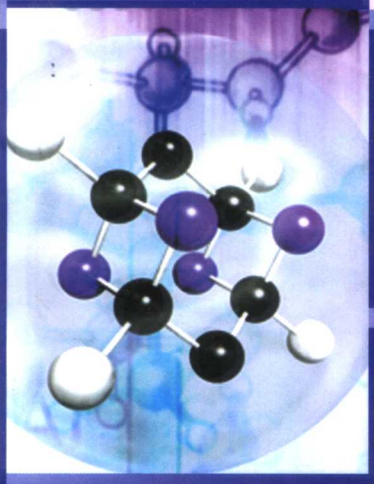
普通高等教育“十一五”规划教材



生物信息学

陶士珩　主编

科学出版社
www.sciencep.com

普通高等教育“十一五”规划教材

生物信息学

陶士珩　主编

科学出版社

北京

内 容 简 介

本书主要涉及生物信息数据库、序列比对、分子系统发育分析、基因组学与基因预测、蛋白质结构与功能预测、转录组与蛋白质组分析以及Perl语言在生物信息学中的应用等内容，力求使读者全面了解和掌握生物信息学领域的重要基础知识与基本操作技能。

本书可用作全国农林院校生命科学各专业相关课程的教材。同时，也供有关科研人员参考使用。

图书在版编目(CIP)数据

生物信息学/陶士珩主编. —北京：科学出版社，2007
（普通高等教育“十一五”规划教材）
ISBN 978-7-03-019771-9

Ⅰ. 生… Ⅱ. 陶… 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（2007）第131997号

责任编辑：甄文全 彭克里 刘 晶/责任校对：刘小梅
责任印制：张克忠/封面设计：科地亚盟

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

双青印刷厂印刷

科学出版社发行 各地新华书店经销

*

2007年8月第 一 版 开本：787×1092 1/16
2007年8月第一次印刷 印张：17 1/4
印数：1—3 000 字数：400 000

定价：28.00元

（如有印装质量问题，我社负责调换〈环伟〉）

前　言

什么是生物信息学？无论是生物信息学的初学者还是其研究者都难以准确回答这个问题，这是由生物信息学专业的特点所决定的。生物信息学领域涉及面十分广阔，在核酸层面，从单个基因进化分析到全基因组的分析；在蛋白层面，从蛋白质一级序列的进化分析到复杂的蛋白质三维结构分析；从控制基因表达的顺式作用元件及启动子的研究到基因的表达的峰度及时间的微矩阵研究。有关生物信息学的定义，见仁见智。按照目前公认的观点，生物信息学可以广义地定义为：利用数学、计算机科学以及统计学等知识来组织和理解生物分子（主要指核酸和蛋白）中所蕴含的信息。简而言之，生物信息学就是分子生物学的信息管理工具的集成。生物信息学对于现代生物研究有着巨大的推动作用，是当代生物学工作者必备的专业基础之一。

本书各章通过介绍生物信息数据库、序列比对、分子系统发育分析、基因组学与基因预测、蛋白质结构与功能预测、转录组与蛋白质组分析以及 Perl 语言在生物信息学中的应用等内容，力求使读者全面了解和掌握生物信息学领域的重要基础知识与基本操作技能。考虑到本书主要面对农林专业师生，本书在编写时侧重于方法的讲解和应用，而对于算法的理论背景则简要叙之，此举并非忽略生物信息学基础理论对该领域发展的重要意义与价值，而是意在使广大读者跳出艰深的理论樊笼而关注应用，因为应用性是生物信息学的重要特征之一。

本书由陶士珩担任主编，庞红侠、吴宪明、郭凤霞和范丙友担任副主编。全书一至十章分别由山西农业大学唐中伟、青岛农业大学穆平、西北农林科技大学庞红侠、甘肃农业大学郭凤霞和范丙友、四川农业大学刘汉梅、西北农林科技大学范三红、河南科技大学范丙友、内蒙古农业大学陈有君、沈阳农业大学吴宪明和王振东以及内蒙古农业大学张文广编写。陶士珩、庞红侠、吴宪明、郭凤霞等共同统稿，吴宪明担任最后校对。本书得以顺利出版，全赖全体作者齐心协力、努力合作。

