩国料理
李时珍雕像
图书馆紫金港分馆
鼎文便利超市
余杭塘路
紫金港
印象城购物中
2号线

二、报告人简介

王　炜 南京大学物理学院教授（1992 年），博士生导师（1995 年），长江计划特聘教授（1999 年度，理论物理），南京大学生物物理研究所所长,匡亚明学院院长。1996 年度获国家杰出青年基金，1997 年获香港求是基金会-杰出青年研究奖(物理学)，科技部 973 项目“非线性科学及其重要应用”首席科学家（2007-2011），科技部 973 项目“与激光聚变、自然灾害和深空探测等相关的非线性动力学斑图和轨道稳定性研究”首席科学家（2013-2017）。现任《Proteins》，《中国科学：物理-力学-天文》，《Chinese Physics Letters》等杂志编委。

科研工作主要涉及凝聚态物理与生物交叉学科的研究：蛋白质折叠、聚集动力学以及复杂相互作用下生物分子组装微结构特性；生物网络系统信息过程的物理机制和动力学特性。在 Nature 子刊、PNAS、PRL 和 JACS 等期刊上发表学术论文 200 余篇。

报告题目：　Fluctuations in native states of proteins--critical behaviors

李　飞 浙江大学教授、博士生导师。1977 年 1 月。1996、2003 年分别获南京农业大学学士、博士学位， 2005 年清华大学自动化系博士后出站，2006 年美国冷泉港实验室访问学者。2005 年破格提升为南京农业大学教授，2006 年前起任博士生导师。2015 年 8 月起，任浙江大学教授、博士生导师。入选浙江大学求是特聘教授、教育部新世纪优秀人才、全国百篇优秀博士论文获得者、江苏省杰出青年基金、浙江省 151 人才第二层次、江苏省 333 高层次人才培养工程第三层次、清华大学优秀出站博士后。曾任第四届国际昆虫生理生化学术会议执行主席兼秘书长、第三届国际昆虫基因组大会及第六届国际生理生化学术会议秘书长。

报告题目：昆虫基因组数据库

徐辰武 扬州大学二级岗位教授农学博士，博士生导师，享受国务院政府特殊津贴专家，《生物统计与试验设计》国家精品课程和国家精品资源共享课程负责人。先后主持国家重点研发计划课题 1 项、国家 973 计划课题 2 项、国家自然科学基金项目 7 项；发表学术论文 160 余篇，其中以通讯作者在 Trends in Plant Science、Plant Physiology、New Phytologist 等刊物上发表 SCI 论文 50 余篇；先后入选全国农业科研杰出人才和创新团队带头人、教育部“新世纪优秀人才支持计划”、江苏省“333 高层次人才培养工程”中青年科学技术带头人和江苏省高校“青蓝工程”省级中青年学术带头人和创新团队带头人等人才计划。主要从事作物数量性状遗传分析新方法和作物分子设计育种研究工作。

报告题目：关联作图和基因组预测的多元统计方法

朱峰 博士，百人计划岗研究员，博士生导师，国家青年千人计划项目入选者，中国化学会计算机化学委员会委员，中国计算机学会生物信息学专业组委员。2011 年在新加坡国立大学获得博士学位。主要从事生物信息学、药物设计和基于大数据的精准医疗研究，已在 Nature Biotechnology 等生物医药领域期刊上发表 SCI 论文 50 篇。主持国家重点研发计划和重庆市重点产业共性关键技术创新重大专项各一项，并已完成国家自然科学基金青年项目和重庆市自然科学基金计划各一项。

报告题目：Identification of Next Generation Innovative Therapeutic Target from OMICs Data

代琦 浙江理工大学生物信息学学科组主任、硕士生导师。2012 年入选浙江省高校中青年学科带头人；2013 年入选浙江理工大学“521”拔尖人才。2014 年在美国德州大学达拉斯分校 Micheal. Zhang 实验室从事研究。主持国家级项目 3 项，浙江省项目 2 项。担任国际期刊《Computational and Mathematical Methods in Medicine》、《Journal of Co4mputational Biology and Bioinformatics Research》等编委。现阶段主要研究方向有功能基因组分析，生物信息智能化处理，肿瘤早期分子诊断中的信息处理等。

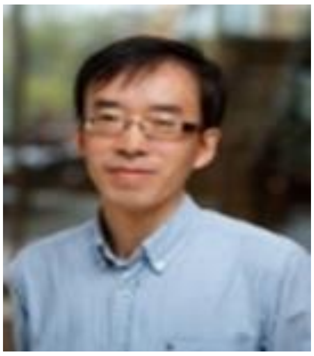
报告题目：Bioinformatics Platform for Genomic Island Detection and Analysis

丁彦蕊，江南大学数字媒体学院教授，计算机学会会员。主要研究方向为生物信息学，重点分析蛋白质的耐热机制。主持国家自然科学基金项目 2 项，“863”重点项目子项目 1 项，江苏省环境检测科研基金项目 1 项，教育部留学回国人员基金 1 项；参与项目主要包括“863”项目、科技部重大仪器专项、国家自然科学基金、江苏省科技成果转化专项资金 A 类重点项目、江苏省科技攻关项目等。以第一作者或通讯作者发表 SCI 收录论文 17 篇，申请发明专利 36 项，获得软件著作权 1 项。

