DOI:10.16262/j.cnki.1000-8217.2012.02.017

·学科进展与展望·

生物信息学重大基础科学问题及关键技术
——第51期“双清论坛”综述

赵 屹[1] 谷瑞升[2] 杜生明[2]

(1 中国科学院计算技术研究所，北京 100190；2 国家自然科学基金委员会生命科学部，北京 100085)

[摘 要] 第51期“双清论坛”围绕“生物信息学重大基础科学问题及关键技术”进行了深入研讨，会议结合国内外研究现状和国内的研究基础及优势，提出了未来我国生物信息学应当关注的重要科学问题。

[关键词] 生物信息学，基础科学问题，关键技术，双清论坛

国家自然科学基金委员会(以下简称“自然科学基金委”)第51期“双清论坛”于2010年11月8—10日在湖北省武汉市召开。本次论坛由华中科技大学理学院承办，论坛主题为：生物信息学重大基础科学问题及关键技术。中国科学院生物物理所陈润生院士担任论坛主席，中国科学院基因组研究所吴仲义院士、中国科学院上海马普学会计算生物学研究所韩敬东研究员为论坛副主席。

来自近30个大学、科研院所的50多位数学、物理、信息和生物学等领域的专家、学者出席了此次论坛。

1 生物信息学简介

生物信息学是伴随着人类基因组计划发展而产生的一门涉及生物学、数学以及计算机科学的交叉学科。关于生物信息学的定义，20世纪90年代，美国人类基因组计划曾给出一个比较完整的解释：生物信息学是一门交叉学科，包含了生物信息的获取、加工、存储、分配、分析、解释等在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具来阐明和理解大量数据所包含的生物学意义。生物信息学旨在揭示“基因组信息结构的复杂性及遗传语言的根本规律”，是本世纪自然科学和技术科学领域中“基因组”、“信息结构”和“复杂性”这3个重大科学问题的有机结合。

迄今为止，生物信息学的发展主要经历了3个阶段：最初是前基因组时代，其研究方向主要集中在生物学数据库的构建，检索工具的开发和应用以及对DNA和蛋白质序列的比对和分析；第二阶段，即基因组时代，这一阶段工作主要集中在对核苷酸序列的测定、分析以及发现新基因，还包括基于网络和交互界面的大量数据库的开发和应用以及对基因组序列信息的提取分析等；随着2006年人类基因组图谱的绘制完成，生物信息学的发展进入当前的后基因组时代，这个时期主要的工作包括蛋白组学的研究、表观组学、疾病表型组学及人类基因组注释等的研究。在后基因组时代，如何快速准确地获取生物体的遗传信息仍然是生命科学的研究中一个重要问题。基因组包含了一个生物体的全部遗传信息，是全面揭示物种复杂性与多样性的源泉。遗传信息的获取是进行基因组、转录组、代谢组、蛋白组、基因表型组、表观遗传组等研究的基础，随着二代测序技术的发展，基因组、转录组的遗传信息能够被准确地检测，从而为生物信息学的研究提供了丰富的资源。因此测序技术的快速发展极大地推动了生物信息学的研究进展。

2 生物信息学研究现状及发展趋势

2.1 国外研究现状及发展趋势

生物信息学在国外发展得非常迅速，自1988年美国成立了国家生物技术信息中心(NCBI)之后，欧洲和日本分别于1993年3月和1995年4月建立了

本文于2011年7月28日收到.

欧洲生物信息学研究所(EBI)和信息生物学中心(CIB)。此外,各种生命科学领域相关的研究机构以及各种制药及生物科技相关的企业也都成立了生物信息部门,各种生物信息相关的公司更是如雨后春笋般地涌现。目前,可以说绝大部分核酸和蛋白质数据库都由美国、欧洲和日本进行维持。他们成立了GenBank/EMBL/DDBJ国际核酸序列数据库,并且每天进行数据交换,同步更新,确保用户在任何一家数据库所得的信息都是最完整最准确的。除了这3大核酸数据库外,还有很多特殊类型的核酸序列数据库,如人类基因组数据库(HGD)、非编码RNA数据库(ncRNA)、表达序列标签数据库(dbEST)、序列标签位点数据库(dbSTS)、核苷酸三维结构数据库(NDB)、人类基因变异数据库(HMGD)等,进一步更细类别的核苷酸数据库如miRBase、tRNAdb等;另外还有蛋白质序列相关的数据库,如蛋白质信息库(PIR)、Munich information center for protein sequences(MIPS)以及Pfam数据库等。

