21世纪高等院校教材
生·物·科·学·系·列

Bio informatics

生物信息学

张阳德 编著

科学出版社
www.sciencep.com

内 容 简 介

本书科学性和新颖性较强，作者通过对近年来教学、科研中素材的总结，引用分析了大量国内外文献资料，全面、系统地介绍了生物信息学相关的概念、产生背景、发展历史、研究目标、国内外研究现状以及人类基因组计划和蛋白质组信息学等前沿课题，并对生物学数据库的建立、核酸序列分析技术、蛋白质结构预测和分子设计技术、生物信息学软件的开发与应用等内容作了较为全面的阐述。

本书经多位院士和专家审阅，其出版对生物信息学在我国的普及、研究与推广应用将起到积极的促进作用。

本书适合于基础医学、临床医学、生物科学、计算机科学等相关专业学生使用，也可为广大教学、科研人员参考使用。

图书在版编目(CIP)数据

生物信息学/张阳德编著. —北京：科学出版社，2004.9
21世纪高等院校教材——生物科学系列

ISBN 7-03-012319-0

Ⅰ.生… Ⅱ.张… Ⅲ.生物信息学-高等学校-教材 Ⅳ.Q811.4

中国版本图书馆CIP数据核字（2003）第091384号

责任编辑：周 辉 单冉东/责任校对：包志虹
责任印制：安春生/封面设计：周 灿

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
源海印刷有限责任公司印刷
科学出版社发行 各地新华书店经销
*
2004年9月第 一 版 开本：B5（720×1000）
2004年9月第一次印刷 印张：22 3/4
印数：1—6 500 字数：428 000

定价：35.00元

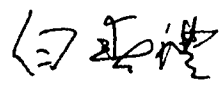
（如有印装质量问题，我社负责调换〈路通〉）

序 一

21 世纪是生命科学、新材料技术、信息技术突飞猛进的发展时期。生物信息学作为跨越生命科学和信息科学两大热点领域的学科展现了它蓬勃的生命力。生物信息学是包括生物信息的获取、处理、贮存、分发、分析与解释的所有方面的一门学科，它综合运用数学、计算机科学和生物学的各种工具，研究、了解大量的生物学意义，已成为整个生命科学发展的重要组成部分，特别是在“后基因组时代（post genome era)”，面对人类基因组计划所产生的宠大的分子生物学信息，生物信息学的重要性越来越突出，无疑将会为生命科学研究带来革命性的变革。

生物信息学是一门理论概念与实践应用并重的学科，在我国，生物信息学随着人类基因组研究的展开才起步，但已显露出蓬勃发展的势头，许多研究单位已经开始或准备开始从事这方面的研究工作。

张阳德编著的《生物信息学》，系统地阐述了生物信息学的理论、技术与方法，内容涉及生物信息学的发生、发展和前沿动态，反映了国内外该领域的最新进展。它将成为生物信息学相关专业的本科生、研究生，以及广大生物学研究者的良师益友。书中阐述的生物信息学的科学思维与方法，将对我国生物信息学领域专业人才的培养、训练提供很新的教学模式，可推进改变我国生物信息学专业人才匮乏的局面，该著作是生物信息学专业的一本很好的教材。



序 二

当今，生物科学技术的迅猛发展，无论从数量上还是从质量上，都极大地丰富了生物科学的数据资源。数据资源的急剧膨胀迫使人们寻求一种强有力的数据分析管理方法，以利于对已知生物学知识的储存和进一步加工利用。由于计算机技术和网络技术的迅速发展，并日益渗透到生物科学的各个领域，一门崭新的、拥有巨大发展潜力的生物信息学也应运而生并迅猛发展。它的诞生和发展是顺应形势所需，以至人们在意识到它的存在之前就已经离不开它了！

国内外非常重视生物信息学的发展，各种专业研究机构如雨后春笋般涌现出来，生物科技和制药公司内部的生物信息学部门也都配备完善。随着社会对生物信息需求迅猛增长，美国等发达国家也面临着资信少、人才匮乏的局面。张阳德教授编著的《生物信息学》，详细介绍了生物信息学的概念、发展历史以及国内外发展的最新动态，编著者总结了近年在研究生教学中的素材，分析了国内外大量文献资料，全面、系统地介绍生物信息学。本书科学性和新颖性较强，全书图文并茂，极具指导价值。

生物信息学是当前生物学领域的研究热点，在未来的若干年内它将变得更加重要、越来越引起人们的重视。本书的问世，有益于培养生物信息学领域的专业人才，衷心希望该书能成为广大医疗、教学、科研人员的良师益友。

