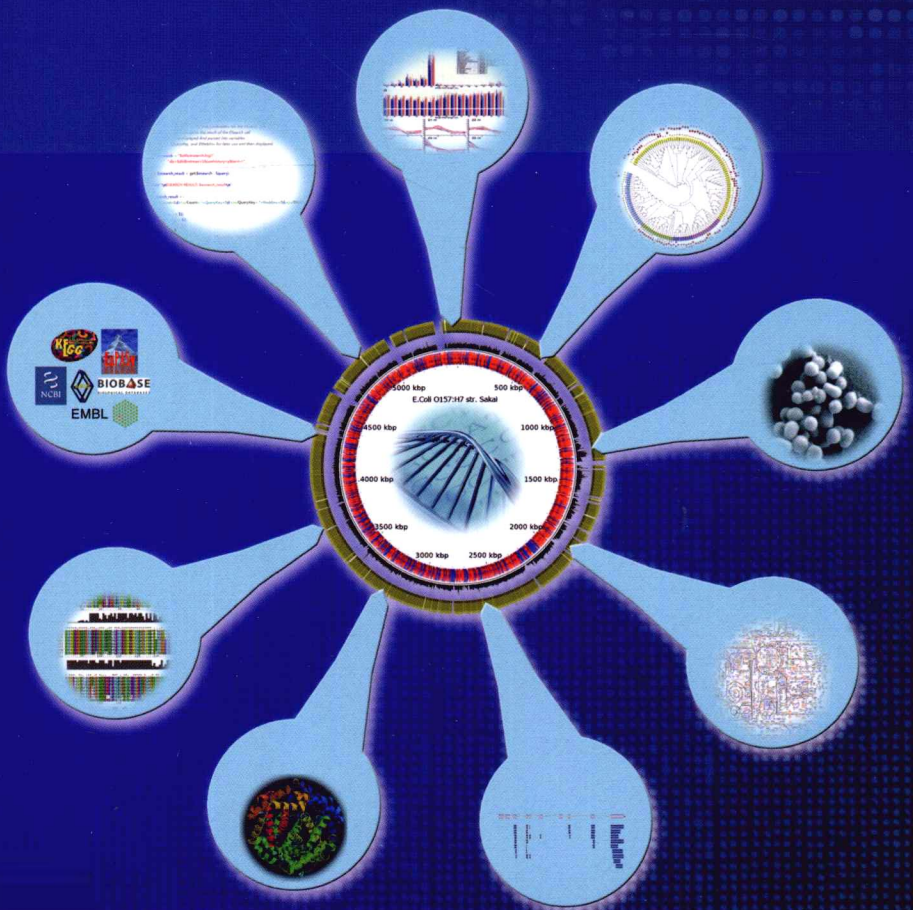
普通高等教育“十二五”规划教材

生物信息学

Bioinformatics

陈　铭　主编



科学出版社

普通高等教育“十二五”规划教材

生物信息学

陈　铭　主编

科学出版社

北　京

内 容 简 介

本书由16所985、211高校联合编写而成，系统全面地介绍了生物信息学的基本概念与内容。全书共12章，内容涵盖分子生物学数据库、DNA/氨基酸序列比对、基因结构与功能、蛋白质结构与功能、系统生物学、合成生物学、计算生物学等生物信息学中的重点问题。第一章回顾了生物信息学历史，第二章介绍了分子生物学相关信息资源，第三章至第六章叙述了从序列比对分析到蛋白质结构预测再到基因组学、蛋白质组学的研究，第七章和第八章介绍了系统生物学与合成生物学的研究内容及成果，第九章介绍了分子进化分析方法，第十章讨论了统计学习与推理等基本知识，第十一章讨论了生物信息学基本的编程基础，第十二章叙述了第二代测序技术的基本概念与分析应用。

本书可用作高等院校生物信息学专业的教材，也可作为科研院所相关专业学生、研究人员的参考用书。

图书在版编目(CIP)数据

生物信息学/陈铭主编. —北京：科学出版社，2012.1
普通高等教育“十二五”规划教材
ISBN 978-7-03-033205-9

Ⅰ. ①生… Ⅱ. ①陈… Ⅲ. ①生物信息论-高等学校-教材
Ⅳ. ①Q811.4

中国版本图书馆CIP数据核字（2011）第277158号

责任编辑：单冉东 刘 晶 / 责任校对：郑金红
责任印制：张克忠 / 封面设计：陈 敬

科学出版社出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
新科印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2012年1月第 一 版 开本：787×1092 1/16
2012年1月第一次印刷 印张：20 插页：4
字数：460 000
定价：48.00元
（如有印装质量问题，我社负责调换）