目前国际上最大的集研究、开发和服务为一体的分子生物信息机构:欧洲分子生物学网络组织(European Molecular Biology Network, EMBNet),通过计算机网络实现英国、德国、法国以及瑞士等国家生物信息资源的共享。在共享网络资源的同时,国家自身的生物信息也不断地得到发展,比如建立生物信息学研究机构、搭建各种具有专业特色的一级、二级或更高级的数据库、研发新的生物信息分析技术和方法,开放于全世界,为本国及其他国家的生物(医学)研究的发展起到了重要的推动作用。

从数据分析技术的角度来讲,有几个比较有代表性的例子,早在1962年,Zuckerkandl和Pauling将序列变异分析联系到演化关系,开创了分子演化的研究领域;1964年,蛋白质结构预测的研究由Davies的工作开始;1970年,Needleman和Wunsch发表了两序列比对算法的文章,期刊*Computer Methods and Programs in Biomedicine*诞生;1974年,Ratner首先对分子遗传调控系统进行理论处理;继第一批小RNA(tRNA)序列发表之后,1975年,Pipas和McMahon首先提出运用计算机技术预测RNA二级结构;伴随着1976年之后大量生物学数据分析技术的涌现,*Science*早在1980年第209卷就发表了关于计算分子生物学的综述;1990年,NCBI开发的序列相似性搜索程序blast方便了基因和基因特征的识别;2000年6月26日,被誉为生命"阿波罗计划"的人类基因组计划,美、英、日、法、德、中六国科学家经过艰苦努力,在全球同一时间宣布完成人类生命的蓝图,这是人类科学史上又一个里程碑式的事件。近些年来,伴随着3代测序技术的到来,测序数据的通量呈指数级地增长,传统的分析方法已无法满足高通量数据分析的要求,因此,迫切需要云计算等新的计算技术应用到生物信息学领域中来。

从专业出版物来看,起初生物信息并没有专业领域的期刊,这些文献都分散在其他领域的期刊中;到了1970年,出现了*Computer Methods and Programs in Biomedicine*生物信息相关期刊;到1985年4月,才有了第一种生物信息学专业期刊—*Computer Application in the Biosciences*;现在,生物信息学相关的专业期刊数目甚多,其中主要包括印刷版期刊和网络版期刊两种,如*Acta Biotheoretica*, *Bio Informatics Technology & Systems*, *Bioinform Newsletter*, *Briefings in Bioinformatics*, *Journal of Computational Biology*, *Genome Biology*, *Genome Research*, *OMICS*, *PLoS Computational Biology*, *Bioinformatics*, *Bioinformation*以及*BMC Bioinformatics*等。

从网络资源来看,生物信息学网站非常多,大的网站有国家级研究机构的,如NCBI, UCSC, Pubmed等;小的有专业实验室的,比如某个工具的在线服务。大型网站一般提供生物信息学相关新闻、数据库服务和在线服务,小型科研机构的网站以介绍自己的研究成果为主,有的免费提供自己机构设计的算法的在线服务。总之,基本都是面向生物信息学专业人士,各种分析方法虽然很全面,却分散在不同的网站,我们可以根据研究的需要来合理地做出选择。

从测序技术上来看,测序技术也在不断地更新换代,从1977年桑格等发明的双脱氧测序法到现在的第2代和第3代测序,技术不断的革新,测序通量也大幅度攀升,而单位碱基的价格也是急剧下降。相对于传统测序的96道毛细管测序,高通量测序一次实验可以读取40万到400万条序列。读取长度根据平台不同从25—450 bp,不同的测序平台在一次实验中,可以读取1—14 G不等的碱基数,这样庞大的测序能力是传统测序仪所不能比拟的。其中具有代表性的产品一个是罗氏公司(Roche)的454测序仪(Roch GS FLX sequencer),另一个就是2006年美国Illumina公司推出的Solexa基因组分析平

台(Genome Analyzer platform)，以及 2007 年 ABI 公司推出了自主研发的 SOLiD 测序仪(ABI SOLiD sequencer)。近期出现的 Helicos 公司的 Heliscope 单分子测序仪、Pacific Biosciences 公司的 SMART 技术和 Oxford Nanopore Technologies 公司研究的纳米孔单分子技术，被认为是第 3 代测序技术。与前两代技术相比，他们最大的特点是单分子测序。第 3 代测序通量，以 Pacific BioSciences 公司推出的 SMART 机器测序能力为例，它平均 15 分钟能够读取 4000 GB 的碱基。目前测试人体整个基因组的价格只需要 3000 美金。可以想象，不久的将来，我们的个人基因组会像验血报告一样成为医生治疗病人的参考数据。