报告题目：蛋白质热稳定性的生物信息学研究

孙 啸 东南大学教授，博士生导师。目前任江苏省生物医学工程学会副理事长，生物信息学专业委员会主任，中国细胞学会功能基因组信息学与系统生物学分会常务理事，中国医药生物技术协会生物医学信息技术分会常务委员。主要从事生物信息学研究，重点研究高通量 DNA 测序数据挖掘和表观基因组信息分析。在 Nucleic Acids Res.、Bioinformatics、DNA Res.等期刊上发表论文 100 多篇，出版专著 2 部，获得发明专利 2 项，获得软件著作权 5 项。

报告题目：多样本混合测序的编码设计和解码算法



刘鹏渊 浙江大学转化医学研究院，浙江大学医学院附属邵逸夫医院教授、博士生导师，国家第二批青年千人计划入选者。长期从事癌症遗传和基因组学、统计遗传学和生物信息学的研究。已在癌症遗传学、人类疾病遗传学、统计遗传学和生物信息学领域发表了 76 篇包括 Nature Genetics，PloS Medicine，American Journal of Human Genetics 和 Cancer Research 等 SCI 论文，第一、二作者 SCI 论文 38 篇，最高论文影响因子 F=32.7（Nature Genetics），共被 SCI 刊物引用 1300 多次。2010 年获教育部自然科学奖一等奖。

题目：肿瘤中 tRNA 来源小片段的基因组图谱

陈 兴 中国矿业大学信息与控制工程学院教授（直接破格），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，江苏省“六大人才高峰”高层次人才，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，江苏省自动化学会生物信息专业委员会（筹）主任，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，江苏省人工智能学会智能系统与应用专业委员会委员。担任 45 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 杂志副主编、SCI 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)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）等六家 SCI 杂志首席特约编委。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 78 篇（SCI 论文 73 篇，EI 检索 3 篇，影响因子累计约 318），其中第一作者 39 篇, 通讯作者 58 篇。以一作或者通讯发表中科院一区论文 21 篇，以一作或者通讯发表中科院二区以上论文 43 篇，以一作或者通讯发表 JCR 一区论文 45 篇。发表论文被 Nature Reviews Genetics、Nature Chemistry、Nature Reviews Endocrinology 等引用共计 1957 次，目前 8 篇论文为 ESI 高被引论文，1 篇论文为 ESI 热点论文，其中被影响因子 5 以上杂志他引 300 余次，H-因子为 25，21 篇论文引用次数超过 30 次，12 篇论文引用次数超过 50，4 篇论文引用次数超过 100，单篇最高引用次数为 359，参编专著 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、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、重点基金、面上项目、青年基金、数学天元基金、江苏省第十四批“六大人才高峰”高层次人才项目、中国矿业大学越崎学者人才引进项目等 16 项重要项目。担任八家国际生物信息学会议的程序委员会成员，担任《2012－2013 运筹学学科发展报告》编写组成员。

报告题目：大数据时代下基于网络算法和机器学习的生物信息学研究

罗洁 浙江省农业科学院，数字农业研究所/公共实验室，生物信息学研究室副主任;美国圣犹达儿童医院，访问交流;清华大学，生物学专业（生物信息学方向）;首都医科大学，生物医学工程专业，免试进入硕博连读

浙江省生物信息学学会理事（候补）。主持或参与省部级以上项目 6 项，发表 SCI 论文 9 篇，其中 1 篇（IF>5）。

报告题目：基于因果关系分析的特定表型相关调控网络研究

宋晓峰 南京航空航天大学自动化学院教授，主要研究领域为转录组学与计算系统生物学。近几年来曾参加过国家自然科学基金、省部级科研项目以及其它各类项目十多项，现主持国家自然科学基金两项，教育部博士点基金与江苏省自然科学基金各一项。在国内外权威学术期刊上发表过 80 多篇学术论文，其中在国际著名的 Nucleic Acids Research, Journal of Theoretical Biology, Journal of Medical Virology, BMC Genomics 等 SCI 检索期刊发表论文 26 篇，目前担任：江苏省生物医学工程学会理事；江苏省生物医学工程学会生物信息学专业委员会委员；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委员会委员；

报告题目：环形 RNA 编码蛋白潜能的生物信息学研究

吴世华 浙江大学生科院副教授，浙江大学思源天然药物与生物毒素研究中心天然产物实验室负责人。1994-2001 年昆明理工大学精细化工学士和应用化学硕士，2004 年浙江大学有机化学博士。2011-2012 年受浙江大学骨干教师计划资助在美国洛杉矶希望之城国家医学中心癌症生物学系访问。近年来致力于抗肿瘤天然产物（生物毒素）的提取分离、结构鉴定、作用机制及其靶向应用研究。自 2001 年开始研发逆流色谱以来，迄今已设计制造了多种类型的逆流色谱仪，并构建了一系列的应用方法。本报告将重点探讨本实验室建立和发展的一些天然产物的分离及模式识别方法和技术。

报告题目：天然产物分离及模式识别技术

甘卓慧 温州医科大学低氧研究所副教授。早年在浙江大学生物医学工程学院获得学士和硕士学位，赴美后在马凯特大学获得生物医学工程专业系统生物学方向博士学位，博士毕业后在加州大学圣迭戈分校生物工程学院进行系统生物学研究，主要研究方向是组学与生物信息学结合，研究缺氧对生物体影响及其调控机制。

报告题目：高通量基因表达数据分析工具的整合、流程自动化和跨物种分析比较

三、论文摘要

Bioinformatics Platform for Genomic Island Detection and Analysis

代琦

浙江理工大学生命科学学院

Abstract: Genomic islands (GIs) that are associated with microbial adaptations and carry sequence patterns different from that of the host are sporadically distributed among closely related species. This bias can dominate the signal of interest in GI detection. However, variations still exist among the segments of the host, although no uniform standard exists regarding the best methods of discriminating GIs from the rest of the genome in terms of compositional bias. In the present work, we proposed a robust software, MTGIpick, which used regions with pattern bias showing multiscale difference levels to identify GIs from the host. MTGIpick can identify GIs from a single genome without annotated information of genomes or prior knowledge from other datasets. We also constructed a genomic islands database and provided an on-line genomics island analysis platform.