尽管出版前再三修改，由于编者水平有限，错谬之处在所难免，敬请国内同行及读者指正。

编　者

2007 年 8 月

目　录

1 绪　论

1.1 生物信息学的历史

生物信息学就其概念而言，可以追溯到1956年在美国田纳西州Gatlinburg召开的首次“生物学中的信息理论研讨会”。20世纪80年代末期，林华安博士认识到将计算机科学与生物学结合起来的重要意义，开始留意要为这一领域构思一个合适的名称。起初，考虑到与将要支持他主办一系列有关生物信息会议的佛罗里达州立大学超大型计算机计算研究所的关系，他使用的是“compbio”。之后，他又将其更改为兼具法国风情的“bioinformatique”，看起来似乎有些古怪。因此不久，他便进一步把它更改为“bioinformatics（或bio/informatics）”。但由于当时的电子邮件系统与今日不同，该名称中的“-”或“/”符号经常会引起许多系统问题，于是林博士将其去除。1987年，林华安博士正式把这一学科命名为“生物信息学”（bioinformatics）。

20世纪60年代，Zucherkandl和Pauling就开创了分子进化这一个全新的研究领域，主要是通过序列分析研究序列变化与进化之间的关系。这一时期，Dayhoff和他的同事们收集了当时已知的氨基酸序列，这就是“蛋白质序列与结构图册”，这一蛋白质数据库后来成为著名的蛋白质信息源PIR。

20世纪70年代到80年代初期，生物化学技术有了很大的发展，DNA测序方法的出现，产生出许多生物分子序列数据。如何根据序列推测结构和功能，这一生物学问题促使了一系列著名的序列比较方法的出现。其中，Needleman-Wunsch序列比对算法的提出是生物信息学发展中最重要的里程碑。同年，Gibbs和McIntyre发表的矩阵打点作图法也是进行序列比较的一个著名方法。Dayhoff提出的点突变模型的PAM矩阵作为比较氨基酸相似性的得分矩阵，这一矩阵的广泛使用大大地提高了序列比较算法的性能。1980年*Science*第209卷发表了关于计算分子生物学的综述。1981年，Smith和Waterman提出了著名的局部对位排列算法，同年，Doolittle提出关于序列模式的概念。1982年，著名的GCG分子计算工具出现。1985年，FASTP序列分析运算法则发表。1988年，Pearson和Lipman发表了著名的FASTA序列运算法则。1990年，目前常用的数据库搜索程序BLAST建立。1997年，BLAST的改进版本PSI-BLAST投入实际应用。

20世纪80年代以后，作为储存生物信息重要组成部分的生物信息数据库陆续建立。1982年，欧洲分子生物学实验室EMBL诞生，提供核酸序列数据库服务。同年，美国国立卫生研究院下属的国立生物技术信息中心建立了GenBank。1986年，日本核酸序列数据库DDBJ诞生。作为国际上三大生物数据库中心，之后的1988年三方达成一项协议，采用共同的数据库记录格式收集直接提交的数据。1986年，出现蛋白质数据库SWISS-PROT。

20世纪90年代后，由美国、英国、法国、德国、日本和我国共同参加的国际人类

基因组（human genome project，HGP）计划正式启动，这一计划旨在对人类基因精确测序，发现人类所有基因并搞清其在染色体上的位置，破译全部遗传信息。在人类基因组计划进行中，许多生物的基因组研究也相继进行。1995 年，第一个细菌全基因组序列——流感嗜血杆菌测定，这是人类拥有的第一个全基因组信息。1996 年，第一个真核生物基因组——面包酵母基因组被完全测序。同年，Affymetrix 推出其第一块基因芯片（gene chip）。1997 年，第一个实验模式生物——大肠杆菌的基因组完成测序。1998 年，第一个多细胞生物——线虫的基因组被完全测序。2000 年，第一个植物拟南芥基因组和果蝇的基因组被完全测序。2001 年，人类基因组草图在 *Nature* 和 *Science* 同时发表。2002 年，小鼠和水稻基因组草图完成。2003 年，历经十余年的人类基因组计划完成。同年中国科学家首先完成非典型性肺炎病毒全基因组测序。2005 年，历时 6 年的国际水稻全基因组测序圆满完成。2006 年，国际研究组织的科学家完成了牛基因组的测序。这些基因组计划的进行极大地促进了生物信息学的发展。

近十年来，科学家完成了包括人类基因组在内的几十种生物的全基因组测序，产生了海量的数据，而这些生物数据的量以摩尔定律呈指数增长。同时，计算机的发展速度每隔 18 个月就翻一倍。但从生物信息学的发展而言，直到 20 世纪 80 年代后，随着生物科学技术的飞速发展，特别是计算机科学的进步，才极大地促进了生物信息学进展。在 Internet 普及的今天，应用计算机科学和网络技术来解决生物学问题，已经使得生物信息学成为生命科学工作者必需的工具之一。生物信息学成为生命科学的必不可少的一部分，它的诞生和发展是科学和历史的必然。

