863

生物高技术丛书

生物信息学

赵国屏 等 编著

科学出版社

"863"生物高技术丛书

生物信息学

赵国屏 等 编著

科学出版社

2002

内 容 简 介

本书是“863”生物高技术丛书之一。生物信息学是一门新兴学科，它以获取、加工、储存、分配、分析和释读生物信息为手段，综合运用数学、计算机科学和生物学工具，以达到理解数据中的生物学含义的目的。本书力求从各个重要的角度反映生物信息学今天的面貌：比较全面地介绍了生物信息学的若干个主要分支，并特别介绍了与人类基因组研究相关的生物信息学的一些较新的成果；着重介绍了数据库和数据库的查询、序列的同源比较及其在生物进化研究中的应用；以生物芯片中的生物信息学问题为例，介绍与基因表达相关的生物信息学问题；还介绍了蛋白质结构研究中的生物信息学问题，以及与分子设计和药物设计相关的生物信息学技术。

本书可供生物信息学专业和生命科学相关专业的本科生、研究生和教学科研人员阅读学习，也可供相关的科技和应用机构的科研、管理和决策人员参考。

图书在版编目(CIP)数据

生物信息学/赵国屏等编著. —北京：科学出版社，2002.4
（“863”生物高技术丛书）
ISBN 7-03-009895-1

Ⅰ.生… Ⅱ.赵… Ⅲ.生物信息学 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字（2001）第 089676 号

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
新蕾印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2002 年 4 月第 一 版 开本：787×1092 1/16
2002 年 4 月第一次印刷 印张：13
印数：1—4 000 字数：280 000
定价：26.00 元
（如有印装质量问题，我社负责调换〈北燕〉）

“863”生物高技术丛书编辑委员会

《生物信息学》编辑委员会

主　编：赵国屏

编著者（按姓氏汉语拼音排序）：

陈　军　陈凯先　陈润生　顾红雅　蒋华良

来鲁华　李亦学　陆祖宏　罗静初　罗小民

孙　啸　王金玲　赵国屏

丛书序 I

生物技术是20世纪末期，在现代分子生物学等生命科学的基础上，发展起来的一个新兴独立的技术领域，已被广泛应用于医疗保健、农业生产、食品生产、生物加工、资源开发利用、环境保护，对农牧业、制药业及其相关产业的发展有着深刻的影响，成为全球发展最快的高技术之一。在近20余年的时间里，各种生物新技术不断涌现。20世纪70年代创建了重组DNA技术和杂交瘤技术之后，动植物转基因技术、细胞大规模培养技术，以及近几年的基因组学、蛋白质组学、生物信息学、组合化学、生物芯片技术和自动化药物筛选技术等相继发展起来。可以说，生物技术的范围在不断地扩展，进入了蓬勃发展的新阶段。

我国的生物技术在“国家高技术研究与发展(863)计划”的支持下，经过全国生物技术科技人员15年的努力拼搏，在农业生物技术和医药生物技术的研究和开发方面都取得了很大的进展。一方面，我们在研究上取得了一批国际影响的创新成果，并获得一批拥有了自己知识产权的专利；另一方面，在开发上已有一批生物技术产品进入市场，还有相当一批产品正在研究开发中；海洋生物技术和环境生物技术也已起步。目前，生物技术研究和产业化已引起了全社会的关注，并将成为我国21世纪的一个新兴支柱产业。

在辞别20世纪，迈入21世纪之际，“863”计划生物领域专家委员会回顾我国生物技术发展历程，展望生物技术发展前景，编写了“‘863’生物高技术丛书”。借此机会，我希望所有从事生物技术研究和开发的科技人员，要进一步团结拼搏，增强创新意识，注重成果转化，为我国生物技术不断发展壮大做出新的贡献！

2000年7月15日

丛书序 Ⅱ

生物技术是20世纪末人类科技史中最令人瞩目的高新技术，为人类解决疾病防治、人口膨胀、食物短缺、能源匮乏、环境污染等一系列问题带来了希望。国际上科学家和企业家公认，信息技术和生物技术是21世纪关系到国家命运的关键技术和作为创新产业的经济发展增长点。

