中国科学 C辑:生命科学 2008 年 第 38 卷 第 10 期: 882 ~ 890
www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS
专题

中国基因组生物信息学回顾与展望

顾坚磊[①②], 周雁[①②*]

① 复旦大学生命科学学院, 上海 200433;
② 国家人类基因组南方研究中心, 上海市-科技部共建疾病与健康基因组学重点实验室, 上海 201203
* 联系人, E-mail: zhouy@chgc.sh.cn

收稿日期: 2008-10-03; 接受日期: 2008-10-08
国家高技术研究发展计划(批准号: 2006AA02Z335)和上海市科学技术委员会(批准号: 07DZ22915)资助项目

摘要 基因组生物信息学是随着大规模基因组测序而兴起的一门交叉学科, 其研究对象是基因组数据. 因此, 对基因组数据的各方面的深入了解, 有助于把握基因组生物信息学的来龙去脉. 本文从基因组数据展开, 围绕着这些数据的收集、储存、分析和比较, 分别列举了相关的数据库、算法和软件包. 今天, 由新一代测序技术所带来的对数据处理、算法设计和功能信息挖掘等技术和研究方面的挑战, 应该及时得到充分的重视.

关键词
基因组学
生物信息学
新一代测序技术

1 简介

基因组生物信息学是随着大规模基因组测序而兴起的一门交叉学科. 其主要研究内容是利用计算机系统收集、储存、分析和比较各种基因组信息, 包括开发和建立一系列相关的数据库、算法和软件包. 生物信息学一方面运用计算机和数理学科的知识和技术服务于生物学数据研究, 另一方面为传统的数理研究开辟了新的研究方向, 这一“桥梁”作用使得生物信息学像其他交叉学科一样, “左右逢源”, 在近 10 年中发展特别迅速.

1.1 基因组生物信息学产生的背景

基因组生物信息学的研究对象是基因组数据. 对基因组数据各方面的深入了解, 有助于把握基因组生物信息学的来龙去脉.

DNA测序工作起始于 20 世纪 70 年代中期[1]. 在 1990 年正式启动人类基因组计划之后, DNA测序工作由量变到质变, 出现了一系列革命性的发展. 当该计划于 2003 年全面完成后, 生物学家并没有局限于人类的基因组计划, 而是利用人类基因组计划所解决的一些诸如物理图谱构建、高通量序列测定、序列拼接等关键技术, 运用于其他生物的基因组序列的研究计划. 至今, 已经有较完整的全基因组序列数据的物种包括 39 种类病毒、2,115 种病毒 、58 种古菌、1,269 种细菌(包括 51 种蓝藻)、69 种真菌、29 种原生生物、10 种植物和 78 种动物(图 1). 在这 3720 个物种中, 病毒由于基因组很小, 积累了大量物种的基因组数据. 原核生物因为其基因组紧凑测序量相对较小, 已经测序的物种占非病毒物种数的 8 成. 对原生生物和真菌的基因组测序工作由于基因组比较庞大, 除了酵母以外主要集中于和人类关系较密切的物种, 如治病物种和工业用菌等. 植物由于其多倍体化比较普遍, 对较大较复杂的植物基因组单纯用全基因组鸟枪法测序难以胜任, 必须构建物理图谱[2], 使得项目进度相对比较缓慢, 基因组测序工作还停留在基因组较小的少数模式物种阶段. 动物(特别是后口动物, metazoa)由于进化、生物学和与人类关系密切等等方面的因素, 虽然不少物种的基因组并不

图 1　基因组数据汇总

小，但是研究进展十分可观. 脊椎动物如人(*Homo sapiens*)、小鼠(*Mus musculus*)、马(*Equus caballus*)、牛(*Bos taurus*)、家狗(*Canis lupus familiaris*)等; 昆虫如果蝇(*Drosophila melanogaster*)、家蚕(*Bombyx mori*)和面粉甲虫(*Tribolium castaneum*)等, 基因组数据已深入到目(Order)的水平, 为比较基因组和功能基因组研究提供了很好的系列数据.

从各种科学发展的普遍规律看, 任何一轮实验数据和观测数据的大积累, 都会引起一场新技术和新学科的革命. 所以, 世界各国对基因组序列测定这场生命科学的新技术革命都报以极大的期望, 相继投入巨额资金进入这一领域. 其中规模较大的有以下几个方面:

肿瘤基因组计划(TCGA): 2006 年 11 月美国国立卫生研究院(NIH)启动“肿瘤基因组计划”. 计划利用 3 年时间花费 1 亿美元, 找出癌症的相关基因, 为将来研究癌症的发生和防治作准备.

