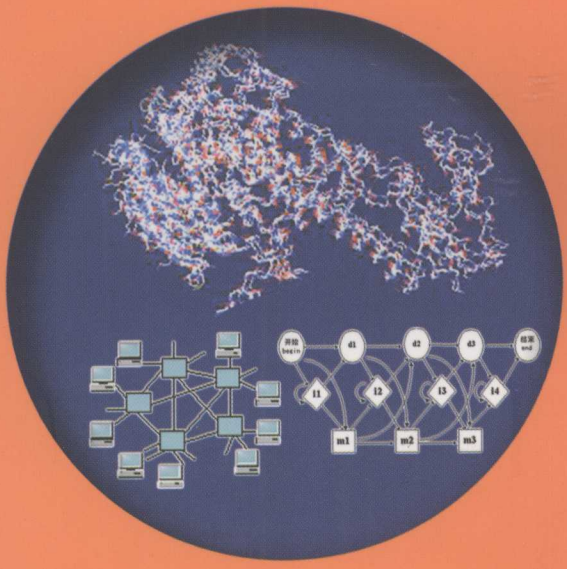
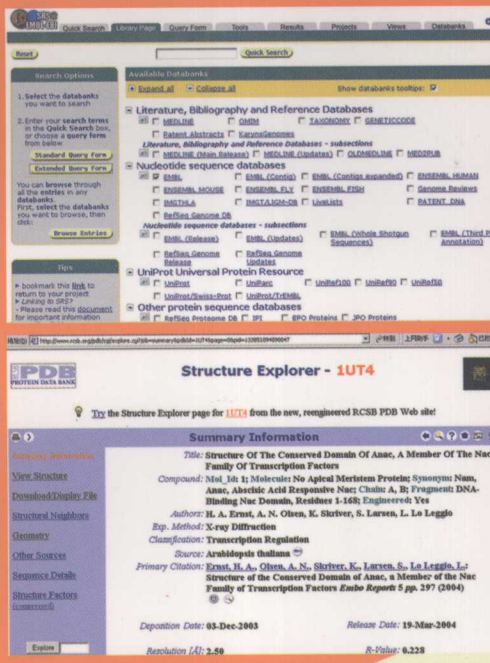
普通高等教育“十一五”国家级规划教材

全国高等农林院校“十一五”规划教材

生物信息学

萧浪涛　主编





中国农业出版社

普通高等教育“十一五”国家级规划教材
全国高等农林院校“十一五”规划教材

生 物 信 息 学

萧浪涛 主编

中国农业出版社

图书在版编目（CIP）数据

生物信息学/萧浪涛主编．—北京：中国农业出版社，2006.9

普通高等教育“十一五”国家级规划教材．全国高等农林院校“十一五”规划教材

ISBN 7-109-11029-X

Ⅰ．生… Ⅱ．萧… Ⅲ．生物信息论—高等学校—教材 Ⅳ．Q811.4

中国版本图书馆 CIP 数据核字（2006）第 102038 号

中国农业出版社出版

（北京市朝阳区农展馆北路 2 号）

（邮政编码 100026）

出版人：傅玉祥

责任编辑 李国忠

中国农业出版社印刷厂印刷 新华书店北京发行所发行

2006 年 9 月第 1 版 2006 年 9 月北京第 1 次印刷

开本：720mm×960mm 1/16 印张：17.75

字数：316 千字

定价：25.00 元

（凡本版图书出现印刷、装订错误，请向出版社发行部调换）

主　　编　萧浪涛（湖南农业大学）
副 主 编　宋东辉（山西农业大学　天津科技大学）
　　韦朝领（安徽农业大学）
　　李合松（湖南农业大学）
　　王若仲（湖南农业大学）
编写人员　（按姓氏笔画排序）
　　马　飞（辽宁师范大学）
　　王若仲（湖南农业大学）
　　韦朝领（安徽农业大学）
　　刘素纯（湖南农业大学）
　　李合松（湖南农业大学）
　　宋东辉（山西农业大学　天津科技大学）
　　黄见良（华中农业大学）
　　萧浪涛（湖南农业大学）
　　彭克勤（湖南农业大学）
　　蔺万煌（湖南农业大学）
　　戴小鹏（湖南农业大学）

前　言

生物信息学是一门由生命科学、数学和计算机科学相互渗透形成的新型交叉学科，是生命科学在后基因组时代不可或缺的研究工具。本教材是普通高等教育“十一五”国家级规划教材和全国高等农林院校“十一五”规划教材，充分吸取现有国内外相关教材与著作的长处，并特别对生物信息学所牵涉的多学科相关知识点进行了全面梳理与优化，充分体现了循序渐进和深入浅出的特色。同时在使用软件绘制高质量插图和方便双语教学等方面均做了较大的努力。

