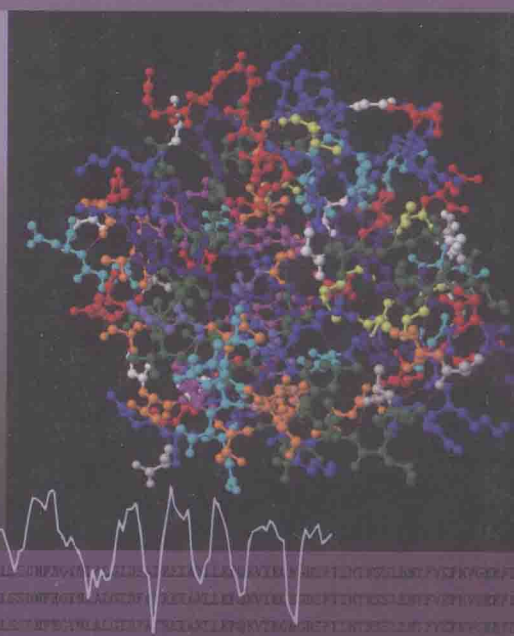
普通高等教育“十一五”国家级规划教材

生物信息学

主编 刘 娟



高等教育出版社

普通高等教育“十一五”国家级规划教材

生物信息学

SHENGWU XINXIXUE

主编 刘 娟

编者 （按姓氏拼音排序）

段谟杰 李 论 刘 娟 陆 枫

罗 飞 马 闯 周 到

高等教育出版社·北京

内容简介

本书围绕目前生物信息学研究与应用的主要内容，以丰富的实例，重点介绍了相关数据库和软件的功能、应用策略和使用方法。具体内容包括：核酸与蛋白质序列数据资源、序列比较与相似序列搜索、分子系统发育分析、基因组结构注释、蛋白质结构分析、蛋白质序列分析、Microarray基因表达数据分析、蛋白质组数据分析、生物信息学在疾病相关基因与药物发现中的应用，以及生物信息导航资源。本书试图综合介绍生物信息学研究解决的问题、基本方法、现有成果与存在的问题，特别是能使读者把握生物信息学自身的特点和分析解决问题的基本途径，使不同专业背景读者都能有一定的收获。

本书适合作为生命科学、计算机科学等相关专业的教材使用，也可供相关科研人员参考使用。

图书在版编目（CIP）数据

生物信息学 / 刘娟主编 . -- 北京 : 高等教育出版社，2014.12

ISBN 978-7-04-040975-8

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

中国版本图书馆 CIP 数据核字（2014）第 204427 号

策划编辑 王 莉　　责任编辑 孟 丽　　封面设计 姜 磊　　责任印制 尤 静

出版发行	高等教育出版社	咨询电话	400-810-0598
社　　址	北京市西城区德外大街4号	网　　址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印　　刷	三河市华润印刷有限公司	网上订购	http://www.landraco.com
开　　本	787mm × 1092mm　1/16		http://www.landraco.com.cn
印　　张	19.5	版　　次	2014年12月第 1 版
字　　数	460千字	印　　次	2014年12月第 1 次印刷
购书热线	010-58581118	定　　价	38.00元