肠道元基因组计划: 美国 NIH 在 2007 年 12 月 19 日宣布正式启动被誉为人体第二基因组计划的“人体微生物群系项目(HMP)”. 2008 年 4 月 11 日, 欧盟宣布启动人类元基因组第七框架资助项目. 目的是研究对人体健康有着巨大影响的人体肠道内微生物的种属组成和代谢特点.

千人基因组计划(1000 genome project): 2008 年 2 月, 美国、英国和中国同时宣布启动千人基因组计划. 通过对 1000 个人类个体基因组的测序绘成一张高精度的遗传变异图谱, 将能让研究者更快地锁定与疾病相关的基因变异点, 从而能够利用这些遗传信息, 更快地提出常见疾病的诊断、治疗和预防的新策略.

1.2　中国基因组生物信息学的发展历程

中国的生物信息学先驱自 20 世纪 80 年代就开始结合蛋白质结构等方面的工作开始致力于生物信息学技术的应用和研究. 而真正被视为基因组生物信

息学开端的是 1993 年, 在国家自然科学基金委员会的资助下, 中国涉足人类基因组计划. 在这之后, 1997 年我国成立了第一家生物信息中心——北京大学生物信息中心(www.cbi.pku.edu.cn). 它是欧洲分子生物学网络组织EMBnet的中国国家节点, 几年来与多个国家的生物信息中心建立了合作关系, 并为国内外用户提供了多项生物信息服务. 而后中国人类基因组研究北方中心(北京)、南方中心(上海)和华大基因研究中心(北京)相继成立. 为中国开展基因组生物信息学的研究创造了数据条件. 1999 年 9 月, 中国获准加入人类基因组计划, 承担测定人类基因组全部序列的 1%(位于 3 号染色体短臂上), 成为第 6 个国际人类基因组计划参与国. 基于人类基因组计划工作的实施, 中国在基因组测序方面得到了长足的进步. 先后测定了水稻(*Oryza sativa*) [3]、腾冲嗜热杆菌(*Thermoanaerobacter tengcongensis*)[4]、问号钩端螺旋体(*Leptospira interrogans*)[5]、家蚕(*Bombyx mori*)[6]、黄单胞菌(*Xanthomonas campestris pv*)[7]、日本血吸虫(*Schistosoma japonicum*)等一批生物的基因组序列, 并在 2007 年由深圳华大基因研究院等单位完成首个“中国人基因组图谱”. 2002 年 8 月国内第一个以推动我国生物信息学数据共享为目的生物信息学中心——上海生物信息技术研究中心(www.scbit. org)成立. 2006 年该中心收录了我国科学家测定的日本血吸虫基因组工作框架图, 并对全世界科研机构和科学家开放, 极大地丰富了我国的基因组生物信息数据资源, 也为建立我国自己的综合性核酸数据库奠定了基础.

2 基因组生物信息学的发展

2.1 测序技术的突破

基因组数据是基因组生物信息学发展的主要推动力, 而测序仪是提供基因组原始数据的关键设备. 1977 年, Sanger等人[1,8]发明了双脱氧核苷酸DNA测序方法, 并以他的名字命名. 1986 年Leroy Hood和他的同事对Sanger的方法作了改进, 并发明了DNA自动测序仪. Applied Biosystems公司使用这种方法制造出第一台自动测序仪[9], 一天能够测定 4,800 个DNA碱基序列, 由此测序研究具备了实际可行的核心设备. 直至今日市场上主流的测序仪仍然采用这种设计. 但已采用毛细管阵列电泳(Capillaries Arrays Electrophoresis)技术, 这一时期可以视为Sanger测序方法的自动化时代.

