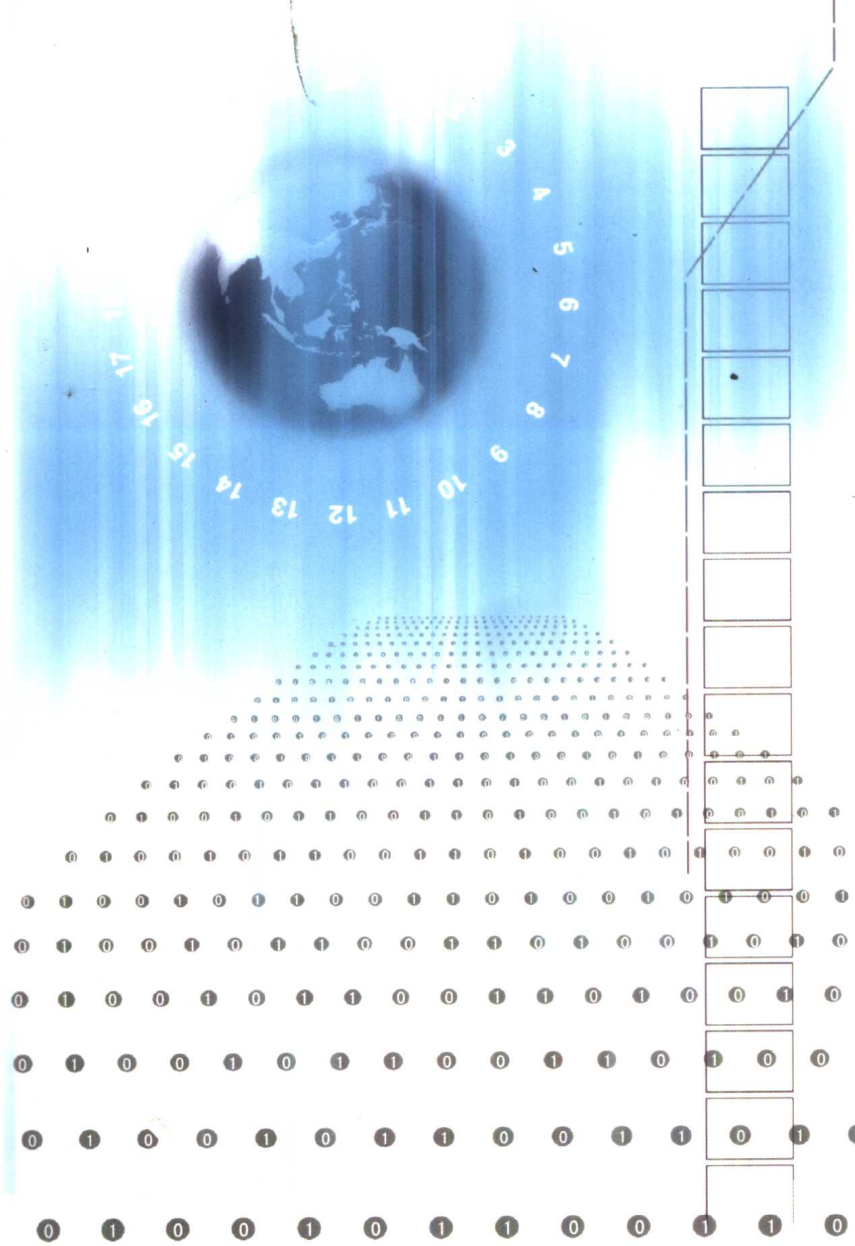
生物信息学

网络资源与应用

黄韧 薛成 等编著



中山大学出版社

生物信息学
网络资源与应用

黄韧　薛成　等编著

中山大学出版社

·广州·

图书在版编目(CIP)数据

生物信息学网络资源与应用/黄韧,薛成等编著.—广州:中山大学出版社,2003.6
ISBN 7-306-02077-3

Ⅰ.生…　Ⅱ.①黄…②薛…　Ⅲ.因特网—生物信息论—情报检索　Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2003)第 036541 号