《生物信息学》编委会名单

主　编　陈　铭

副主编　何华勤　徐　程

编　者（按姓氏汉语拼音排序）

蔡　禄（内蒙古科技大学）
陈玲玲（华中农业大学）
陈　铭（浙江大学）
陈庆峰（广西大学）
何华勤（福建农林大学）
胡　浩（山东理工大学）
纪洪芳（山东理工大学）
林　魁（北京师范大学）
凌　毅（中国农业大学）
刘笔锋（华中科技大学）
宋　凯（天津大学）
唐玉荣（中国农业大学）
魏冬青（上海交通大学）
魏天迪（山东大学）
徐辰武（扬州大学）
徐　程（浙江大学）
徐德昌（哈尔滨工业大学）
杨泽峰（扬州大学）
易图永（湖南农业大学）
袁哲明（湖南农业大学）
张红雨（华中农业大学）
张文广（内蒙古农业大学）
张子丁（中国农业大学）

序

生物信息学（bioinformatics）是20世纪80年代末随着人类基因组计划的启动而兴起的一门新兴交叉学科，体现了生物学、计算机科学、数学、物理学等学科间的渗透与融合。它通过对生物学实验数据的获取、加工、存储、检索与分析，达到揭示数据所蕴含的生物学意义从而解读生命活动规律的目的。

生物信息学不仅是一门科学学科，更是一种重要的研究开发平台与工具。它是今后进行几乎所有生命科学探索，包括生物医药研究开发所必需的重要推手，只有基于生物信息学对大量已有数据资料的分析处理所提供的理论指导，我们才能选择正确的研发方向；同样，只有选择正确的生物信息学分析方法和手段，我们才能正确处理和评价新的实验数据并得到准确的结论。生物信息学已经在生物学、医学、农业、环境科学、信息技术，以及新材料的研究中得到广泛的应用，生物信息学的继续发展也必将为这些领域带来持续性发展与学科前沿突破。

21世纪的科学正呈现出前所未有的技术融合趋势，特别是生物技术与其他高技术的融合，产生了以生物信息为代表的生物技术群。我国也极为重视生物信息学的发展：南、北方人类基因组中心的相继建成，标志着我国生物信息学的研究进入崭新的阶段。国家“973”项目、“863”计划特别设立了生物信息技术主题，从国家需求的层面上推动我国生物信息技术的大力发展。我们有理由相信，我国的生物信息学在21世纪会有巨大的飞跃。因此，生物信息学人才的培养是当前首要的任务，要加强有关学科间协作和加速培养在数学、物理、信息科学、计算机科学及生物学方面均有造诣的生物信息学“双栖”人才。

为此，编写一本适应21世纪人才培养需要且能反映最新进展的生物信息学教材是十分必要的。浙江大学陈铭教授联合各高校青年学者，紧密跟踪学科发展，提炼学科精华，编写完成了这本《生物信息学》。全书涵盖了生物信息学、系统生物学、合成生物学的相关内容，以及应用于第二代测序技术的相关软件和算法。该书内容深入浅出，图文并茂，适合广大生物信息学爱好者和从事生物信息学的研究人员使用。衷心希望《生物信息学》能成为广大青年学生及科技工作者迈向生命科学前沿领域的钥匙和助手，激发年轻学子的求知欲和学习热情，更加崇尚科学、追求创新！

2011年7月22日

前　言

继工业革命和以计算机为基础的信息技术革命之后，一场以基因为基础的生物科学技术革命正在形成并将迅速蔓延，其影响力将丝毫不逊于前两场革命。随着人类基因组计划的顺利实施，生物学的序列数据正呈爆炸式增长，人们惊呼“基于序列的生物学时代已经到来”。我们在挖掘蕴藏在大量序列数据中的生物学规律的同时，逐步完善了“生物信息学”学科的内涵。