本教材共分9章。绪论、第九章、附录和参考文献由湖南农业大学萧浪涛编写，第一章由华中农业大学黄见良编写，第二章由湖南农业大学戴小鹏和彭克勤编写，第三章由安徽农业大学韦朝领编写，第四章由湖南农业大学蔺万煌和刘素纯编写，第五章由山西农业大学、天津科技大学宋东辉编写，第六章由辽宁师范大学马飞编写，第七章由湖南农业大学王若仲编写，第八章由湖南农业大学李合松编写。全书初稿完成后由萧浪涛修改统稿，并于2006年4月在湖南长沙召开编委会会议做了进一步修定。

本教材的编写得到了编写人员所在单位特别是湖南农业大学的大力支持。中国农业出版社教材出版中心提供了热心帮助和指导。美国农业部西部研究中心顾永强博士提供很多参考资料，湖南农业大学黄复深博士、郭兆武博士、周建军副教授、王惠群副教授、湖北孝感学院鲁旭东博士对编写和修改提出了宝贵的建议，湖南省植物激素与生长发育重点实验室晏瑾、刘清、苏益、谭林、库文珍、

胡骁、赵玄之、沈志锦、丁君辉、梁艳萍、童建华、史齐、康朵兰、陈小飞、尹志成、李巍、罗海平等多位研究人员和研究生参与了本书的校对和绘图工作。此外，本教材的编写还参考了国内外多本相关教材与著作，在此一并表示衷心的谢意。

作为一门新学科，生物信息学正不断向深度和广度迅速拓展，新概念、新理论、新技术与新方法层出不穷，虽然本教材力求反映这些新进展，但仍有不少遗漏。为此，编委会为本教材设立了永久网址（http://www.phytohormones.com/bibook），其中收集了很多与生物信息学相关的资料和网络资源，还可通过该网站的论坛和聊天室等工具就教材和教学等有关内容进行充分和长期的讨论交流。由于编者水平所限，加上交稿时间紧迫，本教材缺点和错误在所难免，请同行专家和读者批评指正。

编　者

2006年8月

目　录

绪　论

Preface

近20年来，生命科学尤其是分子生物学的发展日新月异。随着人类基因组计划（Human Genome Project，HGP）和国际水稻基因组测序计划（International Rice Genome Sequencing Project，IRGSP）等大型国际合作研究项目的实施，人类在生命科学领域，尤其是在核酸和蛋白质等生物大分子的序列和功能等方面迅速积累了天文数字般的数据和信息。与此同时，以计算机技术和网络技术为代表的信息科学也在近20年得到了迅猛发展，计算机和互联网已走入千家万户，深入到社会生活的方方面面。生命科学和信息科学作为21世纪的两大领头学科，已经相互渗透并与其他学科交叉而派生出很多新的研究领域。其中，由生命科学和信息科学等学科相结合特别是分子生物学与计算机信息处理技术的紧密结合而形成的交叉学科——生物信息学（bioinformatics）应运而生，并大大推动了相关研究的开展，被誉为“解读生命天书的慧眼”。在今后相当长的时间内，特别是在完成基因组测序以后的“后基因组时代”，生物信息学将在解读基因组序列中的功能信息等方面发挥更大的作用。

生物信息学是一门横跨多个学科领域的新兴交叉学科，整合了生物学、统计学、应用数学、计算机科学、信息科学等学科的最新研究成果，已成为分子生物学、基因组学（genomics）、蛋白质组学（proteomics）、生物芯片等重大前沿领域的基础研究工具。要学习生物信息学这一重要课程，首先必须对这门学科的主要内容、发展趋势、应用领域和学习方法有一个全面的认识，以便为深入学习这门学科的基础理论知识和应用这些知识开展相关研究打下坚实基础。

一、生物信息学的定义与内容

(Definition and Content of Bioinformatics)

生物信息学是采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、储存、传递、检索、分析和解读的科学，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、化学等相互渗透而形成的交叉学科。

生物信息学研究内容丰富，主要包括基因组和蛋白质组（proteome）数据分析、生物芯片信息解读、生物信息数据库、生物学文献等方面，其中很多领域是当今生命科学的重大研究前沿。