2006 年, 新一代基于“非Sanger技术”的基因测序方法相继问世. 相对于以前的测序仪, 新一代的DNA测序仪广泛使用了并行度很高的序列扩增技术和数据读取技术. 其特点是通量大、速度快、准确度高和成本低廉. 如以emPCR(emulsion PCR)[10]和焦磷酸测序技术(Pyrosequencing)[11,12]为核心的Roche公司的454-GS20 和 454-FLX测序仪; 以支持寡核苷酸链接和检测技术为核心的Applied Biosystems公司的ABI SOLiD; 和以单分子阵列原位扩增测序技术为核心的Illumina 公司的Solexa Genome Analysis System测序仪等. 2008 年美国Helico公司开发出了单分子DNA测序仪(single-molecule DNA sequencers). 该测序仪能够测定单分子DNA的单碱基, 而不用进行PCR扩增. 它能够用 8 周的时间测序一个人的基因组而只花费 72,000 美元. 新的技术无疑使得基因组测序更加方便和便宜. 在图 1 中可以看到, 基因组数据的加速积累总是紧跟于新测序技术的问世. 目前新一代测序仪已经相继问世并逐步走向成熟, 由此带来的基因组数据新一轮的“爆炸式”积累无疑对数据的存储、组织、分析和比较提出了更高的处理能力、速度和精确度要求.

2.2 数据库

基因组数据量非常庞大, 有组织地收集和管理这些数据是开展各项工作的前提. 基因组数据库内容丰富、名目繁多、格式不一, 分布在世界各地的信息中心、测序中心以及有关的研究机构和大学. 为了便于研究人员共享这些数据, 及时得到最新的实验数据结果, 也为了保证基因组数据的一致性和完整性, 世界各国政府都相继建立专门的机构搜集和管理这些数据, 还有一些企业也提供商业的生物信息服务. 其中最权威的三大国际数据库为 GenBank, EMBL 和 DDBJ. 这些数据库涵盖了从完整基因组到单个基因的序列数据和注释信息. 为保证数据尽可能的完整, GenBank 与 EMBL, DDBJ 建立了相互交换数据的合作关系, 各数据库中的数据基本一致, 仅在数据格式上略有差别. 对于特定的查询, 三个数据库

的返回结果基本一致.

此外, 还有些专门的模式生物基因组数据库, 除了收录基因组数据资源, 还收录分子生物学及遗传学等大量信息, 为相关研究领域提供了共享和交流信息的平台. 如AceDB是线虫基因组数据库, SGD为酿酒酵母基因组数据库(表 1). 我国首次面向全世界发布的日本血吸虫基因组数据库也于 2006 年在上海生物信息技术研究中心建立, 对我国建立自主的国家生物信息科学数据共享平台具有重要意义. 随着基因组计划的普遍实施, 几十种动物、植物基因组数据被陆续公布, 各物种的基因组项目信息可以在NCBI 的 Genome Project DataBase(http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj)中找到. 基因组的结构和注释信息可以在UCSC Genome Browser (http://genome.ucsc.edu/), Ensembl Genome Browser (http://www.ensembl.org/) 和 NCBI 的 GenomeMap Viewer(http://www.ncbi.nlm.nih.gov/mapview/)中找到.

在各个物种的基因组数据基础上出现了一些针对特殊研究方向提供标准数据集的数据库, 如ASDB[13], Transfac[14], ooTFD[15]和INE[16]等; 国内的如DoriC[17], PlantTFDB[18], 和ProTISA[19]等(表 1). 这些数据库各自提出了一套方法, 结合数据自动分析处理技术、人工整理和实验验证等手段, 旨在提供尽可能真实可靠的数据, 为相应领域的算法研究和实验设计提供踏实的数据基础. 核酸研究杂志(NAR)还为此专门设立了一年一度的数据库专刊. 如果说