生物技术是指有机体的操作技术。它从史前时代起就一直为人类所开发利用，造福于人类。在我国的悠久历史中，传统的生物技术在民族经济的发展中一直起重要作用，特别是农业。据传，在石器时代的早期，神农氏曾传授人民如何种植谷物，并实行轮作制度；在石器时代的后期，我国早就善于酒精发酵；在公元前221年的周代后期，我国就能制作豆腐并酿制酱油和醋，其所用基本技术沿用至今。公元前200年，在我国最早的诗集——《诗经》中就提到过采用厌氧菌进行亚麻浸渍处理。早在16世纪，我国的医生就知道，被疯狗咬可以传播狂犬病。公元10世纪，就有了预防天花的活疫苗，到了明朝(1368～1644)，这种疫苗就广泛用于大量人群接种，此后，这种疫苗接种技术通过有名的丝绸之路传入欧洲国家。

1953年Watson和Crick提出了脱氧核糖核酸(DNA)的双螺旋结构模型，阐明了它是遗传信息的携带者，从而开辟了现代分子生物学的新纪元。DNA分子是所有生命机体发育和繁殖的蓝本。众所周知，一切生命活动主要是蛋白质的功能，而蛋白质是由基因编码的。60年代初就破译了“遗传密码”。生命现象千姿百态，但生命体的本质却有高度的一致性。它们的蛋白质都是由20种氨基酸以肽键连接而成，核酸都由4种核苷酸以磷酸二酯键构成，其遗传密码在整个生物界也基本一致。于70年代，科学家们发展了一种新技术，也就是众所周知的DNA重组技术。它向人们提供了一种手段，人们可以在试管内，根据人们的意愿来操作基因、改造基因，新的基因信息可以转入一种简单的生命体中，如大肠杆菌，或转入另一种机体，借以提供一种手段来改造谷物和家畜品种，或生产有效药物，制作疫苗和一系列自然蛋白质，或进行基因治疗。显然，新生物技术是一场革命，是生产力的一次解放，被认为是20世纪人类的一项最伟大贡献，它必将深刻地促进世界经济的发展。

广义的新生物技术包括基因工程、细胞工程、发酵工程和酶工程，但新技术的核心是基因工程技术，它能带动其他生物技术的发展，最具有革命性。

近20年来，国际上生物技术飞跃发展，特别是基因操作技术、生物治疗技术、转基因动植物、人类和其他生命体基因组工程、基因治疗技术、蛋白质工程技术、生物信息技术、生物芯片技术等。生物技术的创新正在带动着生物技术巨大产业的发展，它包括基因药物、重组疫苗、生物芯片、生物反应器、基因工程抗体、基因治疗与细胞治疗、组织工程、转基因农作物、兽用生物制品、生物技术饲料、胚胎移植工程、基因工程微生物农药、环保、海洋生物，以及现代生物技术对发酵、制药、轻工食品等传统产业的改造等领域。

目前，生物技术产业与信息产业相比较还处于发展初期，至1998年全世界共有生物

技术公司3600余家，主要集中在美国和欧洲，其中年产值超过10亿美元的有约20家。生物技术产业在20年中市场总值增加了50多倍；涨幅最快是在近10年，例如美国在1980年生物技术产品的销售额还处于零增长，1991年达到59亿美元，1996年为101亿美元，1998年增至147亿美元；目前，生物技术仍保持25%左右的增长速度，20%左右的融资率和12.5%就业增长率以及8.76%平均股市涨幅。另一方面，也要看到，美国的1300余家生物技术公司中上市公司为300家，而赢利的公司约为20家，这是由于生物技术产品的研究和开发周期较长，因此从整体看生物技术产业还处在投入阶段。从另一方面来看，尽管美国公司的赢利公司不多，但赢利公司的数量却在稳步上升。

1999年全球生物技术产品的总销售额约为500亿美元，而产生的间接经济效益超过3000亿美元，全球有一半以上的人直接享用过生物技术产品。其主要产品为医药产品、农产品和食品。