刘德培

前　言

“21 世纪是生命科学的世纪”，随着许多模式生物的基因组序列和基因目录的完成以及人类基因组计划进入一个高速收获的时期，基因和许多分子的数据呈指数级上升，而分子水平和整个生物系统的信息水平之间却出现了一道鸿沟；在现代生命科学迅猛发展的过程中，跨学科、跨领域的新思想、新方法不断涌现；利用信息技术剖析生物现象的本质已成为生命科学研究工作者们关注的焦点。计算机科学、网络和数理科学所提供的方法与手段，使信息科学快速发展。多学科融合的生物信息学，正是填补了分子水平和生物系统的信息水平之间的这道鸿沟，成为生命科学研究领域中的一颗璀璨的明珠。

生物信息学是一门研究生物和生物相关系统中信息内容和信息流向的综合性系统科学。通过对生物信息的计算处理，人们能从众多分散的生物学观测数据中获得对生命运行机制的详细而系统的理解。当今的生物学不仅仅是基于观察和实验的科学，理论和计算也将在其领域中发挥巨大的作用。生物信息学的发展反映了科学知识的深化和研究方式的转变，在短短几年内已影响了生物、医学、农业等众多领域。可以预计，生物信息学将渗透到更多的领域，对未来军事和国防的影响也不容忽视。它将从根本上改变生物医学的模式，全方位地推动生物医学的革命。如何利用生物信息库和生物计算手段，是多学科研究人员需要掌握的一种新的基本技能。

生物信息的快速发展，容易使人们在浩如烟海的信息中迷失方向。在这种形势下，亟需一本尽可能全面反映这个领域的书籍以充当在生物信息汪洋大海中游弋的导航图。这就促使我们编写了这本《生物信息学》教材。全书共分 10 章，第一章概要介绍了生物信息学的产生背景、研究目标、国内外研究现状和展望。第二章到第五章详细介绍了生物信息学的基础知识，包括相关的生物知识、生物学数据库及部分算法。第六章全面介绍了核酸序列分析技术，内容包括序列翻译与 ORF 预测、核酸序列分析框架、数据库搜索、生物信息学软件等，并有新测定 DNA 序列的分析实例。第七章介绍了蛋白质结构预测和分子设计技术。第八章和第九章分别介绍了人类基因组计划和蛋白质组信息学。第十章引出了生物信息学的一些前沿问题，包括生物芯片、药物设计、基因诊断与治疗。书末还附录了生物信息学名词解释和自测习题。全书循序渐进，由浅入深，面向不同层次的读者，相信能够对生物信息学在我国的普及、研究与应用起积极的推动作用。

本书的写作，得到了白春礼院士、刘德培院士、魏于全院士、钟南山院士、黄伯云院士、何继善院士、裘法祖院士、胡冬晖教授、曹雪涛教授、裴雪涛教

授、陈志南教授、朱桢教授、王小宁教授的热情支持和指教，在此表示衷心的感谢，并感谢白春礼院士、刘德培院士百忙之中为本书作序。

为了使本书能具有新颖性、科学性、反映最新进展，本书经多次修改，工作量大，为保证出版时间，潘一峰、任力锋、李异凡、罗湘建、张蕾、李沐纯、郭妍、赵劲风、罗育林、周健、彭健、廖明媚、张浩伟、刘蔚东、刘勤、李年丰、何剪太、金鑫、曹兴、翟登高等作了大量的校编工作，为本书的出版和质量保证作出了贡献。本书既介绍了生物信息学的基础知识，又侧重于生物信息学的应用和前沿及发展方向，确保了科学性、系统性和先进性。本书适合作为生物医学及其相关领域学生使用的教材，而且可以作为生物医药领域科研人员的参考书。

本书引用了有关本学科进展的国内外文献资料，在此对这些参考文献作者致以衷心的感谢！同时，敬请读者对本书的不足之处给予指正。

张阳德

2004年于长沙

目　录

第一章 概 论

1.1 生物信息学产生的背景

诺贝尔生理学或医学奖得主 R. Dulbecco1986 年 3 月在 *Science* 上发表文章《癌症研究的转折点：测序人类基因组》，认为要彻底阐明癌症的发生、演进、侵袭和转移的机制，必须对人体细胞的基因组进行全测序。经过 3 年多的讨论，美国政府于 1990 年 10 月正式启动一项耗资 30 亿美元的 15 年计划，预期到 2005 年完成人类基因组大约 30 亿个碱基的全序列测定，这就是被称为生命科学“登月计划”的人类基因组计划（Human Genome Project，HGP）。