肿瘤中 tRNA 来源小片段的基因组图谱

刘鹏渊

浙江大学转化医学研究院

摘要： 我们首次系统性地分析了 TCGA 数据集中的 15 种不同类型的癌症 8000 多个肿瘤样本里的一种新型的 tRNA 来源小片段 （tRFs）的综合图谱特征。我们的一系列的生物信息学证据表明 tRFs 是具有生物功能的一种非小编码 RNA 片段，这一点至今仍被严重忽视。我们发现 tRFs 在多个肿瘤类型中是普遍失调的。我们的结果也表明来自不同癌症类型的基于 tRFs 的分类亚型具有非常强的共性，所形成的涵盖不同肿瘤类型的 supercluster 的群体样本中某特定长度 tRFs 表达上升，肿瘤相关信号被激活。我们也发现通过整合已建立的临床决策分析和 tRFs 的分类亚型可以进一步提高肿瘤病人的预后预测效果。此外，利用我们发展的多维度整合分析策略，我们鉴别了 11 个高信度的驱动肿瘤发生发展的 tRFs。通过结合生物信息学和实验学方法首次证实了一个来源于 tRNA IleAAT 5′ 端的 20nt tRF 通过促进肿瘤细胞增殖，转移和侵袭以及促进细胞周期进展来驱动肿瘤发生和发展。我结果表明 tRFs 是一类具有调控功能的分子，需要进一步研究和探索并且极具潜力作为一种新的诊断和预后标记以及治疗靶点。

基于因果关系分析的特定表型相关调控网络研究

罗洁

浙江省农业科学院数字农业研究所

摘要：随着高通量测序技术的迅猛发展，我们可获得大量与表型（性状）相关的基因型。然而这些基因型与表型是否有因果关系，以及基因型如何通过基因调控网络最终影响表型，这些问题还不清楚。我们提出了一个“基因型->基因表达->表型”的因果驱动模型来连接基因型和表型，以此深入揭示基因型影响特定表型的表达调控网络。我们以 BXD 体系小鼠为例，研究其在酒精和紧张焦虑下的调控网络。（1）识别出 14 个共表达基因网络及其对应的表达模式。（2）进一步识别出 4 个共表达基因网络模块，涉及 Υ-氨基丁酸受体（GABA）信号通路、谷氨酸信号通路、神经肽信号通路及 cAMP 依赖的信号通路。（3）通过基因网络模块的特异性分析，识别出于酒精、紧张焦虑以及两者交互作用特异相关的基因模块。（4）根据因果关系模型，把基因型、基因表达和表型连接起来。以此揭示酒精和紧张焦虑下的调控网络。基于我们所构建的因果关系模型在小鼠上的成功应用，我们计划将此模型进一步应用在植物的相关研究中。

环形 RNA 编码蛋白潜能的生物信息学研究

宋晓峰

南京航空航天大学自动化学院

摘要：真核细胞在 DNA 转录后通过反向剪接机制形成的环形 RNA 具有重要生物学功能。已有研究表明部分环形 RNA 具有编码蛋白的能力，但调控其翻译活性的分子机制尚不清楚。有研究证实人工合成的含有“内部核糖体进入位点”（Internal Ribosome Entry Site，IRES）元件和开放阅读框的环形 RNA 可以在体外翻译产生蛋白质，有研究也发现部分内源性环形 RNA 翻译多肽分子。随着第二代测序技术、Ribo-seq、ChIRP 技术、蛋白质谱技术等各种高通量实验技术的发展，以及国际上多种环形 RNA 数据库的建立，我们能够从系统生物学与生物信息学角度，深入探讨环形 RNA 分子在其编码蛋白潜能上的序列与结构特征。

高通量基因表达数据分析工具的整合、流程自动化和跨物种分析比较

甘卓慧

温州医科大学基础医学院，温州，325005

摘要：高通量基因表达检测技术的发展、普及以及大量数据的公开为我们提供了进行基因表达数据深入挖掘研究的良好契机。然而，基因表达数据处理本身具有的复杂性以及对计算机技术的较高要求，限制了众多研究者对相关数据的利用挖掘。因此，我们面向微阵列数据和高通量测序数据，开发了包括数据下载、质控、预处理、数据对齐和归一化以及综合分析等步骤的自动化分析流程，并支持分析结果基于基因的跨检测平台、跨物种比较。流程整合了大量分析工具，自动化了分析过程中一系列操作，标准化了分析流程并大大降低了对分析人员的编程技术要求。运用此流程，我们对从 GEO 数据库遴选的缺氧刺激后小鼠、果蝇等物种的基因表达数据进行了分析，并筛选出了数个缺氧的保守反应基因。

Abstract：The rapid development and application of high-throughput gene detection technique and the large amount of public accessible gene expression data provides significant opportunities for the scientific community to conduct bioinformatics analyses within and across multiple datasets. However, the processing complexity of high-throughput data and the required programming skills prevent the data utility by the public. Thus, we constructed automatic workflows to facilitate and standardize the analyses of microarray and sequencing data which covers data downloading, quality control, data pre-processing, alignment and mapping, normalization, meta-analyses, as well as cross-species comparison for multiple datasets. Using the constructed workflows, we analyzed datasets acquired from GEO Database regarding gene responses to hypoxia in different species, and screened multiple conserved genes responding to hypoxia.