表 1 常用的基因组数据库

名称	网址
综合性基因组数据库	
UCSC Genome Browser	http://genome.ucsc.edu/
Ensembl Genome Browser	http://www.ensembl.org/
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/mapview/
模式生物基因组数据库	
大肠杆菌基因组数据库 EcoGene	http://ecogene.org/
酵母基因组数据库 SGD	http://www.yeastgenome.org/
疟原虫基因组数据库 PlasmoDB	http://plasmodb.org/plasmo/
日本血吸虫基因组数据库(中国)	http://www.chgc.sh.cn/japonicum/
线虫信息资源 AceDB	http://www.acedb.org/
果蝇基因组数据库 FlyBase	http://www.fruitfly.org/
家蚕基因组数据库 SilkDB(中国)	http://silkworm.genomics.org.cn
斑马鱼信息库 ZFIN	http://zfin.org/cgi-bin/webdriver?MIval=aa-ZDB_home.apg
鸡基因组数据库 ChickVD(中国)	http://chicken.genomics.org.cn
小鼠基因组数据库 MGI	http://www.informatics.jax.org/
拟南芥信息资源 TAIR	http://www.arabidopsis.org/
水稻基因组数据库 BGI-RIS(中国)	http://rice.genomics.org.cn
可视化基因组序列数据库 The Z Curve database(中国)	http://tubic.tju.edu.cn/zcurve/
中国的一些标准数据集	
原核生物基因组翻译位点数据库 ProTISA	http://mech.ctb.pku.edu.cn/protisa
细菌基因组 OriCz 区数据库 DoriC	http://tubic.tju.edu.cn/doric
同源盒结构基因数据库 HomeoDB	http://homeodb.cbi.pku.edu.cn/
多种植物转录因子数据库 PlantTFDB	http://planttfdb.cbi.pku.edu.cn/
基因组中自然反义转录本数据库	http://natsdb.cbi.pku.edu.cn/
非编码 RNA 和蛋白质相互作用数据库	http://bioinfo.ibp.ac.cn/NPInter
发生可变翻译起始为点的基因数据库	http://bioinfo.au.tsinghua.edu.cn/atie/
RNA 编辑位点数据库	http://bioinfo.au.tsinghua.edu.cn/dbRES/

三大国际数据库和各个物种基因组数据库的建立是为了更好地服务于基因组结构和功能数据的提交、存储和交流, 成为了生物学研究不可或缺的组成部分, 那么针对特殊研究方向构建标准数据集则是生物信息学作为一个学科自身独立发展的需要, 并对生物学研究起到了指导和引领作用.

各种各样来源不同的生物信息增加的同时, 也为人们的使用带来了问题. 如何有效地集成这些资源, 使得用户能够方便、高效的利用这些不同数据源的数据并挖掘出有用的信息是生物信息学中一个重要的研究方向. 关于生物数据的整合,目前已有一些工作和产品. 出现了不少“知识库”(knowledgeBase), 如宾夕法尼亚大学计算机系的BioKleisli系统, IBM研究院的DiscoveryLink系统, 曼彻斯特大学计算机系的TAMBIS系统, GSK公司和IBM研究院的TINet系统等. 我国的生物学信息研究者们也开发研究出了一些有关生物信息数据整合的系统, 如复旦大学等单位研制开发了以Ontology为核心的生物信息学数据的集成系统——BioDW[20]. 由此, 生物信息学所提供的数据库已经从原始数据的提交、存储、组织和交流功能, 上升到了进行生物学研究和科学思辨的辅助工具.

2.3 算法工具

面对“海量”的基因组数据, 基因组生物信息学所处理的对象十分复杂和庞大, 不可避免地涉及大量的分析工作如分析DNA语义和识别基因等,要用到大量数学方法. 其中动态规划(dynamic programming)、隐马尔可夫模型(hidden Markov models)和聚类算法是较经典的算法解决方案. 动态规划是生物信息学中一种基本的优化方法, 是一种解决多阶段决策问题的最优化方法. 动态规划将比较复杂的问题划分为若干阶段, 通过逐段求解, 最终获得全局最优解. 动态规划在DNA序列或者蛋白质序列的比对、基因识别、RNA结构预测、隐马尔可夫模型求解、生物分子探针优化设计等方面有着重要的应用[21]. 隐马尔可夫模型是由马尔可夫链(Markov Chain)发展扩充而来的一种随机模型, 隐马尔可夫模型可以被理解为一个双重随机过程, 一个是系统状态变化的过程, 另一个是由状态决定输出的随机过程. 隐马尔可夫模型被广泛用于寻找新基因的软件中. DNA序列模型的统计规律是未知的, 而隐马尔可夫模型能自动寻找出其隐藏的统计规律, 因此它具有独特的优越性. 这些算法和聚类等算法都有很成熟的数学理论支持, 并且常常应用于除序列数据之外其他各种数据的处理过程.