我国自1986年实施"863"计划以来的15年中，现代生物技术的开发研究与产业化进入飞速发展阶段：二系法杂交稻的开发与推广对我国的粮食增产起了重要作用，2000年已推广5000万亩以上。1993年我国第一例转基因作物抗病毒烟草进入了大田试验，1997年第一例转基因耐贮存番茄获准进行商品化生产，至1999年5月共有6种转基因作物投放市场。2000年我国转基因抗虫棉花种植面积超过550万亩。1990年我国研制了第一例转基因家畜，1991年山羊克隆获得成功，生物技术饲料添加剂已经实现了规模化生产。我国自1989年第一种基因药物——重组α1b干扰素获准投放市场以来，至1999年我国已有18种基因药物和疫苗获准进行商业化生产，另有26种基因药物处于临床前或临床Ⅰ、Ⅱ期试验，我国生物技术医药产业已初具规模。我国已列为人类基因组计划国际大协作的成员国，承担完成1%的任务，美、英、日、法、德、中科学家于2000年6月26日宣布人类基因组全部DNA序列的工作框架图已经完成。我国在国际上首先发现神经性耳聋的基因，基因治疗已有4个项目进入临床试验阶段，生物芯片技术的开发研究与产业化正在与国际上同步发展。15年来我国在生物技术领域中取得的成就是举世瞩目的，同时还培养了一大批中青年科技人才，为21世纪初"S-863"计划的实施和生物高技术产业化奠定了扎实的基础，也将为21世纪初我国的经济建设做出应有的贡献。

本丛书是在科学技术部中国生物工程开发中心、"863"计划生物技术领域专家委员会的领导下，由在第一线从事"863"生物高技术研究与开发的科技人员撰写的系列丛书。本丛书包括了农、医生物技术的各个方面，不仅基本上概括了近10年来国际上的研究进展和发展趋势，而且还全面反映了我国"863"计划实施15年来在生物技术领域取得的进展和成果。本丛书的出版无疑将进一步推动我国生物技术开发研究和产业化的进程，促进我国经济的持续发展。同时，本丛书也是培养新一代青年生物技术科学家的重要教科书。

2000年1月16日

前　言

生物信息学（bioinformatics）是一门新兴的交叉学科。它所研究的材料是生物学的数据，而它进行研究所采用的方法，则是从各种计算技术衍生出来的。在历史上，生物信息学也曾经被称为“计算生物学”。随着基因组研究的日益深入，生物学数据积累出现了前所未有的飞跃。首先，数据增长的速度之快，已经只有计算机芯片计算能力的增长能与之相匹配（Moore 定律，每 18 个月翻一番的指数增长）；其次，数据的本质出现了从生理生化数据向遗传信息飞跃以及进一步向遗传与结构功能相互关系信息的飞跃。因此，基因组研究启动以来的十年，是生物学研究真正从往日的以描述、定性研究为主的“经典”模式中脱胎，逐步进入以机制、定量研究为主的“信息生物学”模式的十年，是生物信息学技术不断发展的十年。

我国生物信息学的研究和应用最早应追溯到分子生物学时代和计算机时代之前在生物统计方面进行的工作，譬如群体遗传学方面的工作。虽然这方面的工作具有极大的发展潜力，但是，没有分子生物学提供遗传学研究的工具，没有现代的计算机和计算技术提供数据处理的平台，这些工作只能停留在理论模建的阶段。“文化大革命”之后，随着分子生物学特别是蛋白质晶体结构解析能力的提高和蛋白质工程技术的发展和应用，在国家“863”计划等高科技计划的支持下，以蛋白质分子结构的计算及模拟为代表的“计算生物学”技术在我国有了一定的发展。进入 20 世纪 90 年代后期，随着基因组研究在我国的蓬勃发展，我国科学工作者不失时机地开始发展基因组信息技术。应该说，在过去的五年中（第九个五年计划期间），我国基因组信息技术的发展，特别是普及的速度是前所未有的。本书的出版，从一个侧面反映了我国科学家在这方面努力的成果。

生物信息学不仅是一门新兴的学科，随着基因组研究的发展，它又是一门覆盖面极广的综合性学科。本书力求从各个重要的角度反映生物信息学今天的面貌。第一章导论，除比较全面地介绍了生物信息学的各个分支外，强调了与人类基因组研究相关的生物信息学的一些较新的成果。第二、三章着重介绍了数据库和数据库的查询，这是生物信息学和生物信息技术的基础。第四章着重介绍序列的同源比较及其在生物进化研究中的应用，这是今天的实验生物学家运用最为普遍的生物信息技术。第五章以生物芯片中的生物信息学问题为例，介绍与基因表达相关的生物信息学问题，可以预见，随着大规模基因表达谱和蛋白质组研究的发展，这一内容将获得更为广泛的关注。第六章介绍蛋白质结构研究中的生物信息学问题，这些问题对于研究生物分子的结构与功能关系的读者一定是有吸引力的。第七章介绍与分子设计和药物设计相关的生物信息学技术，这一点可能是今后生物信息学应用研究中最为吸引人的部分之一，也是我国今后生物信息学发展的一个重要方面。