联系方式：gzh@wmu.edu.cn

基于基因表达数据和 MRI 数据的乳腺癌影像基因组学关联分析

明文龙，袁少勋，李海涛，孙啸*

生物电子学国家重点实验室，生物科学与医学工程学院，东南大学，江苏南京 210096

背景：影像基因组学是近几年伴随生物医学大数据爆发而兴起的前沿研究领域，旨在从以基因数据为代表的生物大数据和以影像数据为代表的医学大数据中，提取重大疾病的相关信息，进而转化成医学知识，并最终指导人类重大疾病的诊断和防治。乳腺癌是一类严重威胁女性健康的恶性肿瘤，影像基因组学的出现提供了新的诊断思路，即通过建立影像表型和基因活动的关联，实现乳腺癌的非侵入性诊断。因此，本研究目前关注于乳腺癌的基因表达数据和 MRI 数据的关联分析，寻找与影像表型特征存在强关联性的基因和基因活动。

数据与方法：通过在 TCGA 数据库和 TCIA 数据库中筛选，既具有癌症组织和癌旁组织基因表达（RNASeq）数据，又具有 MRI 数据的乳腺癌患者，最终获得 25 例样本。之后，基于感兴趣区域法（ROI）对乳腺癌 MRI 进行影像特征的提取；通过基因表达差异分析，获得乳腺癌表达差异基因，使用 Spearman 等级相关分析直接分析影像特征和表达差异基因间的相关性，利用基因集富集分析（GSEA）获得影像特征和肿瘤相关基因集间的关联性，最后使用加权基因共表达网络（WGCNA）挖掘影像特征和基因表达模块间的相关性。

结果：从乳腺癌 MRI 初步提取出 94 个影像特征，包括一阶统计特征、形状特征、纹理特征等。基于 RNASeq 数据的基因表达差异分析，筛选出 4446 个表达差异基因。计算表达差异基因与影像特征间的 Spearman 等级系数，筛选存在强相关性的基因-影像对（$|r|>0.7$），最终得到 36 对强相关性的基因-影像对。GSEA 结果显示，有 40 个影像特征存在基因集富集的情况（$FDR<0.25$），这些影像特征主要集中在一阶统计特征和纹理特征。WGCNA 将乳腺癌表达差异基因按照表达趋势分成了 23 个基因模块，分析结果显示 23 个基因模块与 94 个影像特征间存在广泛的相关性。

结论：通过乳腺癌影像基因组学的初步分析，可以看到影像表型特征与疾病潜在的基因活动存在密切的相关性，这为乳腺癌的临床非侵入性诊断提供了可能；同时也为之后，融入更多的组学数据、挖掘更深入的特征或模式提供了工作基础。

基金项目：江苏省重点研发计划（编号：BE2016002-3）

***通讯作者联系方式：**xsun@seu.edu.cn

微生物导电菌毛的进化、结构和电子传递机制研究

束传军，肖可，孙啸*

东南大学生物电子学国家重点实验室，南京，210096

背景：产电微生物可在外界电子受体不能进入细胞的情况下将呼吸链延伸到细胞外，将电子传递到胞外受体 [1]。产电微生物 *Geobacter* 菌属的导电菌毛在产生电能的同时，可以降解工业和生活废水中的有机物质，促进环境中毒性或放射性金属污染物质的还原，从而实现环境的生物修复，是目前研究的热点 [2]。*Geobacter sulfurreducens* 菌毛由于具有金属样的导电率，吸引了诸多学者。目前对导电菌毛的导电机制认识还非常有限，其在胞外电子转移过程中的作用尚有争议，所以本文主要关注导电菌毛蛋白的进化来源，结构，可能的胞外电子传递机制。

材料与方法：为了研究导电菌毛的特征和起源，我们分析了它们的序列、结构和系统发生树。为了构建纳米导线的结构，我们提出了一种预测和注释纳米导线结构的新方法。导电菌毛蛋白的同源基因是使用 PilFind 搜索的。序列特征和蛋白质三级结构分别用 MAFFT 和 QUARK 分析。通过构建系统发育树来探索导电菌毛蛋白的起源。我们发现截断型菌毛的同源蛋白和 GSU1497 同源蛋白的结构组成分别与完整型的 N 末端和 C 末端相似。截断型菌毛的同源蛋白和 GSU1497 同源蛋白可能通过基因分裂从完整型菌毛蛋白进化而来。

结果与结论：我们重构了淋球菌与脑膜炎奈瑟菌的菌毛结构，结果显示预测的结构与生物学解析的构象很接近（偏差在 2 埃左右）。因此，我们的方法可以成功的预测四型菌毛的结构。利用该方法，我们构建了 *Geobacter uraniireducens* 和 *Geobacter sulfurreducens*（GS）菌毛的三维结构。通过分析菌毛结构细节，发现来自不同 GS 亚基的 Phe1、Phe24 和 Tyr27 可以构成 pi-pi 堆叠现象。本研究表明菌毛具有特定的序列特征，结构特征和进化起源，为导电菌毛结构预测提供了新的方法。我们的研究结果有助于进一步揭示和理解沿 *Geobacter* 物种的导电菌毛进行远程电子传递的分子机制。

参考文献：

1. D. R. Lovely, *Annu Rev Microbiol*, 2012, **66**.
2. G. Reguera, K. D. McCarthy, T. Mehta, J. S. Nicoll, M. T. Tuominen and D. R. Lovley, *Nature*, 2005, **435**, 1098-1101.
3. D. R. Lovley, *Energ Environ Sci*, 2011, **4**, 4896-4906.
4. N. S. Malvankar, M. Vargas, K. P. Nevin, A. E. Franks, C. Leang, et al., *Nat Nanotechnol*, 2011, **6**, 573-579.