物 料 号　40975-00

数字课程（基础版）

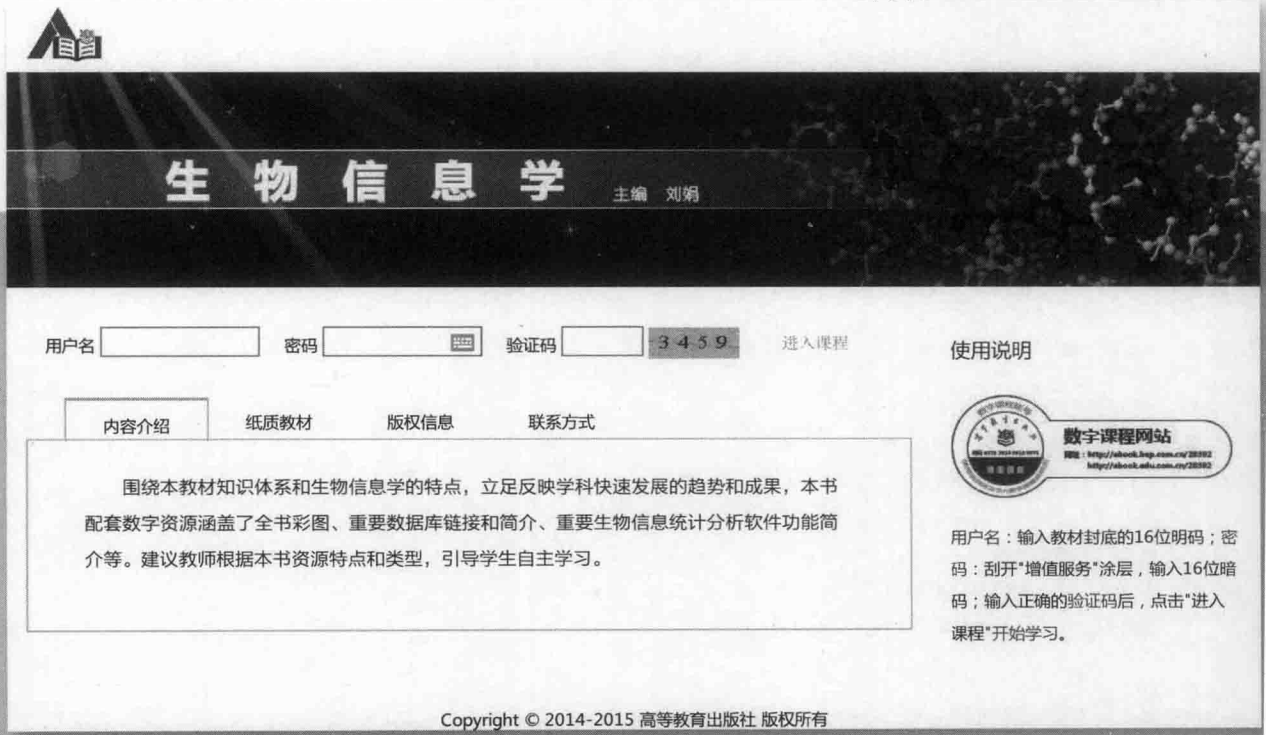
生物信息学

主编　刘娟

登录方法：

1. 访问http://abook.hep.com.cn/40975
2. 输入数字课程用户名（见封底明码）、密码
3. 点击“进入课程”

账号自登录之日起一年内有效，过期作废
使用本账号如有任何问题
请发邮件至：lifescience@pub.hep.cn



http://abook.hep.com.cn/40975

前　言

生物信息学是生命科学、信息科学、统计学等多学科交叉的学科，自人类基因组计划以来迅速发展。促使该学科突飞猛进发展的根本原因就是数据。一方面，海量的实验数据驱动数据的表示、管理和分析方法的发展。另一方面，及时、有效和充分地利用海量的生物信息资源来进行生物机制的理解和生物功能的发现已经成为现代生物、医学、药学等领域研究的必备方法。利用生物信息学的资源和手段开展研究，已经成为多学科研究人员必须要掌握的一种基本技能。鉴于此，我们编写了这本《生物信息学》教材，对生物信息学的常用数据资源、常见的分析手段和工具，以及如何利用数据资源开展应用研究等进行了较为系统的介绍。

全书共分 11 章，各章都配有一定数量的思考题，帮助读者巩固和消化本书内容。第一章简要介绍了生物信息学的发展历史和本书的主要内容。第二章介绍了常用的序列数据库资源及如何从数据库中获取序列数据。第三章比较全面地介绍了序列比对的基本概念、经典的序列比对算法以及有关序列对工具的使用方法等。第四章介绍基因结构的预测方法和 Ensembl 的基因组结构注释流程。第五章系统发生树的基本概念以及分子系统发生树的构建方法。第六章介绍了蛋白质结构的相关知识以及蛋白质结构的预测方法。第七章介绍了蛋白质序列分析方法与蛋白质功能预测方法。第八章介绍了微阵列数据的有关概念和微阵列数据的分析方法。第九章主要介绍了蛋白质质谱数据分析方法和蛋白质 - 蛋白质互作预测方法等。第十章主要介绍了疾病相关的数据资源和疾病基因发现的生物信息学方法。第十一章简要介绍了 SNP 芯片及深度测序的基本概念。