遗憾的是，生物信息学的许多重要组成部分未能在本书中得到反映，这固然与本人的能力有限有关，也与我们的一些科学家工作繁忙，无暇顾及写作有关。好在本书只是旨在对生物信息学作一般性的介绍，读者如果通过阅读本书，感觉到生物信息学的重

要，并对生物信息学研究的入门有一定的认识，本书的作者们也就感到是完成了任务。

本书的作者们都是在科研第一线从事生物信息学或与生物信息学相关研究的科学家。我对于他们在百忙中完成这一写作任务表示深切的感谢！由于时间限制，我们写作和编辑中难免有错误或问题，希望得到同行们的批评和指正。

我国生物学家正在积极参与基因组的各个层次上的研究工作，他们对发展生物信息学研究、应用生物信息技术具有强烈的需求。另一方面，我国又有特别优秀的物理学和数学基础，我国已经有一批物理学家和数学家积极地投入了生物信息学的研究。因此，生物信息学的研究在我国有望取得突破性成果，这对于增强我国在基础研究领域的实力，在某些方面占据国际领先地位是十分重要的。生物信息学成果的应用也会产生巨大的社会效益和经济效益，为实现我国的社会发展、人民幸福、国家富强贡献力量。本书作者们愿与读者们一起努力，为开创生物信息学发展的大好局面而继续努力。

赵国屏

2001 年 9 月 17 日

目　录

丛书序 I
丛书序 II
前言
第一章　生物信息学：导论 …… (1)
一、什么是生物信息学？ …… (1)
二、生物信息学的研究现状与发展趋势 …… (4)
三、生物信息学的生物学内涵 …… (6)
（一）基因与基因组的信息学 …… (7)
（二）基因表达的信息学：大规模基因功能表达谱的分析 …… (20)
（三）生物大分子的三维结构信息：蛋白质结构模拟与分子设计 …… (21)
（四）代谢和疾病发生途径的信息 …… (22)
四、生物信息学的信息学内涵 …… (23)
（一）生物信息数据库 …… (23)
（二）分析工具的发展 …… (26)
五、生物信息学的应用与发展研究 …… (26)
（一）与疾病相关的基因信息及相关算法和软件开发 …… (26)
（二）建立与动、植物良种繁育相关的基因组数据库，发展分子标记辅助育种技术 …… (27)
（三）研究与发展药物设计软件和基于生物信息的分子生物学技术 …… (27)
六、生物信息学研究和发展中的交叉学科和大科学特点 …… (28)
（一）实验生物学家和计算生物学家 …… (28)
（二）三种科学文化的融合 …… (29)
（三）跨越整个生命科学的大科学 …… (29)
第二章　分子生物信息数据库 …… (33)
一、分子生物信息数据库简介 …… (33)
（一）基因组计划和数据库 …… (33)
（二）分子生物信息数据库种类 …… (37)
二、基因组数据库 …… (41)
（一）GDB …… (42)
（二）AceDB …… (42)
三、序列数据库 …… (43)
（一）核酸序列数据库 …… (43)
（二）EMBL 和 GenBank 数据库格式 …… (44)
（三）常用蛋白质序列数据库 …… (50)