生物信息学(Bioinformatics)自上个世纪末诞生以来，发展非常迅猛，在数据上，提供共享数据库平台，供全球共享；在测序技术和分析技术上不断革新，为基因组、转录组、蛋白组等组学研究的深入提供了更多的数据支持，从而不断地促进生命科学向前飞速发展。

2.2　国内研究现状及发展趋势

近几年来，国内对生物信息学的研究方向涉及到基因组、转录组、蛋白组、疾病表型组、表观遗传组及进化组等；同时，结合网络模型，又从中延伸出对各种组学中组件间的网络进行研究，比如蛋白相互作用网络，转录调控网络，以及双色网络等。随着第 2 代、第 3 代测序的发展，大量数据的产生，又进一步深化了对各种组学的研究和认识。

在基因组研究方面，测序与拼接[1,2]的物种低到微生物、微生物群落，高到动植物的全基因组，甚至开始对多倍体小麦进行测序和拼接；测序和拼接技术的发展对于基因的注释、重注释[3]以及对基因差异表达分析[4]提供了足够多的数据；基因的重注释及功能的研究对于基因相互作用的研究提供了更多的证据[5]。

对转录组的研究其中很大一部分是对非编码 RNA 的研究，基因组中的非编码序列是可以表达的，其表达产物是非编码 RNA，与其相应的基因就称为非编码基因。非编码 RNA 按其大小可分为短于 50 nt 的小分子非编码 RNA，比如 microRNA[6-9]，长于 200 nt 则称为长非编码 RNA。越来越多的事实证明非编码 RNA 具有重要的生物功能。非编码 RNA 种类和功能的多样性越来越引起研究人员的关注。非编码 RNA 通过不同的模式发挥功能，如调节转录模式，调节蛋白活力，改变 RNA 的加工方式等。非编码 RNA 的发现和功能研究使研究者从更高的层次对基因组的功能元件以及组织的机制进行认识。非编码 RNA 的结构及其预测[10,11]对于认识 RNA 的功能也起到至关重要的作用。以 Illumina 公司的 Solexa 测序仪为代表的高通量、短序列测序技术可以在一次测序反应中获得数百万甚至上千万小分子 RNA 序列，极大地推动了小分子非编码 RNA 的发现和功能研究。然而，在这数百万小分子 RNA 中，我们仅可以对其中的 30%甚至更少进行分类，而对绝大多数小分子 RNA 的种类和功能还一无所知。而且在基因预测以及蛋白质结构等分析方面建立起来的系列算法和软件，多不适用于非编码 RNA 的研究，这一方面增加了非编码 RNA 的研究难度，另一方面也给我们中国科学家提供了一个绝好的机遇。面对高通量产出的数据，通过分析发现新的非编码 RNA，开发预测新的非编码 RNA 的算法及软件包。

对蛋白组学的研究包括蛋白质结构预测[12-14]、基于结构预测蛋白质功能以及蛋白质结构异常与疾病等。对疾病表型组的研究有复杂疾病[15,16]、复杂疾病与计算系统生物学[17]、计算医学、个人医学等。表观遗传组包括磷酸化修饰[18,19]、组蛋白甲基化修饰[20]等。其中进化组中有对 microRNA 进化的研究[21]、物种亲缘关系的研究[22]等。

在功能基因组研究的过程中，通过基因芯片、蛋白质组学技术以及 RNA-seq 等技术，研究特定组织和特定时期的基因表达，从而进一步根据这些资料来发展相应的算法进行分析，得到复杂的生物网络。如：蛋白质-蛋白质相互作用的网络、基因表达调控网络、代谢作用网络以及信号转导网络等。而未来的生物学的网络调控机制应当是由蛋白质和非编码 RNA 两类元件构成的，因此将双色网络的概念运用到生物网络的分析当中，通过整合非编码 RNA 以及蛋白质相关的网络调控信息，可对复杂的生物网络有更深入的认识。