HGP 的主要任务是：人类基因组以及一些模式生物体（细菌、酵母、线虫、果蝇等）基因组的作图、测序和基因识别。该计划一经提出，很快扩展成为世界范围的研究计划，并以惊人的速度前进。经过美、英、日、法、德和中国科学家的共同努力，至 2000 年 6 月 26 日完成了工作草图；至 2001 年 2 月 12 日完成并公布了准确、清晰、完整的人类基因组图谱。这是人类科学史上又一个里程碑式的事件，它预示着完成人类基因组计划已经指日可待。生物信息最基本的表达形式是一维的分子排列顺序，即序列，包括核酸序列和氨基酸序列，最基本的仍是 DNA 序列。截至 2002 年为止，仅登录在美国 GenBank 数据库中的 DNA 序列总量已超过 280 亿碱基对。基于 cDNA 序列测定所建立起来 EST 数据库的记录也已达数百万条。在这些数据的基础上派生、整理出来的数据库已达 500 余个。与其同步的蛋白质的一级结构，即氨基酸序列也飞速增长；还有蛋白质的高级结构，目前为止已有一万多种蛋白质的空间结构以不同的分辨率被测定。这一切构成了一个生物学数据的海洋。这种科学数据的急速和海量的积累规模，在人类的科学研究历史中是空前的。

人类基因组计划的直接结果是获得了海量的不连续的数据。由于计算机数据库等技术的迅速发展，从 20 世纪 80 年代初开始建立了美国的 GenBank，欧洲的 EMBL 和日本的 DDBJ 等国际性 DNA 数据库，用户可以通过光盘或其他存储媒体以及通过 Internet 获得这些序列，包括最新的序列。蛋白质的一级结构即其氨基酸序列，也建立了相应的数据库，其中著名的有 PIR 和 SWISS-PORT 等；迄今为止已有约 6000 种蛋白质的结构被阐明，记录这些详尽空间结构的数据库为美国的 PDB。美国国立图书馆生物信息研究中心（NCBI）的 Entrez 不但有 Sequences 数据库，还有大量的文献信息。除这些大型主要数据库以外，还有相对较小的专门性数据库，如 GenProEc 为大肠杆菌基因和蛋白质数据库。这些林林

总总、信息各异的数据库，由 Internet 网连接，构成了极其复杂、规模巨大的生物信息资源网络。

数据并不等于信息和知识，但却是信息和知识的源泉。如何收集、存储、分析这些数据，尤其是如何从这些不连贯的数据中获取有用的生物学信息是问题的关键所在。生物数据量的迅猛增长，既受益于数理科学和计算机科学所提供的方法与手段，也呼唤着多种学科的共同努力。于是，伴随着美国国立卫生研究院（NIH）的人类基因组计划，生物信息学应运而生了。

1.2 人类基因组计划

从分子生物学的中心法则可以知道：DNA 是携带遗传信息的主要载体，DNA 转录成为 mRNA，mRNA 再翻译成蛋白质，由蛋白质行使各种生物功能。几乎所有生物的遗传信息都存在由 A、G、C、T 四种碱基组成的 DNA 序列中。对于人类，由 30 亿碱基组成的遗传信息存在于 23 对染色体中。如果每一个碱基作为一个字符，30 亿碱基就相当于 3000 本每本 1000 页每页 1000 字的天书。

HGP 的主要目标是提供公开、完全、高质量的人类基因组全序列。HGP 由美国能源部（DOE）和美国国立卫生研究院（NIH）提出并提供资助，于 1990 年 10 月 1 日正式启动，原计划 15 年，根据每个碱基的花费 1 美元来计算，预计耗资 30 亿美元。但由于技术的改进，计划提前完成，于 2001 年 2 月 12 日完成并公布了准确、清晰、完整的人类基因组图谱（*Nature*[1]，2001 年 2 月 15 日；*Science*[2]，2001 年 2 月 16 日）。但在染色体的着丝粒和端粒处的序列，由于存在很多重复序列，目前的方法还难于测出。HGP 和“曼哈顿原子弹计划”及“阿波罗登月计划”并称为 20 世纪的三大著名计划。曼哈顿原子弹计划从微观层次入手，而阿波罗登月计划则是从宏观的角度出发，都抓住了物质、能量、信息的核心问题。我们说，生命活化物质，生命耗散能量，生命依存信息。HGP 就是要破译生命的遗传信息密码，现在已经取得了初步的成功。中国在 1999 年 9 月 1 日正式承担了 HGP 的 1%，也就是第三号染色体上 3000 万碱基对的测定。