责任编辑:周建华　封面设计:湘羊　责任校对:舟雨　责任技编:黄少伟

中山大学出版社出版发行
(地址:广州市新港西路 135 号　邮编:510275
电话:020-84111998、84037215)
广东新华发行集团股份有限公司经销
中山大学印刷厂印刷
(地址:广州市新港西路 135 号　邮编:510275)
787 毫米×1092 毫米　16 开本　25.25 印张　624千字
2003 年 6 月第 1 版　2003 年 6 月第 1 次印刷
定价:48.00 元

《生物信息学网络资源与应用》主要编写人员

黄　韧（广东省实验动物监测所，广州，501260）
薛　成（广东省实验动物监测所，广州，501260）
任瑞文（广州军区军事医学研究所，广州，510507）
徐晓立（塔里木农垦大学，新疆，843300）
蒋红霞（华南农业大学，广州，510642）

内容简介

本书系统地叙述了生物信息学网络数据资源的操作方法和应用。

全书共八章以及四个附录，介绍了生物信息学发展历史概况和主要网络数据库的查询操作方法；提供了生物序列比对的基本原理和数据库搜索方法，生物信息数据传递、格式转换，DNA、RNA和蛋白质分子序列比对和引物设计等方面的常用软件使用方法，包括一些实例；书后还附有生物信息学网络资源与应用中常见且重要的中英文词汇对照和数据库、软件及科研机构的网址和简介。

本书实用性和操作性强，文字叙述和操作界面显示并重，满足读者无师自通的目的，可作为高校相关专业高年级本科生和研究生的教学参考书，也适合生命科学研究人员和相关学科工作人员参考使用。

前　言

从达尔文进化论开始，孟德尔遗传定律的发现、分子生物学中心法则的建立直至人类基因组计划图谱的公布，人类在认识生命的进程中步步为营，现正面临破译、解读、调节和开发利用基因组功能的新阶段。一方面，生物技术的快速进步，似乎使我们操纵生命有随心所欲的快感；另一方面，生物信息的扑朔迷离依然使我们洞察生命像雾里看花般茫然。

生物信息学将当今发展最迅速、影响最大的生物学和信息学交汇融合，以生命为对象，以信息为特征，利用数学和信息工具，将分子生物学、生物化学、生理和病理学等数据综合，并与生命现象、结构联系起来透视本质，为生命科学研究带来了新的发展动力，催生了公共的生命信息体系。近 20 年来生物分子序列测序技术的进步和基因组工程的展开，生物学数据如爆炸式增长，大规模数据中心随着信息网络的发展似蔓藤春生。截止 2002 年 8 月，GenBank 收录的基因序列已达 1820 万条，约有 227 亿个碱基，同时它还在以平均每 14 个月翻一番的速度增加；1 万多种蛋白质的空间结构以不同的分辨率被测定；基于 cDNA 序列所建立的 EST 数据库也已达 1143 万条。目前，各种一级数据库和由此派生整理的数据库数目超过 500 个，同时也开发了一系列数据比对、搜索、预测等方面的新算法，并建立了大量的检索及结构、功能预测等网络服务，成为教学、科研人员不可或缺的信息技术平台。