(四) 其他蛋白质序列数据库 …… (56)
四、结构数据库 …… (57)
(一) 蛋白质结构数据库 PDB …… (58)
(二) 蛋白质结构分类数据库 SCOP 和 CATH …… (64)
五、二次数据库 …… (66)
(一) 基因组信息二次数据库 …… (66)
(二) 蛋白质序列二次数据库 …… (68)
(三) 蛋白质结构二次数据库 …… (71)
第三章 数据库查询和数据库搜索 …… (75)
一、简介 …… (75)
二、数据库查询系统 Entrez …… (75)
(一) Entrez 系统使用方法 …… (76)
(二) Entrez 系统的特点 …… (80)
三、数据库查询系统 SRS …… (80)
(一) SRS 系统使用方法 …… (81)
(二) SRS 系统的特点 …… (84)
四、数据库搜索简介 …… (85)
(一) 核苷酸碱基和氨基酸残基代码表 …… (85)
(二) 相似性和同源性 …… (86)
(三) 局部相似性和整体相似性 …… (87)
(四) 相似性计分矩阵 …… (88)
五、数据库搜索工具 BLAST …… (89)
(一) 程序简介 …… (89)
(二) BLAST 程序运行实例 …… (92)
第四章 序列的同源比较及分子系统学和分子进化分析 …… (93)
一、简介 …… (93)
二、相似序列的获得 …… (93)
(一) BLAST …… (94)
(二) 与 BLAST 相关的一些知识 …… (97)
(三) 获得同源序列的其他方法 …… (99)
三、多序列比对 …… (101)
四、系统发育分析 …… (104)
(一) 系统树的构建方法 …… (105)
(二) 常用的系统树构建程序 …… (107)
(三) 一些需要注意的问题 …… (111)
(四) COG 数据库 …… (111)
五、其他分子标记在生物系统学中的应用 …… (112)
(一) RFLP (restriction fragment length polymorphism) 标记 …… (113)
(二) PCR 扩增片段长度的多样性 …… (113)

（三）SNP 标记 …… （114）
（四）同工酶 …… （115）
第五章　生物信息学与基因芯片 …… （118）
一、概述 …… （118）
（一）基因芯片简介 …… （118）
（二）基因芯片对于生物分子信息检测的作用和意义 …… （121）
（三）基因芯片研究和应用中所涉及到的生物信息学问题 …… （123）
二、基因芯片设计及优化 …… （124）
（一）基因芯片设计的一般性原则 …… （124）
（二）DNA 变异检测型芯片与基因表达型芯片的设计 …… （126）
（三）cDNA 芯片与寡核苷酸芯片的设计 …… （126）
（四）寡核苷酸探针的优化设计 …… （127）
（五）基因芯片的优化 …… （129）
三、基于芯片的序列分析 …… （129）
（一）测定未知序列 …… （129）
（二）直接检测目标序列 …… （130）
（三）DNA 序列突变检测分析 …… （130）
（四）SNP 分析 …… （131）
四、基于芯片的基因功能分析 …… （133）
（一）基因表达分析 …… （133）
（二）高密度基因表达芯片 …… （133）
（三）基因表达图谱 …… （134）
（四）寻找基因功能 …… （135）
五、基因芯片检测结果的分析 …… （135）
（一）荧光检测图像处理 …… （135）
（二）检测结果分析 …… （136）
（三）检测结果可靠性分析 …… （136）
六、基因芯片信息的管理和利用 …… （136）
（一）基因芯片信息管理 …… （136）
（二）数据集成和交叉索引 …… （137）
（三）数据的可比性和归一化问题 …… （138）
（四）基因芯片信息的利用 …… （138）
七、基于基因芯片的数据挖掘及可视化 …… （138）
（一）数据挖掘 …… （138）
（二）基因芯片的多元数据结构 …… （139）
（三）数据相似程度的量化与距离矩阵 …… （140）
（四）聚类分析 …… （140）
（五）聚类分析结果的树图表示 …… （143）
（六）基因芯片数据的可视化和与数据库的链接 …… （143）