2001 年 2 月份的 *Nature* 杂志上公布了人类基因组图谱。令人意想不到的是，研究结果所显示的人类基因组包含的基因数目，只是原本预料的（6 万～10 万个）几分之一，约 3 万～4 万个基因，是线虫的 3～4 倍，果蝇的 2 倍，全序列比老鼠的也只多了约 300 个核苷酸。

人类基因数如此之少，而人的结构和功能却如此复杂。研究人员对此作了如下解释：①人类基因有 3 万～4 万个，而蛋白质却有 25 万种之多，故推测每个基因平均合成几种蛋白质。②人体蛋白质在合成后分别修饰上了不同的糖苷或其他化学物质，使其产生功能各异的蛋白质。③新数字计算失误？人类基因科学主席 W. Haseltine 博士一直认为人类有 12 万个基因，他至今还坚持 10 万～12 万

个基因数的估计。他认为，两个小组的研究人员分别得出很低的数字，是因为他们采用了错误的方法，寻找基因的方法有所欠缺。例如，艾滋病病毒中的 10 个基因起初只找到 5 个。他说："我个人的看法是，随着人们对染色体持续研究，将提高基因的数目，直至原本估计的 10 万～12 万个。"

人类基因组计划的分析还在很初步的阶段。随着基因组计划的进一步实施，特别是后基因组计划的开展，更加需要对数据进行分析、比较、建模和预测，以推动生物信息学的迅速发展。

HGP 已经完成，进入到后基因组计划，或者说"后基因组时代"（post-genome era），即揭示基因组及其包含的全部基因的功能，以及对基因产物——蛋白质结构和功能的研究和预测（蛋白质组学，proteomics）。这是对应于基因组学的一个概念。蛋白质组（proteome）是指由一个细胞或一个组织的基因所表达的全部相应的蛋白质。蛋白质组与基因组不同，它是一个动态的概念：①不同组织和不同发育时期所表达的蛋白不同；②基因在转录后，还有一系列的修饰、翻译等过程都可以影响蛋白质的表达。因此通过对蛋白质组的研究，更能阐明遗传、发育、进化、功能调控等基本生物学问题，以及与人类健康和疾病相关的生物医学问题。

蛋白质的三维结构是其发挥生物学功能的基础。对蛋白质空间三维结构的研究，以往多依赖于实验手段，如 X-射线衍射法，磁共振等。随着基因组序列的测定，有可能从基因的特征序列快速地推断出可能的蛋白质结构，这就要依赖于计算机模型的建立。在美国加利福尼亚州召开的第四届蛋白结构预测关键评估技术双年会（CASP4）上展示了完全使用计算机程序，不需要人工努力进行的全自动建模技术。虽然这些技术并不是非常成熟，需要进一步的发展完善。但可以说，后基因组或者蛋白质组的研究，将成为 21 世纪生命科学研究的主要任务，而这更离不开生物信息学的发展。

1.3 什么是生物信息学

生物信息学是 20 世纪 80 年代末开始，随着基因组测序数据迅猛增加而逐渐兴起的一门新兴学科，是利用计算机对生命科学研究中的生物信息进行存储、检索和分析的科学。其研究主要是利用计算机存储核酸和蛋白质序列，研究科学的算法，编制相应的软件对序列进行分析、比较与预测，从中发现规律。

(1) 目前对生物信息学的几种定义

由于生物信息学是一门新兴的、正在迅速发展的交叉学科，目前国内外对生物信息学的定义众说纷纭，但它们都是从不同的角度反映了生物信息学这一新学科的主要特点或其主要研究内容。

美国国家基因组研究中心认为，生物信息学是一个代表生物学、数学和计算

机的综合力量的新兴学科（Bioinformatics is an emerging scientific discipline representing the combined power of biology, mathematics, and computers）（http：//www.ncgr.org/gsd）。

美国乔治亚理工大学认为，生物信息学是采用数学、统计学和计算机等方法分析生物学、生物化学和生物物理学数据的一门综合性学科（Bioinformatics is an integration of mathematical, statistical and computer methods to analyze biological, biochemical and biophysical data）http：//www.biology.gatech.edu/bioinformatics/whatis.html）。

美国密苏里大学认为，生物信息学是获知、管理和处理生物学信息的科学与技术（Bioinformatics is the science and technology about learning, managing and processing biological information）http：//cecssrvl.cecs.missouri.edu/bioinformatics/lec01.html。