与此同时，生物信息学还初步形成并展示了作为产业发展和潜在市场的良好前景，以及科学、公共利益和商业利益错综复杂的关系。1997 年，美国专利局宣布允许 EST（短 DNA 序列片断，对基因表达谱具有重要作用）的专利申请，全世界科学家出于对人类公共遗传资源和公共利益的忧虑而表示震惊和沮丧；2000 年美国总统克林顿和英国首相布莱尔针对为商业利益将人类基因组研究成果专利化的行为而呼吁公开人类基因组研究成果；1995 年美国 Amgen 公司肥胖基因交易和 1998 年德国 Beyer 公司对 225 个可能具有药物靶分子位点作用的人类基因位点开发权进行购买；2001 年人类基因组草图的公布，使人类拥有共同的生物信息资源。在完成人类基因组计划的基础上，美国国立健康研究院（NIH）向美国国会建议投资 160 亿美元建立 5060 个生物学与计算机结合起来的中心；法国议会科技决策评估办公室也评估了基因工程、生物信息学和组合化学等的应用前景及法国的对策；全世界出现大批基于生物信息学的公司，利用公开的现成的大量生物信息，进行药物设计、基因工程药物、生物芯片以及代谢工程等开发应用。生物信息学的网络资源成为科学研究的焦点、政府关心的热点和商业淘金的沃土。

至 2001 年，中国完成人类基因组 1%的测序工作，不久前又完成 4．3 亿碱基对的水稻基因组工作图，为生物信息学数据库提供了重要的直接信息源，是我国生物信息学的标志性成就。面对全世界生物信息数字化、网络化和数据开放共享的现状，充分挖掘和发现生命规律，加速基因功能产品开发，成为发展我国生物信息学的重要内容和发展

我国生物信息产业的良好机遇，熟悉和使用生物信息学资源和工具成为我国生命科学研究和有关专业师生共同面对的课题。

我们在开展生物信息学有关工作中，深感不管是网络数据库查询还是信息分析利用方面，都迫切需要适用的中文专著或教学参考书，于是编写了这本《生物信息学网络资源与应用》，内容侧重介绍核酸、蛋白质、基因组三大类数据库和查询、生物分子序列比对、文献检索的工具、预测的工具以及常用的分子生物学应用软件，包括 DNA、RNA、蛋白质的序列分析、引物设计、数据递交和格式转换、综合分析等软件的获得途径和使用方法。

本书涉及的大多数数据库及软件都经作者使用过，在叙述其使用方法和应用中，为了加强可操作性，尽可能地采用图文并茂形式，力求使非计算机专业的生物科技工作者在利用本书时无师自通。但是，我们依然感到，对数据库的挖掘和软件的应用上，需要在实践中不断改进和优化；同时，网络资源的不断充实和快速更新，信息存储与分析利用新策略、新方法的层出不穷，新版本推陈出新；还有，作者本身学识和经验的不足，都可能使本书的内容滞后、产生疏漏与错误，凡此种种，敬请读者不吝赐教和鞭策。

黄韧　薛成

2003 年 2 月

目　　录

第一章

概　述

生物信息学（Bioinformatics）是生物学的一个分支，它采用信息科学、计算机科学、生物数学、比较生物学等学科的观点和方法对生命的现象及其组成分子（核酸、蛋白质等）进行研究，主要研究生命中的本质和规律，包括物质组成、结构功能、生命体的能量和信息交换传递等。人类基因组计划（HGP）和生物医药工业是推动生物信息学发展的两个主要力量。

广义地说，生物信息学包含了对基因组研究相关生物信息的获取、加工、储存、分配、分析和解释等。这一定义包括了两层含义：一是对海量数据的收集、整理与管理，也就是管好这些数据；另一是从中发现新的规律，也就是用好这些数据。具体地说，生物信息学是把基因组 DNA 序列信息分析作为源头，找到基因组序列中代表蛋白质和 RNA 基因的编码区；同时，阐明基因组中大量存在的非编码区的信息实质，破译隐藏在 DNA 序列中的遗传信息规律；在此基础上，归纳、整理与基因组遗传信息及其调控相关的转录谱和蛋白质谱的数据，从而认识代谢、发育、分化、进化的规律。同时，生物信息学还利用基因组中编码区的信息进行蛋白质空间结构的模拟和蛋白质功能的预测，并将此类信息与生物体和生命过程的生理生化信息相结合，阐明其分子机理，最终进行蛋白质、核酸的分子设计、药物设计和个体化的医疗保健设计。该领域已经扩展到对基因组学、蛋白组学、药物筛选和药物化学中获得的大量资料的管理、处理、分析和视图化；还有，生物信息学还包括对不断膨胀的数据库的整合与挖掘，寻求新的信息途径和信息分析方法，综合新技术多层面地观察和阐明生命现象和本质。