近年来，以 Roche 公司的 454 技术、Illumina 公司的 Solexa 技术和 ABI 公司的 SOLiD 技术为标志的第二代测序技术的诞生，极大地促进了生物信息学的发展。生物信息学已广泛应用于基因组研究、蛋白质组研究、药物设计与进化分析等诸多领域。生物信息学的发展离不开人才，正如斯坦福大学的基因学教授大维·波特所讲，“我们需要既懂计算机又懂生物学的人才，就像以前我们需要既懂化学又懂生物学的人才一样”。当前，集二者之优的生物信息学人才已成为最紧缺的人才类型之一。在长期教学与科研实践的过程中，我们逐渐认识到编写一本系统反映生物信息学内涵和前沿研究的教学与科研用书的重要性。

本书相对系统全面地介绍了生物信息学的基本概念与内容。首先介绍了生物信息学的产生与发展概况；然后由浅及深地分析了生物信息学研究的基本内容，从生物学数据库检索与序列比对，到基因组注释、蛋白质结构预测与生物进化分析；之后阐述了由生物信息学而引申出的新的前沿学科，包括系统生物学与合成生物学；最后概括介绍了第二代高通量测序技术的应用。

本书的编写是由长期从事生物信息学教学与科研工作的一线人员共同完成的。编写队伍主要是我国高等院校生物信息学研究的骨干人员，成员大多具有国外学习经历，其学术视野开阔、专业思想先进。同时感谢初稿整理过程中陈迪俊、焦胤茗、原春晖、白琳、白有煌、张钊、刘丽丽等同学的帮忙。

自 2008 秋在京酝酿、筹备，到 2011 年夏季定稿，本书先后得到了教育部、科技部、相关院校、科研单位专家和领导的大力支持，他们为本书的编写提出了诸多合理建议；张先恩先生为本书作序，在此一并表示诚挚的谢意！

由于作者水平和能力有限，在编写过程中难免存在不足和错漏，恳请同仁不吝赐教，以便及时改正。勘误表及更多信息可访问 http://www.cls.zju.edu.cn/binfo/textbook。

编　者

2011 年 7 月

目　录

第一章　生物信息学的概念及其发展历史

本章提要：自从1990年美国启动人类基因组计划以来，人与模式生物基因组的测序工作进展极为迅速。美国最新公布的GenBank数据库版本拥有的DNA序列总量已超过1265亿碱基对，与其同步增长的还有氨基酸序列，序列信息像潮水般向人们涌来。因此，有人说，基于序列的生物学时代已经到来。生物学家面临的最主要的一个困难就是处理浩瀚的数据，序列数据并不等于信息和知识，却是信息和知识的源泉，关键在于如何从中挖掘它们，这就催生了一门新兴的交叉科学——生物信息学。21世纪是生命科学的世纪，离不开生物信息学的发展。生物信息学是计算机与信息科学技术运用到生命科学，尤其是分子生物学研究中的交叉学科。

第一节　生物信息学的发展历史

随着基因组计划的不断进展，海量的生物学数据必须通过生物信息学的手段进行收集、分析和整理后，才能成为有用的信息和知识。人类基因组计划为生物信息学提供了兴盛的契机。目前，生物信息学已经深入到了生命科学的方方面面。

欧美等国一直非常重视生物信息学的发展，各种专业研究机构和公司如雨后春笋般涌现出来，生物科技公司和制药工业内部的生物信息学部门的数量与日俱增。但由于对生物信息学的需求是如此迅猛，即使是像美国这样的发达国家也面临着供不应求、人才匮乏的局面。

目前，各类生物信息学专业期刊门类繁多，包括纸质期刊和电子期刊两种，如 *Bioinformatics*（前身为“*Applications in the Biosciences*”）、*PLoS Computational Biology*、*BMC Bioinformatics*、*Nucleic Acids Research*、*Acta Biotheoretica*、*Bioinformatics Technology & Systems*、*Bioinform Newsletter*、*Briefings in Bioinformatics* 及 *Journal of Computational Biology* 等。

从网络资源来看，国外互联网上的生物信息学网点非常繁多，大到代表国家级研究机构，小到代表专业实验室。大型机构的网点一般提供相关新闻、数据库服务和软件在线服务；小型科研机构一般是介绍自己的研究成果，有的还提供自行设计的算法在线服务。总体而言，它们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，但却分散在不同的网点，分析结果也需专业人士来解读。

目前，绝大部分的核酸和蛋白质数据库由美国、欧洲及日本的三家数据库系统产生，它们共同组成了GenBank/EMBC/DDBJ国际核酸序列数据库，每天交换数据，同步更新。其他一些国家，如德国、法国、意大利、瑞士、澳大利亚、丹麦和以色列等，在分享网络共享资源的同时，也分别建有自己的生物信息学机构、次级或者衍生的具有各自特色的专业数据库及自己的分析技术，服务于本国生物医学研究和开发，有些服务也开放于全世界。