基金项目：国家自然科学基金（No.61472078），国家重点实验室重点研究基金；江苏省重点研发计划（BE2016002-3）

***通讯作者联系方式：**xsun@seu.edu.cn

基于组学数据的阿尔兹海默症亚型风险预测方法

李海涛 ^{1}，袁少勋 ^{1}，吴建盛 ^{2}，孙啸 ^{*1}

1.东南大学生物科学与医学工程学院，生物电子学国家重点实验室，南京 210096

2.南京邮电大学地理与生物信息学院，南京 210003

背景：阿尔兹海默症（Alzheimer disease, AD) 是一种神经系统退行性疾病，其主要特征是记忆和认知功能的渐进性丢失。目前仍缺乏有效的治疗手段来预防、终止或扭转阿尔兹海默症的发生。因此，在 AD 的前期症状——轻度认知障碍（mild cognitive impairment，MCI）——就对疾病的发展加以控制显得极为重要。最近，越来越多的研究发现 AD 具有异质性[1]，因此，该研究希望通过识别 MCI 患者组学数据的不同特征模式分别构建预测模型，对 MCI 患者在 5 年内是否发展为 AD 的风险进行预测。

材料与方法：从 ADNI 数据库中获取被诊断为 MCI 的 125 位患者的 SNP 和基因表达数据，使用相似融合网络算法[2]进行聚类分析：将两类组学数据矩阵分别构建的样本-样本矩阵进行反复迭代后融合为一个矩阵，并对融合矩阵进行聚类，从而对患者亚型分类。随后使用多核学习的变分贝叶斯分类算法（Variational Bayes Multiple Kernel Learning, VBpMKL)[3]作为分类器，对亚型划分的患者进行 5 年病情进展预测。

结果：通过实验发现，当样本被分为两类亚型时，MCI 发展为 AD 的时间差别最为显著，其中一组患者的病情发展较另外一组更为迅速（log rank test, $p < 0.005$）。从而证明基于组学数据对样本进行聚类能够达到很好的效果。并且将 MCI 患者聚类后，基于 VBpMKL 算法使用五折交叉验证分类器性能，预测结果为准确性 91.13%，敏感度 92%，特异性 89.8%，AUC=0.87，从而说明算法的有效性。

结论：根据组学数据特征将 MCI 进行分类，随后根据不同亚型特征构建 VBpMKL 模型进行病情发展的预测，能够显著提高预测准确性，从而帮助后续临床治疗，有效减缓疾病的进展，实现对每位患者的精准治疗。

参考文献：

[1] Qiang W, Yau WM, Lu JX, et al. Structural variation in amyloid-β fibrils from Alzheimer's disease clinical subtypes. Nature, 2017, 541 (7636): 217-221.

[2] Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 2014, 11(3): 333-337.

[3] Damoulas, T. and M. A. Girolami. Combining feature spaces for classification. Pattern Recognition 42.11(2009):2671-2683.

基金项目：江苏省重点研发计划（BE2016002-3）

*通讯作者联系方式：xsun@seu.edu.cn

In silico Classification of Bacterial Glycogen Branching Enzyme Based on N-terminal Domain Organization

Liang Wang[1*], Qinghua Liu[2], James Asenso[2], Michael J. Wise[3, 4*], Xiang Wu[1], Chao Ma[1,5], Yue Huang[1], Daoquan Tang[2, 6]

[1]School of Medical Informatics, Xuzhou Medical University, Xuzhou, 221000, China; [2]Jiangsu Key Laboratory of New Drug and Clinical Pharmacology, School of Pharmacology, Xuzhou Medical University, Xuzhou, 221000, China; [3]School of Computer Science and Software Engineering, University of Western Australia, Perth 6009, Australia; [4]The Marshall Centre for Infectious Diseases Research and Training, University of Western Australia, Perth 6009, Australia; [5]School of Biology and Food Technology, Guangdong University of Petrochemical Technology, Maoming 525000, China; [6]Center of Experimental Animals, Xuzhou Medical University, Xuzhou 221000, China.