（一）基因组和蛋白质组数据分析

基因组学和蛋白质组学的实质就是分析和解读核酸和蛋白质序列中所表达的结构与功能的生物信息。这方面的研究已成为生物信息学的主要研究内容之一。

任何一种生物的全部遗传构成，被称为这种生物的基因组，有关基因组的研究称为基因组学。其中，结构基因组学（structural genomics）注重遗传图谱、物理图谱和测序等方面的研究；功能基因组学（functional genomics）研究基因的表达、调控与功能；比较基因组学（comparative genomics）则主要对不同的基因组之间进行比较研究。

蛋白质组和蛋白质组学的概念是随基因组和基因组学的出现而产生的。蛋白质组概念的提出是由于基因表达水平不能完全代表细胞中活性蛋白质的数量，基因组序列并不能完全描述活性蛋白质所必需的翻译后修饰及反映蛋白质种类和含量的动态变化过程。在一定条件下，某一基因组蛋白质表达的数量类型称为蛋白质组，代表这一有机体全部蛋白质组成及其作用方式。有关蛋白质组的研究称为蛋白质组学。其中，蛋白质组研究技术与方法、蛋白质组的双向凝胶电泳图谱以及对不同条件下蛋白质组变化的比较分析是蛋白质组学的主要研究内容。目前蛋白组学研究的常用方法是利用双向凝胶电泳（two dimensional gel electrophoresis，2-DE）分离复杂的蛋白质组分，并利用专用软件采集和分析凝胶电泳图谱资料，结合氨基酸组成分析和质谱分析对蛋白质斑点进行精确鉴定，以获得蛋白质组成、表达差异、修饰情况等方面的大量信息。

生物信息学在基因组和蛋白质组研究中所起的作用主要有：①基因组信息结构的计算分析，即对基因组数据进行大规模并行计算，并预测各种新基因和功能位点，研究大量非编码区序列的信息结构和可能的生物学意义。②模式生物全基因组信息结构的比较研究，即对已完成全基因组测序的各种模式生物的基因组信息结构进行比较分析，包括同源序列的搜索比对和指导基因克隆。③功能基因组的相关信息分析，包括对基因表达图谱及其相关算法和软件的研究、与功能基因组信息相关的核酸、蛋白质的空间结构的预测模拟以及蛋白质的功能预测。

（二）生物芯片信息解读

生物芯片是近年来由分子生物学发展而产生的一种新技术，是一个学科高度交叉、产业化前景极为广阔的研究领域，具有巨大的理论意义和实用价值。生物芯片通常指通过微加工技术和微电子技术在固体芯片表面构建的微型生物化学分析系统，能够高速率、高通量地完成对细胞、蛋白质、DNA 以及其他生物组分的检测，并实现分析过程的连续化、集成化、微型化和自动化。生物芯片主要包括基因芯片（DNA chip，gene chip，DNA microarray）、蛋白质芯片（protein chip）和芯片实验室（lab-on-a-chip）等。其中，基因芯片是生物芯片中研究最早、最先商品化的产品。它利用核酸双链的互补碱基之间的氢键作用，形成稳定的双链结构，通过检测目的单链上的荧光信号而实现样品的检测。基因芯片可广泛应用于 DNA 测序、基因突变检测、疾病诊断、新基因的寻找、基因转录分析、基因表达检测、药物筛选、农作物育种、环保检测、食品卫生监督、司法鉴定等。生物芯片改变了生命科学的研究方式，堪称一次具有深远意义的技术革命。

生物芯片技术主要包括芯片阵列的构建、样品的制备、生物反应、信号检测、数据处理及分析等环节。信号检测是将芯片置入专用扫描仪中，通过采集各反应点的荧光位置、荧光强弱，再经相关软件分析图像，以快速准确地获取样品中的生物信息。因此，生物芯片技术中整个检测及分析技术环节属于生物信息学的研究内容。

（三）生物信息数据库

生物信息数据库是生物信息学的主要研究内容之一。生物体与生物界的复杂性加上日新月异的生命科学研究产生大量生物学信息，对这些信息的储存、检索、比较分析必须借助于数据库技术，例如各类生物学信息数据库的建立与维护、数据的添加与注释、更新与查询、数据库资料的网络化等。生物信息数据库有以下主要特点：①种类不断增加。除较早出现的核酸序列数据库外，近年来大量出现了基因组数据库、基因图谱数据库、蛋白质序列和蛋白质结构数据库、酶类数据库、免疫学数据库、生物反应数据库、细胞系数据库等多种数据库。此外，还有生物学文献数据库、生物学软件数据库、一体化的综合生物信息学数据库等类型，几乎涉及了生物学研究中的所有研究领域。②结构日益复杂。在数据库的种类迅速增加的同时，生物信息数据库的规模（记录数）和数据结构的复杂程度也在不断增加。例如核酸和蛋白质数据库，其数据除基本的序列数据外，还包含了大量的注释和参考文献以及与其他相关数据库的链接