八、基因转录调控网络分析 …… (144)
(一) 布尔网络模型 …… (144)
(二) 线性组合模型 …… (145)
(三) 加权矩阵模型 …… (145)
(四) 互信息关联网络 …… (146)
第六章 蛋白质结构预测的原理与方法 …… (149)
一、引言 …… (149)
二、影响蛋白质折叠的因素 …… (150)
三、蛋白质结构分析及蛋白质结构数据库 …… (151)
(一) 有关氨基酸残基的信息 …… (151)
(二) 周期性的二级结构 …… (151)
(三) 非同期性的二级结构 …… (152)
(四) 残基间的相互作用及埋藏 …… (152)
(五) 超二级结构 …… (152)
(六) 蛋白质结构数据库 …… (152)
(七) 蛋白质结构域的折叠模式与蛋白质结构分类数据库 …… (153)
(八) 蛋白质的进化 …… (157)
四、二级结构预测 …… (158)
(一) 二级结构预测概况 …… (158)
(二) Chou-Fasman 方法 …… (158)
(三) GOR 方法 …… (158)
(四) 最近邻居方法 …… (159)
(五) 神经网络方法 …… (159)
(六) 基于多重序列比对的二级结构预测 …… (159)
(七) 二级结构预测的准确度 …… (159)
(八) 二级结构在线预测 (online prediction) …… (160)
五、三级结构预测 …… (160)
(一) 同源蛋白质结构预测 …… (160)
(二) 蛋白质折叠类型识别 …… (161)
(三) 蛋白质结构从头预测 …… (164)
六、蛋白质结构预测发展趋势 …… (165)
第七章 生物信息学与药物设计 …… (168)
一、当代生物医药研究所面临的困难 …… (168)
二、现代生物学给生物医药带来的发展契机 …… (168)
三、基因组学、蛋白质组学和生物信息学在药物研究中的应用 …… (169)
(一) 选择药物作用靶标的标准 …… (170)
(二) 候选药物作用靶标的发现 …… (171)
(三) 靶标有效性的验证 …… (174)
(四) 药物作用机制的研究 …… (175)

（五）药物的药代动力学及毒理性质的研究 …… （176）
四、计算机辅助药物设计 …… （177）
（一）间接药物设计 …… （177）
（二）直接药物设计 …… （178）
（三）药物设计实例 …… （179）
五、未来药物研究方法展望 …… （184）
（一）人类基因组和生物信息学的发展，将为药物设计研究开辟更广阔的空间 ……
…… （184）
（二）超级计算机的发展将为复杂生物体系的理论计算和药物设计创造有利的条件
…… （184）
（三）计算机辅助药物设计与组合化学技术相结合将显示巨大威力 …… （185）
（四）基于结构的药物设计将向基于作用机制的药物设计方向发展 …… （185）

第一章

生物信息学：导论

世纪之交，人类基因组计划已经取得了决定性的成功。自全长 1.8Mb 的流感嗜血菌（*Haemophilus influenzae* Rd）基因组序列于 1995 年发表（Fleischmann et al., 1995）以来，数十种模式生物，包括大量微生物（数据参阅：http://www.tigr.org/）以及酵母（Goffeau et al., 1997）、线虫（The *Caenorhabditis elegans* Sequencing Consortium, 1998）、果蝇（Myers et al., 2000）和拟南芥（The Arabidopsis Genome Initiative, 2000）等真核生物的基因组全序列相继公布，至 2001 年的春天，科学家已经公布了人类自身基因组的绝大部分序列（International Human Genome Sequencing Consortium, 2001；Venter et al., 2001）。这一成就意味着，从新世纪开始，人类基因组的研究将全面进入信息提取和数据分析的崭新阶段；人类的遗传语文将被逐渐释读出来，功能基因组和蛋白质组的研究将广泛开展，遗传信息与生物体代谢、发育、分化、进化之间的关系将逐步被人类所认识。当然，在这个已经积累了足够多的基因组数据，但许多重大规律尚未被发现的时候，生物信息学的研究正面临着严峻的挑战和千载难逢的机遇；生物信息学学科正处在它发展成长的关键时刻。因此，普及生物信息学的知识，让更多的科学工作者了解、关心这门科学，善于运用这门科学研究的成果，积极参与这门科学的发展，是必要的、有意义的。本书的编写正是出于这样的一个目的。由于出版过程中的时间差，本章的内容，部分已在本丛书的《基因组科学与人类疾病》一书中关于生物信息学的一章中阐述了。但是，在本书出版时，我们特别依据人类基因组研究工作的进展对内容作了较大的更新，并改进若干行文的方式，使之更符合本书的要求。

一、什么是生物信息学？

生物信息学（bioinformatics）是一门新兴的交叉学科。它所研究的材料是生物学的数据，而它进行研究所采用的方法，则是从各种计算技术衍生出来的（Benton, 1996）。

