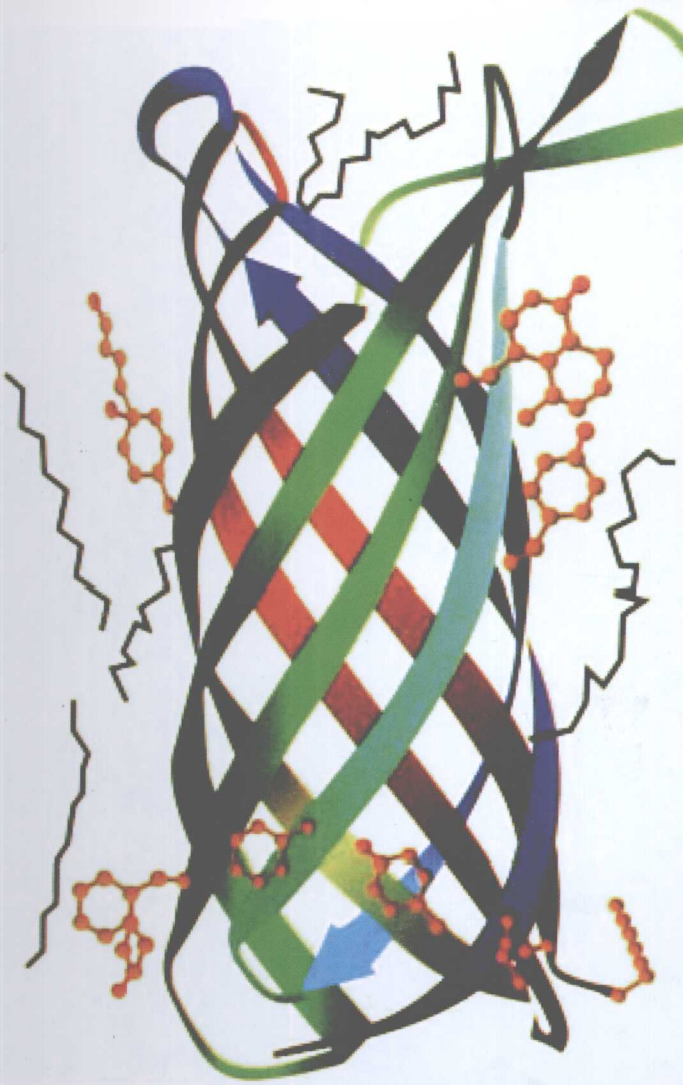
SHENGWU XINXIXUE

生物信息学

许忠能 主编

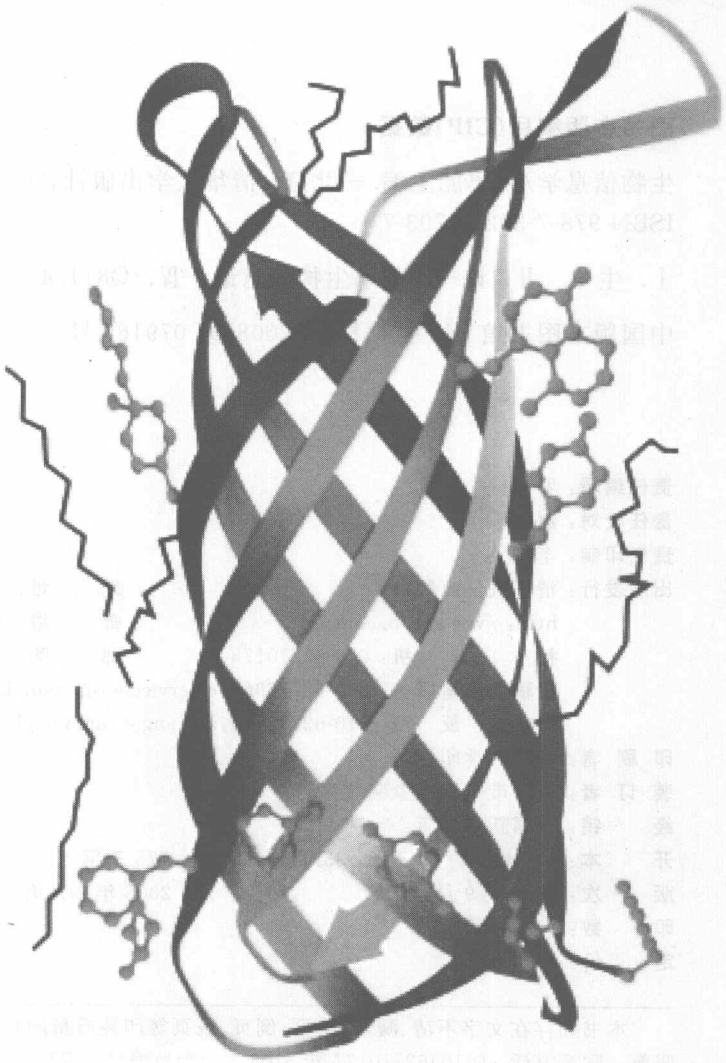


清华大学出版社

SHENGWU XINXIXUE

生物信息学

许忠能 主编



清华大学出版社
北京

内 容 简 介

生物信息学是一门新兴的交叉科学。本书共分16章，详细介绍了生物信息学的定义、研究内容、生物学基础、数据库网络基础、算法与数学基础以及其在序列拼接、基因预测、引物设计、生物进化与分子发育分析、蛋白质结构预测、RNA结构预测、生物芯片、计算机辅助药物设计、生物分子网络与生物系统仿真、DNA计算中的应用与发展状况等内容。

本书结构清晰，系统完整，文笔流畅，既可作为高等院校相关专业师生的教材，也可作为该领域中研究、教学、软件开发等科研人员的参考用书。

图书在版编目(CIP)数据

生物信息学/许忠能主编. —北京：清华大学出版社，2008.9
ISBN 978-7-302-17793-7

Ⅰ. 生…　Ⅱ. 许…　Ⅲ. 生物信息论　Ⅳ. Q811.4

中国版本图书馆CIP数据核字(2008)第079169号

责任编辑：罗　健
责任校对：赵丽敏
责任印制：李红英
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦A座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：清华大学印刷厂
装 订 者：三河市李旗庄少明装订厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：33.75　**字　数**：775千字
版　　次：2008年9月第1版　　**印　　次**：2008年9月第1次印刷
印　　数：1～3000
定　　价：59.00元

编 委 会

前　言