美国加利福尼亚大学洛杉矶分校认为，生物信息学是对生物信息和生物学系统内在结构的研究。它将大量系统的生物学数据与数学和计算机科学的分析理论及实用工具联系起来（Bioinformatics is the study of the inherent structure of biological information and biological systems. It brings together the avalanche of systematic biological data with the analytic theory and practical tools of mathematics and computer science）http：//www.bioinformatics.ucla.edu。

Whatis.com 网站认为，生物信息学是以加快生物学研究为目的而建立计算机数据库和运算方法的科学（Bioinformatics is the science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research）。

中国军事医学科学院欧阳曙光等[3]认为，生物信息学是研究生物信息的采集、处理、存储、传布、分析和解释的科学，它通过综合数学、计算机科学与工程学、生物学的工具和技术揭示大量而复杂的生物数据所赋有的生物学奥秘。

（2）现代生物信息学的定义

现代生物信息学是现代生命科学与信息科学、计算机科学、数学、统计学、物理学和化学等学科相互渗透而形成的交叉学科，是应用计算机技术和信息论方法研究蛋白质及核酸序列等各种生物信息的采集、存储、传递、检索、分析和解读，以帮助了解生物学和遗传学信息的科学。从其研究所涉及的学科上看，生物信息学是集生物学、数学、信息学和计算机科学一体化的一门新的科学；从其研究的主要内容看，基因组信息学、蛋白质的结构模拟以及药物设计是生物信息学的三个重要组成部分，并有机地结合在一起。

生物信息学的核心是基因组信息学。基因组信息学作为一个学科领域，包括基因组信息的获取、处理、存储、分配、分析和解释。基因组信息学的关键是“读懂”人类基因组的核苷酸顺序，即全部基因在染色体上的确切位置及各 DNA

片段的功能。具体说，其内涵包括：①要发展有效的、能支持大量数据信息处理需要的软件和数据库；②需产生若干数据库工具，包括电子网络等远程通讯工具，能容易地处理日益增长的物理图、遗传图、染色体图和序列信息，并在这些数据资料中进行比较；③要研究算法和分析技术，用于解释基因组的信息，例如预测功能基因等。不言而喻，与之相应的很多计算都是大规模的，有些甚至需要发展新一代巨型机才能完成。生物信息学的另一个重要组成是进行蛋白质、RNA 等分子的结构模拟和分子设计，以及随之而来的药物设计。

生物信息学是一门以信息知识为基础的学科，关键资源是知识，关键技术是信息处理。它为揭示人类及重要动植物种类的基因组信息，继而进行生物大分子结构模拟和药物设计，为天然生物大分子的改造和基于受体结构的药物分子设计提供依据。生物信息学不仅对认识生物体和生物信息的起源、遗传、发育与进化的本质具有重要意义，而且可为人类疾病的诊断和防治开辟全新的途径，还可为动植物的物种改良提供坚实的理论基础。此外，很可能通过对影响药物代谢或效应通路、相关基因编码序列的再测序，揭示个体对药物反应差别的遗传学基础。

1.4 生物信息学的研究目标和内容

生物信息学的研究目标：认识生命的起源、进化、遗传和发育的本质，破译隐藏在 DNA 序列中的遗传语言，揭示“基因组信息结构的复杂性及遗传语言的根本规律”，揭示人体生理和病理过程的分子基础，为人类疾病的诊断、预防和治疗提供最合理而有效的方法和途径。

1. 近期任务

(1) 大规模基因组测序中的信息分析

测序仪的采样、分析、碱基读出，载体标识和去除，拼接与组装，填补序列间隙，重复序列标识，读框预测，基因标注等都依赖于信息学的软件和数据库。这也是一个信息的收集、整理、管理、处理、维护、利用、分析的过程，包括建立国际基本生物信息库和生物信息传输的国际互联网系统；建立生物信息数据质量的评估与检测系统；生物信息的在线服务；生物信息可视化等。

(2) 新基因和新单核苷酸多态性（SNPs）的发现与鉴定

①新基因和新 SNPs 的发现与鉴定：在基因组序列上寻找新基因和 SNPs 可以通过理论方法进行预测，如利用国际 EST 数据库（dbEST）和各独立实验室测定的相应数据，经过大规模并行计算发现新基因和新 SNPs 以及各种功能位点。

②SNPs 的意义：SNPs 在不同人种以及同一人种的不同个体（如正常人和患者）之间存在差别，如果我们能绘制每一个人的基因图谱，那么根据各自的基因特点，实行个体化医学，前景应该是非常美妙的。