为了在不损失太多精确度的前提下, 不断提高运算速度, 产生了很多基于概率统计的近似算法(Heuristic), 比如经典的BLAST(basic local alignment search tool)算法. 它是一种基本的局部对位排列搜索工具, 首先由Samuel Karlin和Steven Altschul在 1990 年提出[22], 并在 1992 年进行了较大改进. 之后针对不同的序列比对应用产生了新的统计模型, PatternHunter[23], BLAT[24], BlastZ[25]和megaBlast[26]等算法工具应运而生. 各个主要的生物信息网站都推出了各自的算法工具集, 像NCBI的数据挖掘工具集(tools for date mining)(http://www.ncbi.nlm.nih.gov/Tools/); EMBL的工具集(http://www.ebi.ac.uk/Tools/)等. 国内的算法工具起步虽晚, 但近些年也取得了巨大的进步, 如Biosino开发的基于IBM Splash算法的生物序列模式搜索程序GPAT[27]; 原核生物蛋白编码基因识别程序ZCURVE[28]. 国内的生物信息学工具在魏丽萍的综述文章[29]中有非常详细的记录, 在此就不再一一列举了. 此外, 为方便非生物信息学专业的用户使用, 生物信息学软件应尽量整合在统一、友好的界面下, 形成分析平台. 因此, 我国生物信息学领域合作或自主研发了若干大型分析平台. 例如, 上海生物信息技术研究中心与英国帝国理工大学及所属InforSense公司合作, 建立了基于网络的分布式的数据库和高通量的整合信息分析平台KDE Biosciences, 使生物信息数据处理和挖掘工作自动化、流程化和可视化. 北京大学生物信息中心自主开发的生物信息学网上实验室WebLab通过网络为生物学家及生物信息学家提供一个方便易用、功能强大的“一站式”生物信息学分析平台.

3 基因组生物信息学未来的发展

3.1 深度测序所带来的挑战

新一代的测序仪提供了前所未有的高并行、快速度和低成本测序, 使得基因组研究工作可以迅速向

深度和广度发展. 从而产生了个体基因组(Personal genomics), 肿瘤基因组(The Cancer Genome Atlas), 环境基因组(Environmental genomics, or Metagenomics)和进化基因组(Evolutionary genomics)等多个研究方向. 前三者在本文第一部分有所表述. 2005 年 Poinar等用焦磷酸测序技术手段对猛犸象的古DNA进行了测定, 其中 45%的序列为该猛犸象的DNA序列[30], 有望为今后该基因组的组装提供支持, 这一成果在古生物学中是一个开创性的工作. 此外, 由于测序通量提高和成本降低, 在样品量不多的情况下用测序的方式进行全基因组表达谱分析可以发现新的基因序列, 相对于使用基因芯片有着无法比拟的优势, 尤其在microRNA的研究中发挥了很大的作用. 同样, 当人们朝着 1000 美元重测序一个人全基因组的目标迈进的时候, 传统的SNP分型技术也大有被新测序技术取代的趋势.

虽然, 新测序技术有着"大一统"的美好未来, 但是任何新技术都具有两面性, 我们将从生物信息学数据处理的方面展开讨论.

首先, 在 Sanger 测序时代测序仪的发展主要是提高自动化程度和样品通量, 其核心原理和技术并没有大的飞跃, 只是在外围技术上进行不断地改进, 数据的可靠性有多年研究经验作为保证. 而目前至少有 7 个公司正在发展各自的新一代测序技术(如 Applied Biosystems, GE Healthcare, Helicos BioSciences, Illumina/Solexa, Reveo, Roche/454 Life Sciences 和 VisiGen Biotechnologies 等), 由于专利问题各自都使用不同的技术, 更由于商业利益的问题或者是对新事物认知过程等问题, 常常在初期对各种测序技术局限性的认识不够深. 比如, 虽然 Roche454 测序仪的读长最有希望与传统 Sanger 方法媲美, 但是其单一重复碱基的错误率在商业报告中一直被低估. 当然任何技术都有其局限性, 只要能正确了解各技术的特点和局限, 就可以用组合测序的办法取长补短.

其次, 目前深度测序技术所提供的原始数据由于数据读取方式与以往不同, 而且提供的序列长度较短但是覆盖度很高, 数据的统计特性有较大改变, 所以序列拼接的算法也将需要作相应的调整. 然而要解决数据准确性这个最基本的问题, 还是需要依靠各个公司及购买其产品的合作伙伴, 相对之前人类基因组计划中各方合作的情况显得力量较为分散. 当然参加角逐的这些公司中不乏世界顶尖级的大财团的支持, 其资金实力、人才实力和号召力是值得期待的.