20 世纪 50 年代，DNA 双螺旋结构的阐明开创了分子生物学的时代。以生物学和医学为主要研究内容的生命科学研究从此进入了前所未有的高速发展的阶段。分子生物学和遗传学的文献积累从 60 年代中期的接近 10 万篇迅速增长至 60 年代末期的 20 多万篇，即在 3～4 年间，翻了一番。此后，至 80 年代中期，上升至约 30 万篇，即平均每年增长 6000～7000 篇。至 90 年代中期，文献数已上升至 40 多万篇；即在 10 年中，平

均每年增长 1 万篇。到 2000 年，则增长至约 50 万篇，即在约 5 年间，又增长了 10 万篇（根据http://www.ncbi.nlm.nih.gov 有关 PubMed 数据整理）。与此同时，更为大量的数据已经不再以传统的文献形式发表了；这里，最为典型的是 DNA 序列的数据。美国的核酸数据库 GenBank（Banson et al.，1998）从 1979 年开始建设，1982 年正式运行；欧洲分子生物学实验室的 EMBL 数据库也于 1982 年开始服务；日本于 1984 年开始建立国家级的核酸数据库 DDBJ，并于 1987 年正式服务。即是说，DNA 序列的数据已经从 80 年代初期的百把条序列、几十万碱基上升至 90 年代末的数十亿碱基及包括人类基因组在内的一大批（数百万）EST、cDNA、基因和基因组序列了。至 2000 年底，国际数据库中记录的接近一千万条 DNA 序列的碱基数已超过 100 亿！这就是说，在短短的约 18 年间，数据量增长了近十万倍（数据来自 http://www.ddbj.nig.ac.jp/ddbjnew/statistics-e.html 和 http://www.ebi.ac.uk/genomes/mot/）！事实上，在今天的一个大型的基因组测序中心，每天可进行十万个测序反应，产生出 10^7 的序列数据。参与人类基因组测序的公共部分合作的各国测序中心，自 1999 年 6 月开始进入大规模测序阶段，在短短的 8 个月内，测序能力上升了将近 8 倍。至 2000 年 6 月，这些中心在 6 个星期内的测序量就相当于一个人的基因组。也就是说，每周 7 天，每天 24 小时，每秒即可产出 1000 个碱基的数据！至 2000 年 10 月，从 230 亿（23Gb）的测序数据中产生的草图（draft/pre-draft）总数据数已达 45 亿碱基(4.5Gb)(International Human Genome Sequencing Consortium, 2001)。与之相应，自 90 年代以来，被鉴定的基因数据和被解析的蛋白质结构的数据也摆脱了以往缓慢增长的局面，达到了每两年增长一倍的速度（数据来自http://www.rcsb.org/pdb/holdings.html)。应该指出，由于人类基因组草图的完成以及注释工作的深入，人类疾病基因的定位克隆和鉴定已经大大加快了。同时，由于结构基因组工作的广泛开展，可以预见，蛋白质结构解析的速度也必然大大加快。

与上述生物学数据的海量特征相比，生物学数据的复杂特征则更具有挑战性。生物学数据的复杂性一方面固然是源于生物体的结构和功能以及生命活动过程本身的多样性和复杂性，另一方面则是由生物学研究的“社会学原因”所造成的（Benton，1996)。在生物学研究中，即便存在标准的方法，即便使用商业化的“试剂盒”，在具体实验中，这些方法也往往是以创新的方式或组合加以运用的，因而也就使整个实验方法具有了与众不同的独特性。从信息科学的角度看，除了少数例外，生物学的实验数据，一般是在既无标准词法（semantics)、又无句法（syntax）的条件下生成的。这一情况，必然进一步加剧生物学数据的复杂性。

生物学数据在量（海量）与质（复杂性）方面所提出的挑战是严峻的。事实上，在 20 世纪 80 年代中，人们对于是否应该进行人类基因组大规模测序的争论的一个重要焦点问题就是对于大规模测序所产生的数据进行处理和释读的能力的评估。十分幸运的是，在过去的二十多年里，电子计算机芯片对于数字处理的能力的增长基本符合 Moore 定律（指数增长)。每个 CPU 所含晶体管数从 70 年代初的几千个迅速而稳定地增长到 80 年代末的上百万个，即平均每 2 年翻一番；此后，至 90 年代末，又上升至上亿个，即平均每 2.5 年翻一番（数据来自：http://www.physics.udel.edu/wwwusers/watson/scen103/intel.html)。也就是说，如今的大型计算机的数据处理能力，已经发展到