国内对生物信息学领域的研究也越来越重视，自北京大学于1996年建立了国内第一个生物信息学网络服务器以来，我国生物信息学的研究得到蓬勃发展。较早开展生物信息学研究的单位主要有：北京大学、清华大学、浙江大学、中国科学院生物物理所、中国科学院上海生命科学研究院、中国科学院遗传与发育生物学研究所等。北京大学于1997年3月成立了生物信息学中

心，中国科学院上海生命科学研究院也于2000年3月成立了生物信息学中心，分别维护着国内两个专业水平相对较高的生物信息学网站，但从总体上来看仍与国际水平有较大差距。现在，国内生命科学研究与开发对生物信息学研究和服务的需求市场非常广阔，然而，除华大基因（BGI）外，真正开展生物信息学具体研究和服务的机构或公司却相对较少，仅有的几家科研机构主要开展生物信息学理论研究，提供生物信息学服务的公司也大多进行的是简单的计算机辅助分子生物学实验设计。

表1-1列出了生命科学、计算机科学及生物信息学大事记，从中可以看出其发展进程及中国的贡献。

表1-1 生命科学、计算机科学及生物信息学相关大事记

生命科学	年份	计算机科学
	1642	Blaise Pascal 发明机械计算器
	1858	电报
达尔文的《物种起源》出版	1859	
孟德尔遗传定律	1865	
首次分离得到 DNA	1869	
	1876	电话
Walter Flemming 观察到有丝分裂	1879	
确认孟德尔遗传定律	1900	
疾病可以有序遗传；遗传的染色体理论	1902	
术语“基因”的出现	1909	
染色体理论在果蝇中得到验证	1911	
一个基因一个酶	1941	
DNA 的 X 射线衍射	1943	第一台电子管计算机 ENIAC 研发并于 1946 年诞生
DNA 可以改造细胞的特性；跳跃基因的发现	1944	
	1945	第一个计算机 Bug
DNA 构成基因	1952	第一个编译器的发明
Francis Crick、James Watson 和 Maurice Wilkins 发现 DNA 的双螺旋结构	1953	
人类 46 条染色体的确定；DNA 聚合酶的发现；第一个蛋白质序列（牛胰岛素）被测定	1955	
血红蛋白的一个氨基酸改变可以导致镰状细胞贫血	1956	
DNA 的半保留复制	1958	中国第一台电子管计算机诞生
染色体异常致病被发现	1959	
	1960	计算机 COBOL 处理电话交换
mRNA 将信息从细胞核内传递到细胞质	1961	
	1963	美国信息互换标准代码（ASCII）；鼠标
	1964	BASIC 语言
中国人工合成牛胰岛素结晶；Margaret Dakley Dayhoff 收集蛋白质序列，并在随后一年提出 PAM 模型	1965	
发现第一个限制酶	1968	
	1969	UNIX 操作系统
	1970	Needleman-Wunsch 序列比准算法