国内的研究几乎涉及到生物信息学的各个方面，可见随着对基因结构以及基因功能在整体水平认识的深入，生物信息学的研究已经历了多元化的阶段。

3　"生物信息学重大基础科学问题及关键技术"双清论坛具体内容和建议

3.1　论坛内容

此次论坛共安排了 9 个大会特邀报告和 22 个大会邀请报告。其中 9 个大会特邀报告包括：中国

科学院生物物理所陈润生院士的“非编码研究与生物信息学”；中国科学院基因组研究所吴仲义院士的“microRNA 的进化及其调控的双重功能”；上海生物信息技术研究中心李亦学教授的“Protein phosphorylation plays an essential role in the evolution of vertebrates”；中国科学院基因组研究所王俊研究员的“Sequencing，Sequencing，and Sequencing”；清华大学张学工教授的“对转录调控的一些尝试研究”；上海生物信息技术研究中心陈洛南研究员的“计算系统生物学与复杂疾病的研究”；中国科学院上海马普学会计算生物学研究所韩敬东研究员的“Integrative analysis of transcriptome and regulatory networks”；北京大学生物信息中心魏丽萍教授的“When bioinformatics met personalized medicine”；中山大学生命科学学院松阳洲教授的“Telomere signaling networks in human cells”。22 个大会邀请报告包括：华中科技大学物理学院肖奕教授的“RNA 三级结构预测和动力学研究”；中国科学技术大学生命科学学院刘海燕教授的“生物分子相互作用的分析、模拟与设计”；武汉大学物理科学与技术学院张文炳教授的“RNA 动力学与 RNA 调控”；南京大学生命科学学院王进教授的“microRNA 在基因转录调控网络中的行为”；中国科学院上海马普学会计算生物学研究所 Philipp Khaitovich 研究员的“Systems biology of aging”；中国科学院遗传与发育生物学研究所王秀杰研究员的“非编码 RNA 研究探索”；中国科学院生物物理研究所蒋太交研究员的“后基因组时代的计算结构生物学平台的发展”；同济大学生命科学与技术学院张勇教授的“表观遗传组动态变化与转录调控”；华中科技大学系统生物学系薛宇教授的“蛋白质共价修饰的生物信息学研究：进展与前瞻”；北京工业大学生命科学与生物工程学院王存新教授的“蛋白质折叠、分子对接及其应用问题研究”；北京大学理论生物中心邓明华教授的“Analysis of genetic interaction from EMAP data”；湘潭大学计算机系喻祖国教授的“物种亲缘分析的一种距离方法及生物网络的分形分析”；西安电子科技大学计算机系高琳教授的“复杂网络模块挖掘算法”；北京师范大学生命科学学院林魁教授的“利用 RNA-Seq 技术改进植物基因组注释质量初探”；中国科学院上海马普学会计算生物学研究所朱新广研究员的“系统生物学和 C4 水稻”；哈尔滨医科大学生物信息科学与技术学院李霞教授的“复杂疾病与生物信息学”；清华大学生命科学学院孙之荣教授的“生物分子网络与通路等系统的功能研究”；中国科学院数学与系统科学研究院应用数学研究所巩馥洲研究员的“数学与生物医学交叉应用研究的若干思考”；电子科技大学生命学院郭政教授的“基于高通量技术发现疾病标志的可重复性与癌相关分子改变的功能一致性”；上海交通大学系统生物医学研究院敖平教授的“Two Important Layers of Large Bio-Network Modeling：kinetics of whole metabolic networks and stochastic dynamics of endogenous networks for complex diseases”；中国科学院计算技术研究所赵屹副研究员的“计算医学：面向重大疾病诊治及核酸药物设计的生物信息技术研究”；华中科技大学生命科学与技术学院周艳红教授的“复杂疾病的数据整合和调控网络研究”。

会上专家提问和讨论积极踊跃，气氛热烈，不同学术观点和不同学术视角得到很好的碰撞和交融。报告和交流结束后，专家们结合我国研究特点和现状，就基础性、战略性、前瞻性的科学问题展开了认真、热烈的讨论。

3.2 论坛建议

本次论坛研讨了国内外关于“生物信息学重大基础科学问题及关键技术”的研究现状及未来发展趋势，结合国内的研究基础及优势提出了未来我国生物信息学应当关注的重要科学问题和发展的关键技术，它们是：(1) 基因组组装结构和比较演化的方法研究；(2) 细胞进化的溯祖理论；(3) 复杂系统及性状建模和网络结构与状态分析；(4) 非编码 RNA 系统发现及功能研究中的生物信息学理论和方法；(5) 蛋白质和 RNA 结构修饰与功能预测的方法；(6) 多层次数据整合的理论和方法；(7) 生物信息学特有数据库的构建与应用；(8) 生物信息学新理论、新技术和新方法。

会议建议自然科学基金委加大对“生物信息学”领域的扶持力度，积极稳步地培养国内该领域的研究团队，促进数学、物理、信息和生物学等多学科的相互融合与渗透，推动我国的科学家在生物信息学领域取得原创性研究成果，不断扩大中国生物信息学在国际上的影响力。

致谢 感谢中国科学院生物物理研究所的陈润生院士和华中科技大学的肖奕教授对本文的审阅以及提出的宝贵修改意见。

参 考 文 献

[1] Qin J, Wang J. A human gut microbial gene catalog established by deep metagenomic sequencing. Nature,2010,464: 59—65.