生物信息学融合了生命科学、信息科学等21世纪多个朝阳学科，越来越深入影响科学研究与社会生活的多个方面。开设生物信息学课程有利于提高学生科研理念及科技产业化意识。

许多学者已编写了不少优秀的生物信息学教材与专著，涵盖教学与科研的多个方面，但这些教材与专署往往仅适用于某一专业的学生。本书编写的目标是为不同专业背景的读者提供一本学习生物信息学的入门教科书。编者设计了大量的例子，便于读者理解相关原理。希望为推动国内生物信息学教学贡献绵薄之力。

2005年，许忠能博士在暨南大学教务处的谢绍潮老师的建议下及清华大学出版社罗健编辑的帮助下，以试用多个学期的生物信息学课程教案为蓝本编写了提纲，向在生物信息学及相关领域教学与科研第一线的若干学者发出邀请，希望他们加入本书编委会。令人振奋的是，Ken Chan博士、许龙飞教授、陈宣老师、黄蓓蓓博士生、刘顺会副教授、周杰博士、刘吉平副教授、周晖老师、石宏博士、蒙进芳老师欣然同意共同编写本书。Ken Chan博士参与编写了第2、9章的部分内容；许龙飞教授编写了第3章，同时参与编写了第4、5章的部分内容；刘顺会副教授编写了第14章；刘吉平副教授参与编写了第8、15章的部分内容；黄蓓蓓博士生编写了第7章，并参与编写了第4、6章的部分内容；陈宣老师参与编写了第4章的部分内容；周杰博士编写了第10章；周晖老师编写了第12章；蒙进芳老师参与编写了第13章的部分内容；石宏博士参与编写了第13章的部分内容；许忠能博士编写了第1、11、16章，并参与编写了第2、5、6、8、9、15章的部分内容，同时负责全书的统编、审校工作。暨南大学计算机系蔡利栋教授对第5章进行了审核。

本书得到暨南大学水生生物学国家重点学科相关基金的资助。编委会感谢暨南大学水生生物研究所韩博平教授对本书的大力支持，感谢中山大学邓日强副教授对本书编写过程中提供的帮助，感谢各编者所在单位的相关学生与老师、清华大学出版社、暨南大学教务处对本书的支持。同时，衷心感谢在此过程中，各编者的家人对编写工作的鼓励。