续表

生命科学	年份	计算机科学
	1971	个人电脑
第一个重组 DNA	1972	C 语言
第一个动物基因被克隆	1973	文件传输协议(FTP)出现
DNA 测序工作的开启	1975	微软公司成立
第一个遗传工程公司成立	1976	苹果公司成立
Sanger 研究小组完成了第一个噬菌体全基因组的测序;内含子的发现	1977	
	1978	第一个电子布告栏系统(BBS)的出现
	1979	新闻组(Newsgroup)的出现
中国实现酵母丙氨酸转移核糖核酸的人工合成	1981	第一个计算机病毒 Eld Cloner 出现;Smith-Waterman 序列比准算法;MS-DOS 1.0 发布
	1982	Sun 公司推出第一个工作站 Sun 100;英特尔 80286 处理器
	1983	微软 Windows 系统命名
	1984	互联网节点数超过 1000 个
Kary. Mullis 创立 PCR 技术;生物信息学专业期刊(CABIOS)创刊;德国生物信息学会议(GCB)举行	1985	Bjarne Stroustrup 创建 C++语言
日本核酸序列数据库 DDBJ 诞生;蛋白质数据库 Swiss-Prot 建立;中国开始实施“863 计划”	1986	标准通用置标语言(SGML)ISO 标准公布
	1987	Perl 语言
美国国家生物技术信息中心成立	1988	Pearson 实现 FASTA 程序
	1989	英特尔发布 486 处理器
国际人类基因组计划(HGP)启动;第一届国际电泳、超级计算和人类基因组会议在美国佛罗里达州会议中心举行	1990	Altschul 实现 BLAST 程序;HTTP 1.0 标准发布
	1991	Linux 出现;Python 语言发布
欧洲生物信息学研究所(EBI)获准成立;第一届 ISMB 国际会议在美国国家医学图书馆(NLM)举行;HGP 新 5 年计划,中国开始参与人类基因组计划	1993	英特尔发布奔腾处理器
Marc Wilkins 提出蛋白质组(proteome)的概念;细菌基因组计划	1994	雅虎公司成立;Perl 5 发布
人类基因组物理图谱完成;日本信息生物学中心(CIB)成立	1995	Sun 正式发布 Java;Apache HTTP 项目启动;微软发布 Windows 95 系统
Affymetrix 生产商用 DNA 芯片;北京大学蛋白质工程和植物遗传学工程国家实验室加入欧洲分子生物学网络(EMBnet)	1996	微软发布 IE3.0
大肠杆菌基因组测序完成;北京大学生物信息学中心(CBI)成立;中国科学院召开“DNA 芯片的现状与未来”和“生物信息学”香山会议	1997	微软发布 IE4.0;IBM 深蓝计算机击败国际象棋世界冠军

续表

生命科学	年份	计算机科学
亚太生物信息学网络(APBioNet)成立;瑞士生物信息学研究所(SIB)成立;美国 Celera 遗传公司成立;线虫基因组测序完成;CABIOS 期刊更名为 *Bioinformatics*;中国人类基因组研究北方中心(北京)和南方中心(上海)成立	1998	W3C 发布可扩展标记语言 XML 1.0;微软发布 Windows 98
人类 22 号染色体序列测定完成;中国获准加入人类基因组计划,成为第 6 个国际人类基因组计划参与国	1999	英特尔发布奔 III 处理器
德国、日本等国科学家宣布基本完成人体第 21 对染色体的测序工作;果蝇基因组测序完成;中国科学院上海生命科学研究院生物信息中心(SIBI)成立	2000	微软发布 Windows 2000 和 Windows Me 简单对象访问协议(SOAP)
美国、日本、德国、法国、英国、中国 6 国科学家和美国 Celera 公司联合公布人类基因组图谱及初步分析结果;中国首届全国生物信息学会议(CCB)举行;中国完成籼稻基因组工作框架图	2001	微软发布 Windows XPLinux 内核 2.4
小鼠基因组测序完成	2002	
HGP 完成	2003	微软发布 Windows Server 2003;Linux 内核 2.6
蛋白质组学;解码基因组;大鼠和鸡基因组草图完成	2004	
大猩猩和狗全基因组测序完成;人类 HapMap 项目完成	2005	
我国研制出全球首例骨髓分析生物芯片	2006	
世界首份“个人版”基因图谱完成	2007	谷歌和 IBM 合作推动云计算
千人基因组测序计划启动;拟南芥 1001 株系测序启动	2008	英特尔发布酷睿 i7 处理器
黄瓜、高粱和两个玉米品种的基因组测序	2009	
外显子测序	2010	我国“天河一号”超级计算机以每秒 2570 万亿次的实测运算速度成为全球运算速度最快的超级计算机
	2011	日本超级计算机 RFI 成为全球最快计算机

第二节 生物信息学的研究领域

虽然生物信息学可以理解为“生物学+信息学(计算机科学及应用)”,但作为一门学科,它有自己的学科体系,而不是简单的叠加。需要强调的是,生物信息学是一门工程技术学科。必须注意到,生物信息学的研究内容与研究对象或客体(应用方面)是不同的概念。很显然,生物信息学的研究对象是生物数据。其中最“经典”的是分子生物学数据,即基因组技术的产物——DNA 序列。后基因组时代将从系统角度研究生命过程的各个层次,走向探索生命过程的每个环节,包括微观(深入到研究单个分子的结构和运动规律)和宏观(结合宏观生态学,从大的角度来研究生命过程)两个方向,着重于“序列→结构→功能→应用”中的“功能”和“应用”部分。就研究面来说,