[2] Li R, Wang J. The sequence and de novo assembly of the giant panda genome. Nature,2010,463: 311—317.

[3] Luo Y Q, Fu C, Zhang D Y et al. Overlapping genes as rare genomic markers: the γ-Proteobacteria phylogeny as a case study. Trends Genet,2006,22: 593—596.

[4] Wang L, Feng Z, Wang X et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data, Bioinformatics, 2010,26(1): 136—138.

[5] Zheng H, Hang X, Zhu J et al. REMAS: a new regression model to identify alternative splicing events from exon array data. BMC Bioinform, 2009,10 (Suppl 1): S18.

[6] Zhao Y, He S, Liu C et al. MicroRNA regulation of messenger-like noncoding RNAs: a network of mutual microRNA control. Trends Genet,2008,24: 323—327.

[7] Luo R, Liao S, Tao G et al. *Dynamic analysis of optimality* in myocardial metabolic network under normal and ischemic conditions. Mol Systems Biol, 2006,2: 2006.0031.

[8] Liu L, Luo G Z, Yang W et al. Activation of the imprinted Dlk1-Dio3 region correlates with pluripotency levels of mouse stem cells. J Biol Chem,2010,285: 19483—19490.

[9] Wang Q, Huang Z, Xue H et al. MicroRNA miR-24 inhibits erythropoiesis by targeting activin type I receptor ALK4. Blood, 2008, 111: 588—595.

[10] Zhao P N, Zhang W B, Chen S J. Predicting secondary structural folding kinetics for nucleic acids. Biophys J,2010, 98:1617—1625.

[11] Jiang X, Chen C, Xiao Y. Improvements of network approach for analysis of the folding free-energy surface of peptides and proteins. J Comput Chem,2010,31: 2502—2509.

[12] Zhang H, Chen J, Wang Y et al. A computationally guided protein screen uncovers coiled-coil interactions involved in vesicular transport. J Mol Biol, doi: 2009,10.1016/j.jmb.2009.07.006.

[13] Gong X Q, Wang P W, Yang F et al. Protein-protein docking with binding site patch prediction and network-based terms enhanced combinatorial scoring. Proteins, 2010, 78: 3150—3155.

[14] Duan M, Huang M, Ma C et al. Position-specific residue preference features around the ends of helices and strands and a novel strategy for the prediction of secondary structures. Protein Sci,2008, 17: 1505—1512.

[15] Xu J, Li C, Li Y et al. MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. Nucleic Acids Res, Oct 6, 2010,[Epub ahead of print]

[16] Zhu J, Xiao H, Shen X et al. Viewing cancer genes from co-evolving gene modules. Bioinformatics, 2010, 26 (7): 919—924.

[17] Wang J, Zhang S, Wang Y et al. Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. PLoS Comput Biol, 2009,5(9): e1 000 521.

[18] Wang Z, Ding G, Geistlinger L et al. Evolution of protein phosphorylation for distinct functional modules in vertebrate genomes. Mol Biol Evol,2010 (in print)

[19] Ren J, Jiang C, Gao X et al. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. Mol Cell Proteomics,2009,9: 623—634.

[20] Yu H, Zhu S, Zhou B et al. Inferring causal relationships among histone methylations and gene expression. Genome Res,2008,18: 1314—1324.

[21] Lu J, Shen Y, Wu Q et al. The birth and death of microRNA genes in *Drosophila*. Nat Genet,2008,40: 351—355.

[22] Yu Z G, Chu K H, Li C P et al. Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model. BMC Evol Biol,2010,10: 192.

REVIEW OF THE 51ST 'SHUANGQING' FORUM ON "THE BASIC THEORIES AND KEY TECHNOLOGIES OF BIOINFORMATICS"

Zhao Yi[1]　Gu Ruisheng[2]　Du Shengming[2]

(1 *The Institute of Computing Technology, Chinese Academy of Science, Beijing* 100190;
2 *The Department of Life Science, National Natural Science Foundation of China, Beijing* 100085)

Abstract The 51st 'Shuangqing' forum—"the basic theories and key technologies of bioinformatics" sponsored by National Natural Science Foundation of China was held in Wuhan, Hubei Province in November 8—10, 2010. More than 50 experts and scholars from over 30 universities and research institutes and in the research fields of mathematics, physics, informatics and biology attended this forum to exchange their academic results and ideas. They had thorough discussion regarding the basic theory and key technologies of Bioinformatics. In combination with the international research frontiers and domestic statues and advantages, some prior research fields and suggestions that should be considered in further research were also proposed.

Key words bioinformatics, basic theories, key technologies, 'Shuangqing' forum