本书既可以作为生物信息学及相关专业的本科生和研究生的教材，也可以作为生物医学等领域的科研人员有关各种生物信息资源的使用指导书。同时，也可作为生物信息学专业人士的研究参考书。希望不同层次的人员通过阅读此书，各取所需，各有收获。

本书由来自 4 所高校 7 位一线教师编写而成，具体分工如下：第一、二、三章由华中科技大学陆枫编写，第四章由西北农林科技大学马闯编写，第五章由华中科技大学李论、陆枫编写，第六章由华中科技大学段谟杰和武汉大学罗飞编写，第七章由华中科技大学陆枫和武汉大学刘娟编写，第八章由中南民族大学周到、

华中科技大学陆枫和武汉大学刘娟编写，第九章由武汉大学刘娟、罗飞和华中科技大学陆枫编写，第十章由武汉大学刘娟、罗飞和华中科技大学李论、陆枫编写，第十一章由武汉大学罗飞编写。全书所有章节由刘娟统稿并编撰习题。刘娟和罗飞对本书全部内容进行最后校对审核。

高等教育出版社的王莉和孟丽编辑为本书的出版付出了辛勤的劳动，在此一并感谢！同时，我们也感谢在科研工作中给予我们帮助的同仁，以及在本书撰写和出版过程中给我们提供各种帮助的单位和个人！

本书中也引用了生物信息学的国内外文献资料，在此对这些参考文献的作者致以衷心的感谢！由于作者学科背景不同、学术水平有限、写作经验不足，书中不足之处在所难免，敬请读者批评指正。

刘娟

2014 年 5 月

目　录

1 绪言

生物信息学(bioinformatics)是生命科学、计算机科学、现代信息科学、数学、物理学以及化学等多个学科交叉结合形成的一门新学科,是利用信息技术和数学方法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。自20世纪80年代至今,生物信息学已得到了快速发展,已对基因组分析、大分子的结构和功能、生化途径、疾病发生以及进化相关等诸多重要的生物学问题展开了研究。其研究成果不仅极大地推动生命科学相关学科的发展,还对农学、医药、食品和环境等领域产生巨大的影响,很有可能引发新的产业革命。

1.1 生物信息学的发展历史

生物信息学作为一门独立的学科只有20多年的历史,但是与生物信息学相关的研究却可以追溯至20世纪中期。为帮助读者更好地了解生物信息学的起源和发展历史,下面简要地列出与生物信息学发展相关的部分重要事件。

(1) 萌芽期(20世纪50—70年代)

1953年,J. D. Watson 和 F. H. C. Crick 根据 R. E. Franklin 和 M. H. F. Wilkins 的 X-晶体衍射数据提出了 DNA 双螺旋结构。

1955年,F. Sanger 发表了牛胰岛素的蛋白质序列,是生物信息学发展的基础。

1962年,L. Pauling 提出了分子进化理论。

1967年,M. O. Dayhoff 构建了蛋白质序列数据库。

1970年,用于序列比较的 Needleman – Wunsch 算法发表。

1971年,蛋白质结构数据库(protein data bank,PDB)在美国纽约 Brookhaven 国家实验室创建。

1974年,欧洲分子生物学实验室(european molecular biology laboratory,EMBL)建立。

1977年,A. M. Maxam 和 W. Gilbert 发表了化学降解法,F. Sanger 和 A. R. Coulson 发表双脱氧终止方法,用于 DNA 测序。F. Sanger 等完成了第一个基因组序列——噬菌体 φX174。

(2) 形成期(80年代)