其涉及并参与各个生命科学领域的研究(陈铭,2004)。

1. 分子生物学与细胞生物学

该领域以 DNA — RNA —蛋白质为对象,分析编码区和非编码区中信息结构和编码特征,以及相应的信息调节与表达规律等。由于生物功能的主要体现者是蛋白质及其生理功能,研究蛋白质的修饰加工、转运定位、结构变化、相互作用等活动将推动对基因的功能、表达和调控的理解,对细胞活动及器官、系统、整体活动的调控都很关键。

2. 生物物理学

生物物理学其实是物理学的一个分支,研究的是生物的物理形态,涉及生物能学、细胞结构生物物理学、电生理学等。但这方面的生物数据获取和分析也越来越依赖于计算机的应用,如模型的建立、光谱和成像数据的分析等。

3. 脑和神经科学

脑是自然界中最复杂的组织,长期以来,通过神经解剖、神经生理、神经病理和临床医学研究,获得了大量有关脑结构和功能的数据。近年来,神经生物学研究也取得了大量科研成果,但是这些研究大多是在组织、细胞和分子水平进行的,不能很好地在系统和整体水平上反映人脑活动的规律。随着核磁共振成像和正电子发射断层成像的发展,应用计算机技术,使得我们有可能在系统和整体水平上无创地研究人脑的功能定位、功能区之间的联系及神经递质和神经受体等。由此产生的神经信息学研究将对我们了解脑、治疗脑和开发脑产生重大的作用。

4. 医药学

人类基因组计划的目的之一就是找到人类基因组中的所有基因。如何筛选分离各疾病的致病基因,获得疾病的表型相关基因信息的工作才刚开始。我们需要在现有的基因测序的工作平台上,强化生物信息学平台的建设,从而加快对突发性疫情、公共卫生的监控,以及对致病源进行快速有效的分析和解决。此外,结合生物芯片数据分析,确定药物作用靶,再利用计算机技术进行合理的药物设计,将是新药开发的主要途径。

5. 农林牧渔学

基因组计划也加快了农业生物功能基因组的研究,加快了转基因动植物育种所需生物信息学研究的步伐。通过比较基因组学、表达分析和功能基因组分析识别重要基因,为培育转基因动植物、改良动植物的质量和数量性状奠定了基础。通过分析病虫害、寄生生物的信号受体和转录途径组分,进行农业化合物设计,结合化学信息学方法,鉴定可用于杀虫剂和除草剂的潜在化学成分。此外,通过此方法可以进行动植物遗传资源研究,保护生物多样性;还可以对工业发酵菌进行代谢工程的研究,有目的地控制产品的生产。

6. 分子和生态进化

另一个重要的研究对象就是分子和生态进化。通过比较不同生物基因组中各种结构成分的异同,可以大大加深我们对生物进化的认识。从各种基因结构与成分的进化、密码子使用的进化,到进化树的构建,各种理论上和实验上的课题都等待着生物信息学家的研究。

第三节　生物信息学的主要应用

一、生物信息学数据库

生物信息学很大一部分工作体现在生物数据的收集、存储、管理与提供上,包括:建立国际基

本生物信息库和生物信息传输的国际联网系统；建立生物信息数据质量的评估与检测系统；生物信息工具开发和在线服务；生物信息可视化和专家系统。

比较著名的与生物有关的数据资源有 NCBI、EMBL、KEGG 等。

（一）数据库建设

生物数据库的建设是进行生物信息学研究的基础，尽管目前已有许多公共数据库可供使用，如 GenBank，且它们还同时集成开发了相应的生物分析软件工具，如 NCBI 的 BLAST 系列工具（http://www.ncbi.nlm.nih.gov/BLAST/）。但我们进行专项研究时，往往需要组建新的数据库。建立自己的数据库，就必须分析数据库的储存形式和复杂程度，选择怎么样的数据库，怎么开发信息交流平台，要不要提供相应的分析程序，甚至要不要将各搜索算法硬件化，实行并行计算、显卡处理器（GPU）计算和先进的内存管理以提高速度等。还要考虑到数据库的价格，像 Oracle（http://www.oracle.com）这样大型的数据库比较昂贵，MySQL（http://www.mysql.com）免费但功能可能满足不了要求。目前看来，基于 UNIX 开发的共享数据库 PostgreSQL（http://www.postgresql.org）可能是个不错的选择，此外还可以考虑用 XML 数据库。如果要构建二级数据库，可能还要涉及其他多个数据库的整合和数据挖掘。