Abstract: Glycogen is a water-soluble, highly branched, and homogeneous polysaccharide that is widespread in archaea, bacteria, fungi and animals. In bacteria, glycogen plays important roles in carbon and energy storage. It is also linked with bacterial environmental persistence and pathogenicity. Among the essential genes for bacterial glycogen metabolism, the *glgB*-encoded branching enzyme GBE plays an essential role in forming α-1,6-glycosidic branching points, and determines the unique branching patterns in glycogen. Currently, only the structure of full-length GBE from *Mycobacterium tuberculosis* H37Rv has been solved in bacteria. In contrast, GBE sequences have been extensively studied. Previously, N-terminal analyses based on small sets of GBEs revealed that two types of GBEs might exist: 1) type I GBE with both N1 and N2 (also known as CBM48) domains and 2) type II GBE with only the N2 domain. Several *in vitro* studies have linked N1 domain with transfer of short oligosaccharide chains during glycogen formation, which could lead to small and compact glycogen structures. Compact glycogen degrades more slowly and, as a result, may serve as a durable energy reserve, contributing to the enhanced environmental persistence for bacteria. We were therefore interested in classifying GBEs based on their N-terminal domain organizations via large-scale sequence analysis. In addition, we wanted to understand the evolutionary patterns of different GBEs through phylogenetic analysis. In this study, we initially analyzed N-terminal domains of 169 manually reviewed bacterial GBEs. A previously unreported group of GBEs with around 100 amino acids ahead of the N1 domains was identified. Phylogenetic analysis found clustered patterns of GBE types in certain bacterial phyla. A further study of 9,387 GBE sequences identified more GBEs that might belong to a novel group of GBE. This research aimed to investigate bacterial GBEs based on N-terminal domain organization and trace evolutionary relationships of GBEs among bacterial species. Our theoretical study revealed the limitations of arbitrary classification of GBEs based on N-terminal length. However, it might provide guidance for further experimental study of GBE N-terminal functions in bacterial glycogen structure and metabolism.

Keywords: Glycogen, Branching enzyme, N-terminus, N1, CBM48, Average chain length

***Correspondence author:** Dr. Liang Wang, Email: leonwang@xzhmu.edu.cn; Dr. Michael J. Wise, Email: michael.wise@uwa.edu.au

使用分子动力学模拟 *Geobacter sulfurreducens* 菌毛亚基 PilA 在不同环境中的构象变化

肖可，束传军，孙啸*

东南大学生物科学与医学工程学院，生物电子学国家重点实验室，南京 210096

Geobacter sulfurreducens 菌毛是一种在微生物表面生长的，具有“类似金属的”电导率的蛋白多聚体纤维，具有生态学和材料学等多方面的重要意义[1]。这些菌毛是由大量同一的亚基 PilA 组装成的，这些亚基的结构已通过核磁共振（NMR）在 DHPC 去污剂溶液中获得[2]，其主要结构包括 N 端由 50 个氨基酸组成的 α 螺旋和 C 端长度为 11 个氨基酸的无序卷曲。这其中，α 螺旋被命名为 α1，又可以细分为 α1-N 和 α1-C 两个子区域。尽管两亲的去污剂分子常被使用在膜蛋白的结构测定实验中，以模拟天然的双分子膜并稳定蛋白折叠，但是人们也证明了在去污剂环境中获得的蛋白构象是和去污剂胶束的大小以及厚度相关的，并可能和双分子膜上的构象存在显著差异[3], [4]。这样的影响和差异使得 PilA 的结构验证变得尤为重要。

为探究 PilA 在实验环境和真实内膜环境中的差异和细节，并更好地优化菌毛亚基的结构，我们使用分子动力学的方法模拟了三种环境下的菌毛亚基 PilA 的构象，分别是：作为参照组的完全的水环境，与实际组装环境类似的磷脂（POPE）双分子膜以及与 NMR 实验一致的去污剂（DHPC）胶束环境。模拟的结果表明，暴露在水环境中的亚基蛋白的 α1 螺旋由于疏水区域受到水分子的攻击而完全变形；插入双分子膜的亚基蛋白，其靠近 N 端的完全疏水的 α1-N 区域受到细胞膜的保护而保持完整，但接近 C 端的 α1-C 区域由于非极性氨基酸的非对称分布形成的输水面暴露在水中，导致了构象的改变；而位于 DHPC 胶束中的亚基蛋白的 α1 螺旋，包括 α1-N 和 α1-C 子区域都保持了完整的构象，与已发表的实验结果一致。这是由于 DHPC 分子相较于 POPE 而言拥有更短的烃尾并能够形成更加多变的形状，从而对双分子膜无法接触的 α1-C 疏水面形成有效的覆盖和保护。这些结果暗示在菌毛组装过程可能有其他未知蛋白与亚基相互作用以维持亚基蛋白的在 IV 型菌毛中保守的 α1 螺旋结构，并影响 *Geobacter sulfurreducens* 菌毛的电子传递功能。

参考文献：

[1] G. Reguera, K. D. McCarthy, T. Mehta, J. S. Nicoll, M. T. Tuominen, and D. R. Lovley, “Extracellular electron transfer via microbial nanowires,” *Nature*, 435(7045): 1098–101, 2005.

[2] P. N. Reardon and K. T. Mueller, “Structure of the type IVa major pilin from the electrically conductive bacterial nanowires of Geobacter sulfurreducens,” *J Biol Chem*, 288(41): 29260–6, 2013.

[3] L. Columbus *et al.*, “Mixing and matching detergents for membrane protein NMR structure determination,” *J Am Chem Soc*, 131(21): 7320–6, 2009.

[4] T. A. Cross, M. Sharma, M. Yi, and H.-X. Zhou, “Influence of solubilizing environments on membrane protein structures,” *Trends Biochem. Sci.*, 36(2): 117–125, 2011.