1980年,K. Wüthrich 发明了利用核磁共振技术测定溶液中生物大分子三维结构的方法;EMBL 核酸序列数据库建立。

1981年,用于序列比对的 Smith – Waterman 算法发表。

1982年,创建了 GenBank 数据库;λ 噬菌体基因组序列完成。

1984年,日本国立遗传学研究所(national institute of genetics,NIG)开始信息服务。

1986年,SwissProt 蛋白序列数据库创立。R. Dulbecco 在《科学》杂志上撰文首次提出人类基因组计划的设想。美国能源部正式提出实施测定人类基因组全序列的计划。

1987年,克隆容量可达几百至几千 kb 的酵母人工染色体(yeast artificial chromosomes,YAC)

问世。NIG 发行日本 DNA 数据库 DDBJ 第一版。

1988 年,D. J. Pearson 和 W. R. Lipman 发表 FASTA 算法。美国国立卫生研究院下属的国家生物技术信息中心(National Center for Biotechnology Information,NCBI)成立。国际人类基因组织(the Human Genome Organisation,HUGO)成立。欧洲分子生物学网络组织(European Molecular Biology Network,EMBnet)创立。GenBank、EMBL 与 DDBJ 共同成立了国际核酸序列联合数据库中心,建立了合作关系。根据协议,这三个数据中心各自搜集世界各国有关实验室和测序机构所发布的序列数据,并通过计算机网络每天都将新发现或更新过的数据进行交换,以保证这三个数据库序列信息的完整性。

1989 年,林华安首先采用“bioinformatics”一词。美国成立国家人类基因组研究中心,J. D. Watson 出任第一任主任。美国 Affymetrix 公司研制出了世界首张基因芯片。

(3) 高速发展期(90 年代至今)

1990 年,美国国会批准正式启动人类基因组计划(Human Genome Project,HGP)研究,计划用 30 亿美元的预算在 15 年的时间内完成人类 30 亿碱基对的测序和基因确定;随后法国、英国、意大利、德国、日本、中国等也相继宣布开始各自的 HGP 研究。S. F. Altschul 发表 BLAST 算法。

1991 年,J. C. Venter 在《科学》杂志上描述表达序列标签(expressed sequence tag,EST)的建立和使用。

1992 年,J. C. Venter 在美国马里兰州成立基因组研究所(the Institute of Genome Research,TIGR),成为细菌基因组测序研究的先驱;M. Simon 和同事宣布细菌人工染色体(bacterial artificial chromosome,BAC)在 DNA 克隆中的应用。

1994 年,欧洲生物信息学研究所(European Bioinformatics Institute,EBI)成立。

1995 年,《科学》杂志首次刊登了 TIGR 采用其创立的全基因组鸟枪法(whole genome shotgun,WGS)完成的流感嗜血杆菌(*Haemopophilus influenzae*,*H. inf*)全基因组测序的论文,这是人类完成的第一个单细胞微生物的基因组序列的测定,标志着基因组时代的真正开始。

1996 年,第一个真核生物——酿酒酵母(*Saccharomyces cerevisiae*,*S. cere*)——全基因组测序完成。Affymetrix 开始正式销售商用基因芯片。

1997 年,第一个重要的实验模式生物——大肠杆菌(*Escherichia coli*,*E. coli*)——全基因组测序完成。S. F. Altschul 发表 PSI - BLAST 算法。

1998 年,第一个多细胞真核生物——线虫(*Caenorhabditis elegans*,*C. elegans*)——全基因组测序完成。

2000 年,拟南芥(*Arabidopsis thaliana*)的全基因组测序工作完成,成为植物界第一个被完整测序的物种。黑腹果蝇(*Drosophila melanogaster*)的基因组测序完成。

2001 年,国际人类基因组测序协作组(The International Human Genome Sequencing Consortium,IHGSC)和 J. C. Venter 领导的 Celera 公司分别在《自然》和《科学》杂志同时发表人类基因组草图。