（二）数据库整合和数据挖掘

生物数据库覆盖面广，分布分散且异质。当根据一定的要求将多个数据库整合在一起提供综合服务、提供数据库的一体化和集成环境时，最简单的方法是用超级链接或进行拷贝再整理。但往往简单的链接并不能符合要求，再整理涉及数据下载和更新的问题，而且不是真正意义上的“整合”。目前使用较多的是联合数据库系统，它是 IBM 分布式数据库解决方案的重要组成部分，支持用户或应用程序在同一条 SQL 语句中查询不同数据库甚至不同数据库管理系统中的数据。也有直接基于 Internet 技术而进行远程查询，从而进行文本数据挖掘和再整理的。由于生物的分支学科较多，整合时还需从语义学的角度考虑不同数据库的一致性问题，其实这已经成为了通过标准查询机制来连接数据库的一大阻碍，Ontology 技术可能可以解决这一问题。

二、序列分析

（一）序列比对

生物信息学最基本的操作对象是核酸序列和氨基酸序列。

1955 年桑格（Frederick Sanger）完成了第一个蛋白质——牛胰岛素化学结构的测定。1977 年，他领导的研究小组再一次成功地测定了第一个噬菌体 ΦX174 全基因组 5386 个碱基对的核苷酸序列，并发明了快速测定 DNA 序列的新方法。此后，全世界生物科学研究进入了分子水平。在使用散弹法进行 DNA 测序时，完整的 DNA 链被打散为成千上万条长 600～800 个核苷酸的 DNA 片段，这些 DNA 片段的两端相互重叠，只有依照正确的顺序组合，才能还原为完整的 DNA 序列。对于较大的基因组，散弹法能够迅速地测定 DNA 片段的序列，但将它们组装起来的工作则相当复杂。由于现今几乎所有基因序列均由散弹法测定，基因重组算法是信息生物学研究的重点课题。

比较序列的目的是发现相似的序列，得到保守的区域，它们可能有功能、结构或进化上的关系。对于一个感兴趣的 DNA 或蛋白质序列，寻找到与它同源的序列是基本工作。目前已开发了很多的算法，其中 BLAST 或 FASTA 都是不错的算法。在此基础上开发的 PSI-BLAST 和

megaBLAST 等，针对不同情况有更好的性能。

（二）基因序列注释

越来越多的物种测序工作的开展，迫切需要全基因组的自动注释，这一直都是生物信息学的研究领域。Ensembl 是由 EBI 和 Sanger 研究院合作的一个项目，利用大型计算机根据已有的蛋白质证据来对 DNA 序列进行自动注释。自动寻找基因和调控元件的工作通常需要的步骤包括：翻译起始点和终止点的确定，潜在的阅读框、剪切位点的识别，基因结构的构建，各种反式和顺式调控原件的识别等。除此以外，转录起始位点和可变剪切体的鉴定等工作都可利用计算生物学方法从庞大的基因组数据中提取出生物学信息，把它注释并图形化显示给生物学家。

三、其他主要应用

（一）比较基因组学

各种模式生物基因组测序任务的陆续完成，为从整个基因组的角度来研究分子进化提供了条件。比较基因组学的核心课题是识别和建立不同生物体的基因或其他基因组特征的联系。利用比较基因组学方法可以研究不同物种间的基因组结构的关系和功能。发现基因组中新的非编码功能元件是很有前途的应用。起初，真核生物中基因预测依靠概率模型预测得到，该方法的缺点是会产生很多的假阳性。通过比较不同物种间的同源基因可以大大提高预测的精度和准度。例如，在人类基因预测上，老鼠的基因信息起到了很重要的作用。

（二）基因和蛋白质的表达分析

进入后基因组时代，高通量技术高速发展并得到广泛应用。多种生物学技术可以用于测量基因的表达，如微阵列、表达序列标签、基因表达连续分析、大规模平行信号测序、多元原位杂交法等。所有这些方法均严重依赖于环境并能产生大量高噪声的数据，而生物信息学致力于发展一套统计学工具，以从中提取有用的信息。

通过蛋白质微阵列技术或高通量质谱分析对生物标本进行测量所获得的数据中，包含有大量生物标本内蛋白质的信息，生物信息学被广泛地应用于这些数据的分析。对于前者，生物信息学所面临的问题与 RNA 微阵列数据分析中遇到的问题相似；对于后者，生物信息学将所获得的大量质谱数据与通过已知蛋白质数据库预测的数据进行比较，并使用复杂的统计学方法进行进一步分析。