第一节　生物信息学的发展历史

生物信息学（Bioinformatics）这一名词最早出现于 1991 年的电子出版物中，但是其早期的命名——基因组信息学（Genomics）、计算生物学（Computational Biology）的出现要早得多，在计算机初创期的 1956 年就已经在美国田纳西州的 Gatlinburg 召开过首次“生物学中的信息理论讨论会”；在 20 世纪 60 年代，生物学家 Margaret Dayhoff 就将氨基酸序列的比较用于物种进化的研究，她在研究中建立了一个公共的蛋白质序列数据库系统，随后她将研究的数据输入该数据库，这对以此数据库为基础发展起来的计算机分析工具起了很大的作用。基于 Margaret Dayhoff 的研究和工作的成果，最早的手

工搜集数据的蛋白质数据库于 1962 年建立。而核酸数据库的研究则相对较晚，美国洛斯阿拉莫斯国家实验室 1979 年才建立起 GenBank 数据库；欧洲分子生物学实验室 1982 年开始提供核酸序列数据库 EMBL 的服务；日本也于 1984 年着手建立国家级的核酸序列数据库 DDBJ，并于 1987 年开始提供服务。早期的生物学数据库都还是以常规的出版物或本地计算机为存储媒介，直到 1984 年初，瑞士日内瓦大学医学生物化学系才开发出了第一个分子生物学的 Web 服务器 ExPAsy。

这一时期，遗传密码刚刚被揭示，限制性内切酶、耐热 DNA 聚合酶都还没有被发现，也缺乏快速、高效的测序手段；同时，数据存储、传输及分析手段受到限制。传统的代谢途径研究及测序工作，往往是由单个科学家独立完成的，比如科学家对互不相干的单个领域如癌症、心脏病以及记忆力等方面的研究，虽使得相应领域得到发展，但不同研究者获得的数据都是相互独立、互不相关的。因此，这一时期的生物信息学研究和分析技术是十分初级的，无论是数据的数量还是分析的手段，抑或其数据整合能力，都还不足以对生物学研究产生重大影响。

20 世纪 80 年代，生物学技术，特别是基因工程技术得到迅速发展：大批肿瘤基因与肿瘤抑制基因被发现；基因克隆技术获得突破；基因表达研究技术渐趋完善，神经活动的研究面临新的突破；大规模双向电泳、核磁共振、X 光晶体衍射以及聚合酶链反应等技术的建立与改进，使得现代分子生物学得到了前所未有的发展。这些研究使得人们对生命的许多本质现象的认识逐渐深入，同时发育生物学家在追踪生物不同的发育阶段基因作用模式的过程中发现，多细胞生物所有细胞中都具有相同的一套遗传物质，但每一细胞都仅仅利用了其中很小的一部分，细胞内活化的基因集限定了细胞的生物学功能，并决定了细胞的分化，而且一般来说这一过程是不可逆转的，从而使人们认识到生物功能的执行并不仅仅取决于单个基因，生物体内的基因表达有其复杂的调控机制及时空模式，基因的功能不仅仅是由它所编码的蛋白质决定的，还决定于基因在染色体中的组织方式，而且基因并不是独立而是作为一个群体在起作用的。同时，随着分子生物学技术及核酸测序手段的进步，使得对整个基因组进行测序的工作变得不再遥不可及，人们开始进行众多模式生物的基因组测序工作。其中，20 世纪 80 年代末开始启动的人类基因组计划是生物信息学发展过程中一个划时代的事件，参与此计划的科学家不再仅仅是为了零星地探索细胞活动的过程及其机制或者是单个基因的功能，相反，他们试图从总体上对染色体 DNA 进行分段测序，并最终将它们拼接构成完整的基因组，从而将探讨生命活动奥妙的进程提高到一个新的高度。