生物信息学涉及的领域众多，学科发展迅速，本书的内容难免出现错误，恳请各位读者批评指正。

本书编委会

2008年5月于暨南大学

目　录

第1章　生物信息学概述

本章提要：本章介绍了生物信息学的概况。生物信息学是一门交叉学科，它综合运用数学、计算机科学和生物学的各种工具，来阐明大量数据所包含的生物学意义。生物学、计算机技术、数学等学科是学习生物信息学基本理论及运用相关软件的基础。生物信息学研究的内容包括生物信息的存储与获取、序列比对、测序与拼接、基因预测、生物进化与系统发育分析、蛋白质结构预测、RNA结构预测、分子设计与药物设计、代谢网络分析、基因芯片、DNA计算等。生物数据库、相关软件是生物信息学研究与应用的重要资源。专门机构、期刊、著作在生物信息学的研究、教育、推广应用上起着重要的作用。生物信息学有非常好的发展前景及广泛的应用性，它给从事该专业的人有较高的回报。

生物信息学(bioinformatics)是一门新兴的交叉学科，生物学与医学、数学、计算机科学是其中三个主要组成部分。生物学家与医学家认为生物信息学主要是搞清生物学意义或医学意义，数学家认为算法与数学模型是其核心，计算机专业人员则觉得数据库及相关软件是生物信息学得以发展的基础。可见，这个学科涉及面广，其作用和价值在人类基因组计划的实现过程中得以充分体现。本章即介绍一下这个领域的概况。

1.1　背景与定义

1.1.1　生物学原始数据量的急剧扩增

在近代科学发展中，数学、计算机科学、生物学互相渗透、结合，用以解决许多重大的科学问题。生物学数据量剧增是近二十年来生物学出现的重大问题之一，它需要多学科结合去解决。

由于20世纪初遗传学、分子生物学、生物化学等生物学分支的研究异军突起，使众多学科纷纷在技术与理论上支持生物学研究。物理学领域的一位伟大的科学家Delbrück的著作《生命是什么》引领众多优秀的物理学家投入到生物学研究中去，并在其中产生了若干位诺贝尔生理医学奖得主。应用先进的物理学方法，科学家提出了生物大分子如蛋白质、核酸的结构模型，并最终得到验证。之后，对生物大分子的研究突飞猛进。在此基础上，人类对自身生物背景信息的探求欲望不断加深。1986年，诺贝尔生理医学奖得主Dulbecco在*Science*杂志上提出“人类基因组计划”。经过几年的论证，最终由美国政府出资30亿美元开展此计划，这是美国历史上第三个国家计划。

美国国家科学院人类基因组制图与测序委员会在1988年确定了测序计划及所需资金，经美国国会审批，人类基因组计划从1990年开始，历时15年，由美国国立卫生研究

所、美国能源部资助 30 亿美元来完成。该计划的目的是测定包括人类基因组在内的若干基因组序列，开发相关基因组序列分析技术及储存技术，研究该计划对社会、法律及伦理的影响。实际看来，美国出的 30 亿美元是高收益的，从该计划实施带来的影响使美国稳居现代生物学界领头羊的位置，多国生物学精英汇聚美国，同时该计划为美国培养了一批生物学顶尖人才，仅这些方面的价值就不止 30 亿美元了。由于这项计划对人类的未来意义重大，同时学术界也不希望重要的生物学技术被垄断，因此先后有 5 个国家加盟这项计划，当然这些国家的相关机构也出钱资助。这次计划最初测定的是其他生物的基因组，以寻求合适的方案对人类基因组测序。这些过程推动了 DNA 测序技术的发展，使 DNA 序列量急剧增长。在 2001 年 2 月，人类基因组草图公开发表。2003 年，人类基因组项目成果已达到设定的目标，因此美国政府宣布该项目已完成，比原计划提前了两年。然而测序工作及对测序结果质量的验证工作将在此后多年继续进行。

人类基因组计划的完成，部分得益于数学与计算机技术的广泛应用。如在艰巨的测序过程中，多种序列分析软件起到关键的作用，保证了测序顺利进行。大型国际公用的生物大分子序列数据库不断接受各测序中心提交的序列信息，数据量以指数形式上升。图 1-1 是世界三大序列数据库之一的 EMBL 在人类基因组计划期间核酸序列数量的增长情况。