2002 年,小鼠、水稻基因组草图公布。

2003 年,人类基因组测序计划完成。

1.2 本书内容简介

目前的生物信息学研究，已从早期以数据库建立和 DNA 序列分析为主的阶段，转移到后基因组学时代以比较基因组学（comparative genomics）、功能基因组学（functional genomics）和整合基因组学（integrative genomics）为中心的新阶段。本书对以下生物信息学相关内容给予了介绍。

（1）序列数据资源

序列数据资源贮存了生物信息学研究的原始数据，是生物信息学存在和发展的基础。熟悉并了解这些数据资源将有助于更好地开展生物信息学相关的研究与应用。本书介绍了多个常用的核苷酸和蛋白质序列数据库，以及从这些数据库中获取信息的方法。

（2）序列比对与比对搜索

序列相似性分析是生物信息学最早涉及的问题之一，常用的分析方法是序列比对（sequence alignment）。本书介绍了多个知名的序列比对算法（包括 Needleman - Wunsch 算法和 Smith - Waterman 算法等），以及目前广泛使用的序列比对工具（例如，NCBI 提供的 BLAST 等）。为了进一步帮助读者理解序列比对的结果，本书还从进化的角度描述了两条序列之间的氨基酸（或者核苷酸）是如何被比对和相互比较的。

（3）基因组结构注释

随着人类基因组测序计划的完成及其他物种基因组测序工作的相继展开，公共数据库中已积累了大量的基因组序列数据。为了充分发挥这些基因组序列的应用价值，一项极其重要的工作即是“翻译”出基因组序列中蕴含的生物学知识。本书主要以“翻译”基因组序列中蛋白质编码基因为例，介绍基因组结构注释相关的生物信息学方法和软件。

（4）分子系统发生分析

系统发生关系是表示物种进化关系的参考依据。通过分析分子水平的序列数据，可以了解物种系统发生的关系，目前常用树的形式来表示不同物种间的进化关系。本书结合相关实例主要对分子系统发生树的构建方法和过程进行了介绍。

（5）蛋白质结构

蛋白质的空间结构是其行使功能的基础。在进行蛋白质相关研究时，我们经常会遇到一些难题。比如，跨膜蛋白如何能控制跨膜区域的开合并实现细胞内外分子的转运；又如一些酶被简单的化学修饰后，其活性为何会发生巨大的改变；朊蛋白发生怎么样的构象变化才会导致诸如疯牛病等。要解答上述的问题，都需要深入地了解蛋白质结构。本书介绍了蛋白质结构相关生物信息学的多个重要内容，包括蛋白质结构相关数据库、常用的蛋白质结构分析工具以及利用计算手段预测蛋白质结构的方法。

（6）蛋白质序列分析与功能预测

蛋白质在生命过程中发挥着巨大的作用，它们执行着大部分生物学功能，包括结构功能、酶功能，以及在细胞内或细胞间转运物质的功能等。本书将会介绍用于描述蛋白质功能的 GO 术语，以及基于不同策略进行蛋白质功能注释的生物信息学方法。

(7) 微阵列数据分析

在考察某个具体生命状态下的基因表达水平时,实验人员最想知道的问题是:在这个生命状态中,哪些基因相对于它们正常状态有较为明显的高或低表达? 微阵列(*microarray*)是一种重要的基因表达高通量检测技术。分析微阵列数据时,最大的挑战是如何使用相应的统计学手段判定哪些基因确实存在改变。本书对微阵列的概念、微阵列数据的处理方法和分析软件,以及基于微阵列数据的基因调控网络的构建等内容进行了较详细介绍。

(8) 蛋白质组数据分析

高通量的蛋白质组工程能够大范围地确定蛋白质功能,能确定蛋白质在哪种特殊的生理条件下会出现,还能确定哪些蛋白质之间有相互作用。目前,除了能单独研究一个蛋白质外,同时对数千个蛋白质进行高通量的分析也已成为可能。本书主要介绍了几种对蛋白质进行高通量分析的手段,包括二维凝胶电泳、质谱数据分析、蛋白质-蛋白质互作的生物信息学方法以及分析细胞通路的生物信息学方法等。