随着生物信息学数据的快速积累，各国相继成立了一大批具影响力的生物信息学中心。美国于 1988 年在国会的支持下成立了国家生物技术信息中心（NCBI），其目的是进行计算分子生物学的基础研究，构建和发布分子生物学数据库；欧洲于 1993 年 3 月就着手建立欧洲生物信息学研究所（EBI）；日本也于 1995 年 4 月组建了自己的信息生物学中心（CIB）。与此同时，大量的关于生物信息学的专业期刊也相继开始发行。1970 年出现了第一本提供生物信息学相关内容的专业期刊 Computer Methods and Programs in Biomedicine；到 1985 年 4 月，第一本生物信息学专业期刊 Computer Application in the Biosciences 开始发行。目前关于生物信息学领域较有影响的期刊包括 Bioin-

formatics（免费的网络期刊，http://yiliao.shol.com.cn/journal/j4.htm），Acta Biotheoretica（http://yiliao.shol.com.cn/journal/j1.htm），Bio Informatics Technology & Systems（http://yiliao.shol.com.cn/journal/j2.htm），Bioinform Newsletter（http://yiliao.shol.com.cn/journal/j3.htm），Briefings in Bioinformatics（http://yiliao.shol.com.cn/journal/j6.htm）和 Journal of Computational Biology（http://yiliao.shol.com.cn/journal/j12.htm）等。由中国科学院遗传与发育生物学研究所主办的我国第一本关于生物信息学的专业期刊《基因组蛋白质组与生物信息学报》也于 2003 年创刊。

在序列分析及核酸和蛋白质分子结构预测与算法方面，早在 1962 年，Zuckerkandl 和 Pauling 就将序列变异分析与其演化关系联系起来，从而开辟了分子演化的崭新研究领域；1964 年，Davies 开创了蛋白质结构预测的研究；1970 年，Needleman 和 Wunsch 发表了广受重视的两序列比较算法；同年，Gibbs 和 McIntyre 发表了矩阵打点作图法，用于寻找单条序列的重复片断，从而推测基因功能；Gatlin 在序列比较中引入信息理论，首次发现自然的生物分子序列具有高度非随机的定量证据；1974 年，Ratner 首先运用理论方法对分子遗传调控系统进行处理分析；1975 年，Pipas 和 McMahon 首先提出运用计算机技术预测 RNA 二级结构；随后，出现了许多种序列分析和预测地算法，特别是 20 世纪八九十年代以来，随着生物学数据的迅速增长，生物学数据分析技术也随之大量地涌现，并获得了突飞猛进的发展。

1990 年第一届国际电泳超级计算机和人类基因组会议在美国举行，会议的主要内容就是生物信息学；1994 年国际生物信息学系列会议由 Healthtech 研究所承办，并走向商业化；次年，该系列会议改成年会形式。1997 年，美国专利局允许 EST 的专利申请，引起全世界科学家震惊的同时也预示生物信息商品化；同年，由 DNA 决定逻辑思路的 DNA 计算机出现，表明生物信息与信息技术的真正相互渗透和交叉成为现实，改变了生物学利用信息技术方法的单向现象。2000 年 3 月，美国总统克林顿和英国首相布莱尔针对为商业利益而试图将自己的研究成果申请专利的纷争，发表联合申明，呼吁公开人类基因组研究成果。2001 年 2 月人类基因组序列图谱公开发表，使人类共同站立于基因组学大平台上。紧接而来的后基因组时代，基因结构与功能的关系、基因表达及调控的时空关系成为更令人向往和高深莫测的命题，也是科学家的热情焦点所在。同时，由于基因组功能和产物在健康诊疗及新药开发的巨大潜力而成为新的投资热点和淘金沃土。科学技术和商业行为，共同推动生物信息学向前发展。