除了原始数据可靠性和数据拼接算法效率等问题之外, 数据产出能力大大超过了数据存储、处理和分析能力也是个巨大的挑战. 比如说, 重测序所产生的短序列片段需要比对到基因组的参考序列上, 这一过程用传统的 BLAST 算法显得力不从心. 首先, 序列虽短但数目庞大, 每个序列比对一次, 计算设备的非运算开销浪费严重. 其次, 序列越短单个碱基位置对比对结果影响就越大, 必须考虑碱基位置的可信度才能保证比对结果的可靠性, 而 BLAST 并不考虑碱基位置的质量问题. 最后, 序列片段越短, 整个片段完全落入重复区域的可能性就越大. 这些片段应尽快尽早剔除, 以免造成不必要的运算资源浪费. 所以需要不断提高识别重复序列的算法效率. 除了简单的序列比对工作, 那些需要更复杂处理的过程, 需要更强大运算能力支持的序列功能挖掘算法, 如模式识别、基因家族聚类、进化分析和统计分析等, 都面临着一系列提高处理通量和优化处理算法的挑战.

3.2 环境基因组学数据所带来的挑战

"环境基因组(Environmental genomic/Metagenomic, 又称元基因组)", 即环境中全部微小生物遗传物质的总和, 目前主要指环境样品中的细菌和真菌的基因组总和. 元基因组文库既包含了可培养的又包含了未能培养的微生物基因, 避开了微生物分离培养的问题, 目前元基因组的研究已经涉及到土壤、海洋、热泉口、人体和矿产等领域. 极大地扩展了微生物资源的利用空间. 微生物环境基因组学为最大限度地挖掘微生物资源带来了前所未有的机遇, 极端微生物在极端环境下的代谢特征对人类了解生命起源、生命本质和生命极限, 开发新型药物和生物制品提供了机遇, 其中对极端微生物特征蛋白质结构和功能的认识是关键, 具有十分重要的意义, 已经成为国际生命科学技术研究和开发最重要的热点. 我国参加了中欧合作的肠道元基因组项目和已经完

成了 1000 多人的上海常住居民“营养、菌群与肥胖的病例对照研究”的现场体检和血液、尿液和粪便样品的采集工作. 这是目前国际上规模最大的人类元基因组人群研究项目, 备受国际同行关注.

虽然环境基因组学数据所带来的挑战很多方面与新测序技术所带来的挑战一致, 不过其具体的数据处理有着更为复杂的情况. 首先, 需要整理现有微生物数据, 对全基因组数据、基因片段和蛋白质序列片段以及 16S/18S/ITS序列都需要分门别类地进行整理, 尽可能去除冗余并明确其物种属性和功能信息. 这一工作的运算量和工作量已经相当可观了. 其次, 每个元基因组学文库的测序数据量将远远高于对单个微生物基因组的测序量, 并且不同物种序列混合拼接的难度是相当大的, 这一课题值得长期的深入研究. 目前的解决方法是尽量绕开混合拼接的问题, 将序列与已知序列进行比对, 以确定目的序列的种属和功能特性. 这就带来了第三个挑战, 即如何更准确快速的进行此类比对, 虽然已经有一些研究成果[31], 但是要取得令人满意的进展必须从整理现有数据、研究新的拼接算法和研究比对算法三方面同时入手.

3.3 后基因组时代大规模功能解析

随着基因组测序技术的发展及众多基因组测序计划的提出和实施, 很多动、植物已经完成了基因组测序, 将来会有更多的基因组序列被测定. 因此如何分析这些基因组序列, 挖掘序列数据所蕴含的功能信息, 已经成为基因组学研究的重要内容. 基因组计划的发展为研究者提供了极其丰富的资源. 但是, 迄今为止, 仍有许多基因和DNA元件的功能未被发现. 由此在 2003 年产生了ENCODE项目[32], 并在 2007 年得以扩大研究.

(1) 用比较基因组方法进行基因和 DNA 元件功能注释. 比较基因组学研究跨生物学分类界限的基因组结构和功能的相似性和差异, 已经成为了挖掘潜在功能信息的强有力工具. 之前的比较基因组学较多解决的是少数物种间比较的问题. 现在特别是在微生物基因组研究领域, 常常出现同一个种的不同特色菌株都有全基因组序列数据. 这样的局面使得比较基因组学必须能同时比较多个物种, 运用一些进化方面的算法, 并运用一些网络生物信息学的成果进行网络化比对进行功能信息挖掘已成为必然. 在真核模式生物方面, 这样的研究趋势也已经非常明显. 2007 年美国、丹麦等国研究人员利用 12 种果蝇的基因组数据, 完成首次大规模全基因组比对, 确定出了数千个新的基因和其他功能元件. 从中揭示了新的基因, 新的结构和调控子. 围绕着线虫也有至少 5 个基因组整体比较的计划, 被称为 Roundworm 基因组计划(http:// www.genome.gov/11007952).