指针。对蛋白质序列的注释数据还包括了蛋白质功能、空间结构、结构域与活性中心等大量相关内容。③使用日趋便捷。尽管生物信息数据库的种类、规模和结构日益复杂，但数据库的管理和使用却越来越方便和快捷。这主要得益于数据库硬件和数据库管理软件的不断升级。目前大多数数据库具有自动投送数据、在线查询、在线计算、空间结构的可视化浏览等功能。

（四）生物信息学的其他研究领域

生物信息学的其他研究领域包括：①生物计算，根据生物学数据建立数学模型和计算方法，又被称为计算生物学（computational biology）；②生物学应用软件；③生物学文献。生物信息的来源特别是对互联网上的生物信息学资源的收集、整理也是生物信息学的研究内容之一。例如对互联网上各种核酸和蛋白质数据库的网址、数据特点、查询方法的收集和介绍以及对各种生物信息学研究机构、出版物、论坛、新闻组等相关资源的收集和整理。

二、生物信息学的产生与发展

（Origin and Development of Bioinformatics）

近20年来生物学海量数据的积累催生了生物信息学。1991年出现了“bioinformatics”一词，但标志生物信息学产生的许多事件，早在此前便已发生。纵观生物信息学的发展历史，大致可将它分为以下3个主要阶段。

（一）前基因组时代的生物信息学

属于生物物理学范畴的传统生物信息学可以追溯到很久以前，如研究生物发光、生物电、生物磁和激素等信息物质的传递现象及其相应测定技术。以研究序列比对为标志的现代生物信息学则起源于20世纪70～80年代。

这一阶段的主要成就包括核酸和蛋白质序列的初步分析、生物学数据库的建立以及检索工具的开发。例如Dayhoff的替换矩阵、Neelleman和Wunsch的序列比对（sequence alignment）、GenBank（由美国国立生物技术信息中心建立和维护的核酸与蛋白质序列数据库）等大型数据库的建立，形成了生物信息学的雏形。

（二）基因组时代的生物信息学

以基因组计划的实施为标志的基因组时代（20世纪80年代至20世纪末）是生物信息学成为一个较完整的新兴学科并得到高速发展的时期。这一时期生

物信息学确立了自身的研究领域和学科特征，成为生命科学的热点学科和重要前沿领域之一。

这一阶段的主要成就包括大分子序列以及表达序列标签（expressed sequence tag，EST）数据库的高速发展、BLAST（basic local alignment search tool）和FASTA（fast alignment）等工具软件的研制和相应新算法的提出、基因的寻找与识别、电子克隆（in silico cloning）技术等，大大提高了管理和利用海量数据的能力。

（三）后基因组时代的生物信息学

在多种模式生物基因组测序基本完成的后基因组时代（21世纪初开始），这一时期的生物信息学确立了以综合为特征的相互作用分析方法，是生物信息学日趋成熟的时期，已经成为当今生命科学乃至整个自然科学的重大前沿研究领域之一。今后的主要研究目标是对基因组数据的大规模分析、比较与综合，从基因组信息来揭示生物体的系统功能信息，以推进人们对生命活动基本规律的认识。

三、生物信息学的应用（Application of Bioinformatics）

迅速发展的生命科学以及相关生物技术产业对处理大量生物数据的迫切需求是生物信息学产生和发展的重要基础，也是生物信息学应用的主要领域，与学科的自身发展是一种互动关系。生物信息学的发展和应用能够促进相关生物技术产业的发展，而在其应用实践中又会遇到很多新的课题而推动生物信息学研究。事实上，生物信息学在服务生命科学研究和相关生物技术产业的实践中已有大量成功的先例。

首先，在人类基因组计划、国际水稻基因组测序计划以及其他许多模式生物的基因组计划中，生物信息学发挥了越来越大的作用。其次，在方兴未艾的生物芯片产业中，生物信息学方法是生物芯片数据处理的必要工具。此外，生物制药也是生物信息学应用的重要领域，生物信息学为新药筛选和靶标设计提供了新的方法，大大减少了开发成本，并缩短了开发周期。生物信息学还在流行病学、神经科学、作物育种学等诸多学科领域中展现出诱人的应用前景。近年来，生物信息产业高达数百亿美元的年产值已经证实了生物信息产业的巨大经济价值和发展潜力。

可以预见，生物信息学的发展将对生命科学本身的发展产生革命性的影响，其研究成果将大大地促进生命科学其他研究领域的发展。生物信息学的应