第二节 序列测定技术的发展

生物信息学的形成源于生物分子序列数据的积累。并且分子序列也是生物信息学进行分子识别、比对和预测等分析的基本信息。因此有必要回顾分子序列测定技术的发展。

一、蛋白质序列测定技术

1945 年以前，没有任何蛋白质序列定量测定的方法，在其后的十年中，随着色谱技术及标记技术的发展，1950 年，Pehr Edman 提出了 N 末端顺序降解多肽的技术，并随后发展了著名的 Edman 降解技术，从而展开了蛋白质测序的序幕。1953 年 Frederik Sanger 及其同事首次采用 2，4－二硝基氯苯 N 末端标记，并结合酸水解及纸层析分离技术，首次测得了第一个完整的蛋白——牛胰岛素的氨基酸序列，并于 1958 年获得了诺贝尔奖。随后，促黑素 α 和 β 链、ACTH、赖氨酸与精氨酸血管升压素和催产素（Du Vigneau）等一系列的短链多肽相继被测序。在此基础上，Spackman 在 1958 年发明了氨基酸测序仪，首次允许高精度、定量地分析多肽的氨基酸组成，使大分子量蛋白质分子一级结构研究和序列测定成为可能，利用该技术人们先后完成了人血红素蛋白 α 和 β 链、鸡溶菌酶素、牛核酸酶、烟草花叶病毒外壳蛋白等一系列分子量相对较大的蛋白质的序列测定工作。到 1965 年，已有约 20 个长度为 100 多个氨基酸的蛋白质序列被测序。但这一时期受仪器敏感性的限制，只有那些能够获得克量级的纯化蛋白质，并且其结构比较简单的情况下，才可能对其进行序列分析。其主要限速步骤为繁琐的单向纸层析或纸电泳多肽分离技术以及缺乏灵敏的分析仪及有效的蛋白或多肽纯化手段，并且此时自动蛋白组分收集仪器还未出现，只能手工收集组分，这都大大限制了蛋白质序列测定的进程。1967 年，Edman 等发明了自动 Edman 降解程序，可以用 7 mg 的纯化蛋白在短时间内完成 60 步的多肽连续测序，使得蛋白质测序工作有了一个质的飞跃。1980 年开始使用的自动测序仪，其灵敏度增加了近 10000 倍。目前，随着双向凝胶电泳、电转移印迹法、高效液相色谱法、灌注层析技术、蛋白质组分自动收集仪以及高敏感度的氨基酸分析仪和质谱分析技术的出现，使得多肽分离、纯化以及分析工艺得到了极大的提高，目前已经可以对皮摩尔的样品进行序列分析。特别是双向凝胶电泳和质谱分析技术，已成为目前蛋白质研究的最有效的蛋白分离和分析手段，推动了蛋白质序列分析的迅速进展。据统计，截止 2002 年已测序的蛋白质序列已经达到 108159 条。

二、核苷酸序列测定技术

20 世纪 60～70 年代，科学家们一直致力于研究测定核酸序列的方法，最初使用的方法只能测定核糖核酸，且主要是转运核糖核酸（tRNA）。由于 tRNA 分子链较短，通常只有 74～95 个核苷酸，而且相对比较容易分离纯化单个分子，tRNA 分子序列测定比 DNA 分子序列测定要容易的多。随着对大量序列数据的需求，迫使人们寻求更快、更敏感、更精确、更经济的测序技术。1977 年，在 DNA 聚合酶以及 DNA 荧光标记技术的基础上，Sanger 发明了核苷酸的双脱氧核糖核酸链末端终止法测序，极大地改进了 DNA 测序的速度。荧光自动分析仪的发明标志着大规模测序的开始，它能以每天超过 100000 碱基对的速度测序，同时使其费用控制在 50 美分/碱基以下，并随着 DNA 聚合酶的不断修饰和荧光底物的不断更新，在很长的一段时间里，荧光测序法保持着 DNA