基金项目：国家自然科学基金 No. 61472078，东南大学生物电子学国家重点实验室重点研究基金

***通讯作者联系方式**：xsun@seu.edu.cn

基于表型聚类的 SNP 相互作用网络分析

袁少勋，李海涛，孙啸*

东南大学生物科学与医学工程学院，生物电子学国家重点实验室，南京 210096

背景：基因多效性是指一个基因位点影响两个或多个表型特征的特性[1]，是一因多效。遗传性缺失是指目前研究发现的致病基因仅能解释很小一部分表型变化的遗传规律[2]；研究基因相互作用是解决遗传性缺失的一种有效手段，是多因一效。影像遗传学是指利用医学影像数据中获取的定量表型数据评估遗传变异对疾病内在表型的影响。同时考虑基因多效性和遗传性缺失的影像遗传学研究可以较全面的评估多个遗传因素对多个相关表型的影响。

材料与方法：本研究采用的数据来源于阿尔兹海默病神经影像行动计划（ADNI）数据库中 174 个 AD 患者和 213 个健康对照（NC）的全基因组关联研究（GWAS）数据以及 35 种皮层下组织神经影像学定量表型（QT）数据。研究分为三部分，（1）对定量表型进行 Pearson 相关性聚类，将所有定量表型分为若干表型组；（2）进行基于定量表型的数量性状 GWAS 关联分析，状挖掘表型相关的显著 SNP；（3）基于信息增益理论，构建每一表型组中所有显著 SNP 相互作用网络，并通过去掉某一节点后网络信息增益值的下降程度的大小评估相互作用网络中关键 SNP 的重要性。

结果：依据表型相关性，影像特征可明显聚为五个表型组，分别为脑室部分、基底节部分、内侧颞叶部分、胼胝体部分、间脑部分；每一个表型组内显著相关 SNP 位点相互作用网络关键节点分别为 rs734477、rs1062098、rs7181251、rs3755809、rs10545584，关键节点去除后的相互作用网络信息增益值分别下降 6.52%、12.2%、8.26%、14.66%、4.3%。

结论：通过基于表型相关和 SNP 相互作用网络的影像遗传学的研究，可以综合考虑基因多效性和遗传性缺失的问题，能够较系统的研究表型变化和遗传变异之间的关系。

参考文献：

[1]. Paaby, A.B. and M.V. Rockman, The many faces of pleiotropy. Trends Genet, 2013. 29(2): p. 66-73.

[2]. Manolio, T.A., et al., Finding the missing heritability of complex diseases. Nature, 2009. 461(7265): p. 747-53.

基金项目：江苏省重点研发计划（BE2016002-3）

***通讯作者联系方式：**xsun@seu.edu.cn

基于 R shiny 框架的单细胞数据分析平台

薛继统，周银聪，陈铭*

浙江大学生命科学学院生物信息学系，杭州 310058

摘要：流式细胞仪分析技术在细胞生物学、生物医学等多个领域的研究过程中发挥着至关重要的作用。随着 CyTOF 技术以及单细胞测序技术的发展，高维度，大规模的数据分析方式成为了当下对细胞数据分析的一种硬性需求，因此快速高效地自动化细胞分群技术应时代需求脱颖而出，逐步取代手动“设门”技术的地位，被越来越多的科学家所接受。即便如此，自动分群在一些领域仍然无法替代手动“设门”的方法，比如确认样本中包含的细胞类型，是否存在有特殊的细胞群体等。面对大规模数据分析的需要，自动化地实现对分得群落的细胞类型进行判别也将成为一种需求。另外对免疫学家而言，目前在对数据进行分析时还是很少利用这些自动化工具来进行分析，这里可能的一个因素就是多数这些工具在使用上对编程有一定的需求，而 R 语言作为流式细胞分析常用环境，大量的自动化预处理和分群后分析工具也是依赖于该环境，因此基于 R 环境开发一个用户友好的 GUI 分析平台是一种面向未来的流式细胞分析需求。

我们研究设计了一种实现自动化识别群落细胞类型的工具 PhenoCL。该工具利用 DMT（Divisive Marker Tree）来评估并获得各个群在分群过程中起到重要作用的 marker，用该群落 marker 表达结果去 CL（Cell Ontology）中查询，从而预测可能的群落细胞类型。同时，我们还搭建了一个基于 R 语言 shiny 包的全过程可视化可交互的流式细胞分析平台 CytoSEE。该分析平台全过程都支持交互操作，使用者可以获得更好地用户体验，并时刻掌握分析过程中产生的各个数据动态变化情况。该平台在整合了许多目前被证明最为好用的部分分群方法例如 flowSOM, densitycut 等的同时还自主设计开发了 consboost，这是一种利用 adaboost 机器学习手段和 consensus clustering 方法实现的高精度大规模 CyTOF 细胞数据分群方法。该平台的出现将为高效快捷地实现流式细胞分析全面自动化过程提供新动力。

关键词：CytoSEE；细胞类型预测；Consboost；细胞分群；PhenoCL

***通讯作者联系方式：**mchen@zju.edu.cn

水稻发育过程中的表达及代谢水平分析

李梦婷，孟宪文，陈铭*

浙江大学生命科学学院生物信息学系，杭州 310058

摘要：水稻整个生长周期可以分为三个时期：幼苗阶段、营养阶段和生殖生长阶段。以往的研究缺少对水稻全周期的整合分析，目前整个过程的表达调控变化还尚不清楚。为了探究水稻整个生长过程的生长代谢模式，我们选取水稻发育三个阶段的高通量 RNA-Seq 数据，不仅分析了线性 mRNA 的表达，也挖掘和分析了长非编码 RNA，并进一步构建了它们之间的共表达网络。通过基因表达模式、代谢调控和功能富集分析，我们发现了一些与水稻发育相关的 mRNA 及 lncRNA，我们的研究发现 lncRNA 在水稻发育过程呈动态表达，意味着这类非编码 RNA 可能对水稻发育起作用。综上，本研究的发现为今后的水稻生长发育研究提供了一定的基础。

关键词：水稻生长周期；长非编码 RNA；共表达网络；基因表达模式；代谢调控

***通讯作者联系方式：**mchen@zju.edu.cn

DeepRRC：Identification of Residue–Residue Contacts Using a Deep Learning-based Two-dimensional Scheme

Ming Wen[1], Peisheng Cong[2], Zhimin Zhang[1], Hongmei Lu[1, *] and Tonghua Li[2, *]

[1]College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China, [2]School of Chemical Science and Engineering, Tongji University, Shanghai 200092, China.