经过研究发现, 多个基因组共同注释比单物种基因组单独作注释的准确性和覆盖度要好得多. 运用Ortholog和基因家族Profile构建等方法, 可以跨过物种界限将基因横向联系起来, 相互提示基因功能信息, 如基于KEGG Ortholog(KO)的KAAS服务器[33]和基于SEED[34]数据库的RAST服务器, 以及最近发表的IMG数据库[35].

(2) 用网络生物信息学进行基因功能挖掘. 近年来各种生物学网络理论的研究和通过构建生物网络进行基因功能挖掘的研究正逐渐成为生物信息学领域的研究热点. 要了解细胞的整体状态, 必须以我们所掌握的知识重建复杂生物学网络并进行研究, 从基因组水平上对基因的活动规律进行阐释. 这个变化从根本上改变了传统生物学的思维方式, 形成了一种新的全局方法[36]. 重建网络的方向基本有以下方面:

首先是代谢/调控网络, 如KEGG等已经整理了跨物种的代谢网络图, 并在积极完善各种调控网络图. 其次是基因表达网络, 基因表达存在组织特异性、细胞周期特异性和外界信号的响应特异性等特性, 这些特异性都是由细胞内复杂而有序的调控机制实现的. 对基因表达数据的研究能够构建复杂的表达网络和表达调控网络. 再次是分子相互作用网络, 随着大量分子相互作用数据的出现[37], 根据这些相互作用数据构建的生物学网络规模越来越大. 最后是遗传网络, 通过遗传标记的连锁分析和关联分析等研究可以将一些基因与特定的表型联系起来, 提供功能线索, 这便可以组建遗传网络.

然而, 任何一个网络都面临着数据量过于庞大和信噪比较低的问题, 用单独研究的方法和手段, 难以得到统计上显著性较强的结果. 并且, 任何一方面的数据都不可避免地存在假阳性和假阴性问题, 从

而干扰分析结果. 所以, 整合上述各种网络信息, 从整体网络结研究基因的相互关系, 挖掘基因的功能信息更符合细胞的生命本质, 在乳腺癌的相关基因研究中这种方法就获得了成功[36].

4 结束语

随着基因组以及和基因组相关的各类“组学”和“系统生物学”数据的不断加速积累, 生物信息学的技术和理论研究变得越来越重要. 一方面, 数据不仅在数量上飞速增长, 而且在性质上正日益体现“多元”、“多维”和“复杂”、“异质”等特征; 另一方面, 这些新数据集也正在不断地要求人们提出新的研究课题甚至开拓新的研究领域. 这样, 生物信息学从单纯地借用计算机和数学方面的技术和理论, 对生物学数据进行分析的一种技术手段, 发展成为了包括数据库、算法、软件包和计算与知识挖掘平台的完整体系, 并且在这些领域内出现了诸多研究方向和热点. 在这个方面, 我国基因组生物信息学正面临着从研究到应用, 从技术到管理的巨大挑战. 同时, 随着越来越多的数据被免费在网上公开, 我国与欧美发达国家在生物信息学研究领域的差距正在不断缩小, 生物信息资源的利用度和利用效率, 在很大程度上决定了一个国家生命科学以及生物技术研究和开发的速度和水平. 在基因组研究刚起步时, 有一句发人深省的话“人类正在从殖民时代对国土资源的瓜分, 走向对基因资源的瓜分”. 在过去的 10 余年间, 这种“瓜分”主要体现于“序列”, 其中的大部分可能没有直接的效益. 但是, 在今后的 10 余年间, 这种瓜分将转移到对“知识”的“瓜分”, 即通过生物信息技术对数据挖掘而得到的知识及其挖掘技术的占有. 中国的生物信息学工作者们必需再接再厉, 不断提高自身的研究水平和服务水平; 与全世界同行们一起, 肩并肩地不断在生物学领域发现新规律, 提出新假说, 提供新知识. 相信他们作出重大科学发现和技术创新贡献的日子已经不远了.