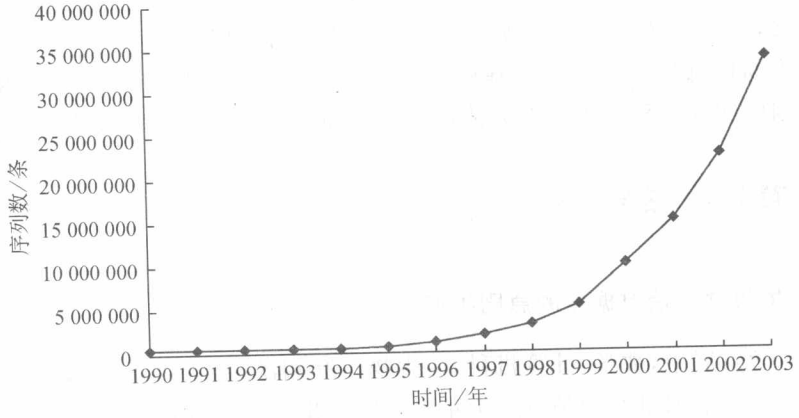


图 1-1 EMBL 数据库在人类基因组计划期间核酸序列数量增长情况

面对庞大的生物学数据，如何去分析、管理它们就变得非常重要。这时，传统的生物学研究手段已经束手无策，一门由生物学、数学、计算机科学紧密结合的技术——生物信息学应运而生，用以对付正在等待管理、分析与开发的海量生物学数据，就这样，生物信息学在人类基因组计划进行过程中产生不断发展壮大。

1.1.2 名词“bioinformatics”的第一次出现

生物信息学研究内容的源头较多，很难找到一个公认的标志性研究内容或研究论文作为该领域开创性的里程碑。有的学者觉得生物信息学的主要内容是计算分子生物学，分子生物学方面的事件左右着生物信息学，所以认为应将 1953 年 Watson 和 Crick 提出

DNA 双螺旋结构模型作为生物信息学的开端。而有的学者认为第一次公开提出将生物与信息结合的事件更有意义，所以 1956 年美国田纳西州的 Gatlinburg 召开的“生物学中的信息理论研究会”应被认为是揭开生物信息学序幕的事件。还有 20 世纪 60 年代 Dayhoff 等完成的《蛋白质序列与结构图册》、20 世纪 70 年代 Needleman 和 Wunsch 提出的序列比对算法、20 世纪 80 年代诞生的三大序列数据库等都是生物信息学的某些研究内容的开创性事件。然而这些事件中仍未提到目前生物信息学的专用名字——“bioinformatics”。

“bioinformatics”一词是由林华安(Hwa A. Lim)博士在 1987 年首创的。林华安博士出生于马来西亚，1981 年由英国伦敦大学帝国学院大学毕业，1986 年获美国 Rochester 大学博士学位。1987 年他到佛罗里达州立大学的超级计算机中心担任基因与生物物理组的主任，并于同年年底开始发表跨学科研究的论文。他认为信息学与生物学相结合是未来科学研究的一个潮流，并构思一个新的名词为这个新学科命名。1987 年，他最终敲定“bioinformatics”一词作为这个新学科的名字。1990 年，林华安博士组织了第一届生物信息学与基因组研究国际会议(Bioinformatics and Genome Research International Conference)，并担任会议主席。此后 10 年，他连续担任该大会主席，于是“bioinformatics”一词也在学术界更加深入人心。

1.1.3 定义

生物信息学的定义随着其研究发展与实际需要而几经改动，一般认为，1995 年美国人类基因组计划第一个五年总结报告中生物信息学的定义较为完整。这个定义是：生物信息学是一门交叉学科，它包含了生物信息的获取、处理、存储、分发、分析和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。

在实际的工作及相关文献中常出现一些词与生物信息学相关，如计算分子生物学、后基因组学、生物数学、生物学、信息学等。由于不少词所包括的内容有交叉或不同的作者对同一名词的解释有所偏重，所以不少初学者容易对这些词产生混淆。以下简单介绍一些相关的专业名词的定义，以示与生物信息学的区别。

生物学(田清涞 2000) 是研究生命的科学，是研究生命现象的本质及探讨生物发生、发展及其活动规律的科学，又称生命科学(bioscience)。生命有几个主要特征：生命体内的化学元素种类与各元素的比例相似，体内有相似的生物大分子，能进行新陈代谢、生长发育与繁殖，会产生遗传和变异，具有应激性。

信息理论(常迥 1993) 是研究信息的产生、获取、度量、变换、传输、处理、识别及其应用的一门科学。来自生物方面的信息源是自然信息源的一个组成部分。

生物数学(徐克学 1999) 是一门介于生物学与数学之间的边缘学科，它以数学方法研究和解决生物学问题，并对与生物学有关的数学方法进行理论研究。

计算分子生物学(João Setubal 和 João Meidanis 1997) 计算分子生物学是开发和使用数学与计算机技术，以帮助解决分子生物学中的问题的一门学科。

生物统计学(李春来，王志和，王文林 2004) 是用数理统计的原理和方法来分析和解释生物界各种现象和实验资料的科学。