(9) 疾病相关研究

寻找疾病相关基因是认识疾病发生机理、研制疾病的基因诊断与防治手段的基础,也是人类基因组研究的重要目标。现在,一些与遗传病有关的重要基因已被分离和测序,另一些常见病,如乳腺癌、结肠癌、高血压、糖尿病和阿尔茨海默病等具有遗传倾向的疾病基因已在染色体遗传图谱上精确定位,寻找致病基因的研究工作也在不断的深入,这些工作为新药的发现和开发提供了可靠的信息和软件工具资源。本书主要介绍了疾病相关的数据资源以及现有的疾病基因预测方法与软件工具。

(10) SNP 芯片及深度测序数据分析

后基因组时代一个典型特征就是生物实验芯片化,深度测序技术的不断发展极大地促进了生物信息学的研究。本章介绍了 SNP 芯片及深度测序技术,介绍了几个重要序列数据格式的规范。在此基础上,通过两个简单的例子,介绍了测序数据分析的基本流程。

1.3 贯穿本书的例子

为了更好地对相关技术与工具资源进行介绍,本书还选用了 J. Pevsner 在其《生物信息学与功能基因组学》中的两个具有代表性的例子,即一个基因和它的对应蛋白质产物——视黄醇结合蛋白(retinol-binding protein,RBP4)以及人类免疫缺陷病毒 1(HIV-1)的 *pol*(polymerase,聚合酶)基因,进行示例性介绍。

(1) 视黄醇结合蛋白

视黄醇结合蛋白是一个相对分子质量小、被大量分泌的蛋白质,能结合血液中的视黄醇(维生素 A)。视黄醇可从胡萝卜中以维生素 A 的形式获得,疏水程度大。RBP4 帮助转运这个配体到眼睛,为视觉系统所用。它有一系列有趣的性质:① 在多个物种中有许多蛋白质和 RBP4 同源,包括人、小鼠和鱼("直系同源")中的蛋白质。② 也有许多人类蛋白质和 RBP4 紧密相关("旁系同源"),它们和 RBP4 的家族称为 lipocalin 家族——一群多样的小配体结合蛋白,它们倾向于分泌到细胞外空间。一些 lipocalin 蛋白具有的功能与 RBP4 不同,例如结合胆固醇(如 apo-

liprotein D)、与妊娠相关(如 pregnancy-associated lipocalin)、催欲(如仓鼠的催欲蛋白)和气味结合(如黏液中的气味结合蛋白)等。③ 有细菌的 lipocalin 蛋白,它们在对抗生素的抗性中起作用。编码细菌 lipocalin 的基因可能是一古老基因,它通过水平基因转移的过程进入真核生物基因组。④ 一些 lipocalin 蛋白的表达水平受到显著的调控。⑤ lipocalin 蛋白小而丰富,并且是可溶性的,它们的生物化学性质已被详细研究,许多蛋白质的三维结构也以 X 线晶体衍射的方法被解析出来。⑥ 一些 lipocalin 蛋白和人类疾病相关。

(2) 人类免疫缺陷病毒

人类免疫缺陷病毒(HIV)是当今世界上最大的公共卫生挑战。HIV-1 基因组仅编码 9 种蛋白质,包括 *pol*。*pol* 基因的特性、其蛋白质产物以及 HIV-1 基因组具有显著特点:① *pol* 基因编码一种 1 003 个氨基酸的蛋白质。该蛋白质是一多结构域蛋白质:单条肽链但有多个结构和功能不同的结构域。② pol 蛋白有反转录酶活性(即 RNA 依赖的 DNA 多聚酶),它也是天冬氨酸蛋白酶,并且还有整合酶(integrase)的活性。有多种活性是多结构域蛋白的典型特征。③ pol 蛋白的模块化特点会影响数据库搜索和多序列比对。④ *pol* 基因以相当快的速度发生碱基替换。一个典型的被 HIV 感染的个体可能会有百万种以上的 *pol* 变种。