参考文献

1 Sanger F, Air G M, Barrell B G, et al. Nucleotide sequence of bacteriophage phi X 174 DNA. Nature, 1977, 265: 687—695[DOI]

2 Paux E, Sourdille P, Salse J, et al. A physical map of the 1-gigabase bread wheat chromosome 3B. Science, 2008, 322: 101—104[DOI]

3 Yu J, Hu S, Wang J, et al.A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science, 2002, 296: 79—92 [DOI]

4 Bao Q, Tian Y, Li W, et al. A complete sequence of the *T. tengcongensis* genome. Genome Res, 2002, 12: 689—700[DOI]

5 Ren S X, Fu G, Jiang X G, et al. Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. Nature, 2003, 422(6934): 888—893[DOI]

6 Xia Q Y, Zhou Z Y, Lu C, et al. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science, 2004, 306: 1937—1940[DOI]

7 Qian W, Jia Y, Ren S X, et al. Comparative and functional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. campestris. Genome Res, 2005, 15(6): 757—767[DOI]

8 Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, 74: 5463—5467[DOI]

9 A History of the Human Genome Project. Science, 2001, 291(5507): 1195[DOI]

10 Williams R, Peisajovich S G, Miller O J, et al. Amplification of complex gene libraries by emulsion PCR. Nat Methods, 2006, 3(7): 545—550[DOI]

11 Margulies M, Egholm M, Altman W E, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005, 437(7057): 376—380

12 Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Res, 2001, 11(1): 3—11

13 Zhou Y, Zhou C, Ye L, et al. Database and analyses of known alternatively spliced genes in plants. Genomics, 2003, 82: 584—595[DOI]

14 Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform, 2008, 9(4): 326—332[DOI]

15 Ghosh D. Object oriented Transcription Factors Database (ooTFD). Nucleic Acids Res, 1999, 27(1): 315—317[DOI]

16 Sakata K, Antonio B A, Mukai Y, et al. INE: a rice genome database with an integrated map view.Nucleic Acids Res, 2000, 28(1):

97—101[DOI]
17 Gao F, Zhang C T. DoriC: a database of oriC regions in bacterial genomes. Bioinformatics, 2007, 23: 1866—1867[DOI]
18 Guo A Y, Chen X, Gao G, et al. PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res, 2008, 36: D966—969[DOI]
19 Hu G Q, Zheng X, Yang Y F, et al. ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes. Nucleic Acids Res, 2008, 36: D114—119[DOI]
20 朱海燕. 试析异构生物信息数据库的整合. 现代情报, 2006, 3: 50—52
21 孙啸, 陆祖宏, 谢建明. 生物信息学基础. 北京, 清华大学出版社, 2005, 23
22 Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. Mol Biol, 1990, 215: 403—410
23 Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics, 2002, 18(3): 440—445[DOI]
24 Kent W J. BLAT-the BLAST-like alignment tool. Genome Res, 2002, 12(4): 656—664
25 Schwartz S, Kent W J, Smit A, et al. Human-mouse alignments with BLASTZ. Genome Res, 2003, 13(1): 103—107[DOI]
26 Zhang Z, Schwartz S, Wagner L, et al. A greedy algorithm for aligning DNA sequences. J Comput Biol, 2000, 7(1-2): 203—214[DOI]
27 Xu Y, Li Y X, Kong X Y. GNU Pattern: open source pattern hunter for biological sequences based on SPLASH algorithm. Zhongguo Yi Xue Ke Xue Yuan Xue Bao, 2005, 27(3): 265—269
28 Guo F B, Ou H Y, Zhang C T. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res, 2003, 31: 1780—1789[DOI]
29 Wei L P, Yu J. Bioinformatics in China: a personal perspective. Computational Biology, 2008, 4(4): w1000020
30 Poinar H N, Schwarz C, Qi J, et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Sicence, 2006, 311: 392—394
31 Dalevi D, Ivanova N N, Mavromatis K, et al. Annotation of metagenome short reads using proxygenes. Bioinformatics, 2008, 24(16): i7—13
32 Weinstock G M. ENCODE: more genomic empowerment. Genome Res, 2007, 17(6): 667—668[DOI]
33 Moriya Y, Itoh M, Okuda S, et al. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res, 2007, 35: W182—185[DOI]
34 Overbeek R, Begley T, Butler R M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res, 2005, 33(17): 5691—5720[DOI]
35 Markowitz V M, Szeto E, Palaniappan K, et al. The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res. 2008 , 36: D528—533[DOI]
36 Pujana M A, Han J D, Starita L M, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet, 2007, 39(11): 1338—1349[DOI]
37 Salwinski L, Miller C S, Smith A J, et al. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res, 2004, 32: D449—451[DOI]