Abstracts: Identifying residue–residue contacts in proteins is a two-dimensional problem and a critical step in the prediction of tertiary protein structures. Many prediction methods based on machine learning have been developed, but their prediction accuracy is still lower than required. Most of these prediction approaches are similar to one-dimensional schemes and they are not true two-dimensional schemes. A major problem when identifying residue–residue contacts is the lack of a useful feature among those explored previously. Thus, new schemes need to employ innovative approaches for residue contact prediction.

In this study, we reported the design and implementation of DeepRRC, a deep learning-based two-dimensional scheme approach for accurately identifying protein residue–residue contacts. First, a deep learning-based classifier (DLC) was built on a collection of widely used features. Next, anoveltwo-dimensional contact profile (TDCP) of each residue pair was obtained via multiple sequence alignment of the query against the template library. A XGBoost classifier was built based on the TDCP and the prediction probability of DLC. Comparison with other state-of-the-art existing tools indicated that DeepRRC improved overall predictive performance. DeepRRC is freely available at https://github.com/Bjoux2/DeepRRC

Keywords: residue-residue contacts, deep learning, two-dimensional sheme

***Correspondence author:** Tonghua Li: lith@tongji.edu.cn, Hongmei Lu: hongmeilu@csu.edu.cn

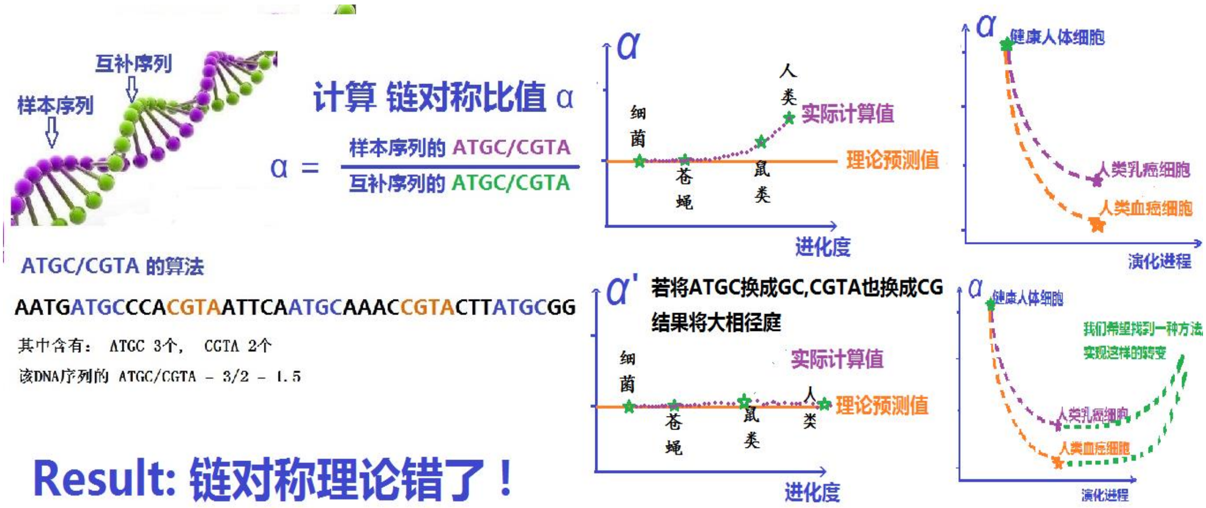
Both Chargaff Second Parity Rule and the Strand Symmetry Rule Are Imprecise DNA 双链的物理&化学：属性将不再对称相似

Zhiyu Chen

Peiyou Education School, Shanghai, China

Abstract: In order to check Chargaff Second Parity Rule, we find the strands are asymmetric in human DNA, this breaks the strand symmetry rule. We calculate the ratio between oligonucleotide ATGC and oligonucleotide CGTA, and we compare the sample sequence average ratio ATGC/CGTA and the complementary sequence average ratio ATGC/CGTA. we find evolution degree bigger, then the strand symmetry deviation will be bigger. sequence and its complementary strand sequence obviously have two different characters, include physical property, chemical property and biological property. It is very important, based on this asymmetry, we can find some new and special theories in biology to explain how chromosome communicates and works in the future. we also find, both leukemia and breast cancer are weakening the DNA's asymmetry degree. Here need more research and check, maybe we can find an easy diagnosing method to leukemia and breast cancer, if this result here is right at last, it will benefit to the world.

Keywords: ATGC/CGTA, Strand Asymmetry, Complementary Sequence, Evolutionary Forces



$$\alpha = \frac{\text{样本序列的 ATGC/CGTA}}{\text{互补序列的 ATGC/CGTA}}$$

ATGC/CGTA 的算法

AATGATGCCCACGTAATTCAATGCAAACCGTACTTATGCGG

其中含有： ATGC 3个， CGTA 2个

该DNA序列的 ATGC/CGTA – 3/2 – 1.5

若将ATGC换成GC,CGTA也换成CG
结果将大相径庭

Result: 链对称理论错了！