（三）生物芯片大规模功能表达谱的分析

生物芯片因为其具有高集成度、高并行处理能力及可自动化分析的优点，可对不同组织来源、不同细胞类型、不同生理状态的基因表达和蛋白质反应进行监测，从而获得功能表达谱。此外，生物芯片还可进行 DNA、蛋白质的快速检测及药物筛选等。由此可见，无论是生物芯片还是蛋白质组技术的发展都更强烈地依赖于生物信息学的理论与工具。鉴于生物芯片固有的缺陷及实验重复性等问题，以及有关表达谱的分析还不很精确，仍需大量的工作来提高对斑点图像处理的能力和系统的分析。

近年来，随着第二代测序技术的使用，人们已普遍运用 RNA-Seq 技术来进行大规模转录组表达谱的分析（第十二章）。

（四）蛋白质结构的预测

蛋白质结构的预测是生物信息学最重要的任务之一。蛋白质的一级结构决定其高级结构，

而后者又决定着它的生物学功能，目标是通过氨基酸序列来预测出蛋白质的三维空间结构。这方面的用途在医药工业上特别突出，如药物设计、设计各种特殊用途的酶等。对于序列同源性大于25%的蛋白质，可以使用比较同源模建的方法预测蛋白质结构，如 SWISS-MODEL 和 Modeller 软件。对于没有合适的模板的蛋白质预测可以使用折叠识别方法。折叠识别方法尝试寻找该目标序列可能适合的已知的蛋白质三维结构。如果前两种方法都无效，则要从头预测（*de novo* modeling），它的缺点是计算量大、耗时，而且仅适用于长度为几十个氨基酸的蛋白质片段，因此该方法目前主要作为前两种基于模板预测法的补充。整体来看，蛋白质结构预测领域还有待发展。

（五）蛋白质与蛋白质相互作用

蛋白质与蛋白质互作涉及蛋白质分子间的联系，而这种联系与生化反应、信号转导、各种网络都有关系。生物学的实验技术有很多种，如免疫共沉淀法、荧光共振能量转移、双分子荧光互补技术，生物学实验的方法往往繁琐且耗时。利用计算机技术有望基于蛋白质的各种性质，如理化性质、初级结构、三维结构等，来对蛋白质互作进行预测，但目前来看，这方面的工作还有很长的路要走。

（六）生物系统模拟

生物体是个复杂的系统，整个系统可以分成多个亚系统。现在的生物学家越来越清楚地认识到网络涉及生物的方方面面，从而兴起了一个新概念——系统生物学。Leroy Hood 认为系统生物学是确定、分析和整合生物系统在遗传或环境的扰动下所有内部元件间相互作用关系的一门学科。模拟生物系统对于更好地理解生命的本质活动至关重要。细胞水平下的代谢网络、信号转导通路、基因调控网络的构建，以及分析和可视化工作都给生物信息学带来了挑战。另外，人工生命或虚拟进化的研究往往致力于通过计算机模拟简单的生命形式来理解进化过程。

（七）代谢网络建模分析

代谢网络涉及生化反应途径、基因调控及信号转导过程（蛋白质间的作用）等。后基因组时代将研究大规模网络的生命过程，又称为“网络生物学”研究。

1. 预测调控网络

尽管目前已有多个代谢网络途径数据库，有些数据可以直接参考使用，而且这些数据库本身除了手工和自动检索文献以补充数据外，也有开发预测工具的，但是都有局限性和准确性的问题，还需要从基因组来预测网络，或有针对性地去整合某些数据，研究其规律，开发算法模型等。已有若干研究小组从事“基因组到代谢网络”的预测。

2. 网络普遍性分析

构建调控网络之后，人们对网络的“图论”方面的属性作了分析，如最短距离、连接度等，试图给出一些重要的结论；也有分析其最小单元的代谢途径等。越来越多的人开始开发专门的软件工具来自动分析大规模网络系统的物理属性，提供路径导航、模式搜索、图形简化等分析手段。

3. 建立模型分析

目前已有若干个比较优秀的代谢网络建模工具，如 Gopasi（http://www.Copasi.org）、E-cell（http://www.e-cell.org）等，它们大都基于代谢控制分析原理，使用常微分方程来求解反应速率。基于标准化数据输出输入考虑，已经组成了合作组，共同支持 SBML（http://www.sbml.org）数据交换。其他形式的建模工具也很多，如用随机方法处理的，因为毕竟确切的动态