1.1.1 生物信息学与生命科学的发展

21 世纪是生命科学的世纪，生物科学是自然科学中发展最迅速的学科之一，人类对于生命奥秘的探索从未停止过。生物学作为自然科学中的一个基础学科，经历了从博物学、生物学、生物科学到生命科学的发展历程，从对自然生物的描述进入了结构功能、系统演化现象本质的研究，建立了生命科学的体系。达尔文发表的《物种起源》是生物学史上第一部关于生物进化的划时代著作。孟德尔发现的遗传学三大定律被认为是生物遗传的最基本规律，而 Watson 和 Crick 发现的 DNA 双螺旋结构及核酸是生命本质的一系列重大发现，为生物学的发展奠定了坚实的基础，从此生物学正式摆脱了博物学的那种仅依靠观察、比较的方法，发展成为一门实验性学科。生命科学作为实验和理论紧密结合的学科，它的研究手段进入了实验与模拟这一综合的研究方法体系。

在生命科学的研究中，一方面，人们认识到用物理、化学和生物学方法研究生命的物质基础、能量转换、代谢过程等的重要性；另一方面，人们也认识到需要用信息科学的方法来研究生命信息，理解生命的工作机制，揭示生命的奥秘。DNA 分子和蛋白质分子携带了遗传信息、功能信息及进化信息，成为生物信息的主要研究物质。

生命科学研究的层面进入了分子水平，DNA 是遗传信息的载体。DNA 的核苷酸序列包含蛋白质的氨基酸序列编码信息、基因表达调控信息等遗传信息。遗传信息存储在 DNA 四种碱基字符组成的序列中，包含生物体生长发育的各种遗传信息。因此，DNA 序列是包含着最基本的生命信息的载体。蛋白质是一切生命的物质基础，没有蛋白质就

没有生命。蛋白质由 20 多种氨基酸组成，由于氨基酸组成的数量和排列顺序不同，使人体中蛋白质多达 10 万种以上。它们的结构、功能千差万别，形成了生命的多样性和复杂性。蛋白质功能的多种多样是由蛋白质的空间结构决定的。蛋白质结构作为一种重要的生物大分子信息与蛋白质序列密不可分，因此蛋白质序列可以说是包含结构信息的重要数据。

生物信息学的对象就是生物学数据，更直接地说就是分子生物学数据。随着分子生物学及其实验技术的日趋成熟，生命科学研究进入了前所未有的高速发展时期。以 DNA 序列为主的生物数据，其增长速度更是爆炸性的。今天在一个大型基因组测序中心，每天测序可产生以千万计的序列数据，每秒产出的碱基可达 1000 个。以人类基因组计划为开端的测序计划以及酵母、大肠杆菌、线虫、果蝇、小鼠、拟南芥、水稻、玉米等其他一些模式生物的基因组计划的相继完成或全面实施，使有关核酸、蛋白质的序列和结构等的分子生物学数据呈指数增长。由生物数据所产生的库多种多样，生命科学产生的生物数据储存在相关领域的数据库中，包括核酸序列数据库、蛋白质序列数据库、蛋白质结构库、酶学数据库、蛋白质双向电泳数据库、小分子配体化合物数据库，等等。

在海量生物数据产生的同时，计算机处理数据的能力增长也是以摩尔定律指数增长。计算机技术为海量生物数据的处理提供了有力的技术支持。各种数据计算方法的出现和网络的普及为生物信息学的发展拓宽了道路。

对于分析和处理生命科学产生的生物数据的生物信息学概念，有不同的定义。从广义上来说，生物信息学是指运用数学、计算机科学和生物学工具对生物学数据进行储存、检索、分析和处理，以达到揭示生物数据中的生物学内涵的目标。狭义的定义是指利用计算机技术作为研究手段和工具对生物学数据进行管理和分析。简单而言，生物信息学是从生物数据中提取新知识的学问。

生物信息学以基因组 DNA 序列信息分析作为出发点，辨别隐藏在 DNA 序列中的基因，找到基因组序列中代表蛋白质和 RNA 基因的编码区，阐明编码区的信息实质，破译遗传语言，认识遗传信息的组织规律。同时掌握基因调控信息，从而认识发育、进化的规律。

当前核酸分子和蛋白质分子序列的一级结构序列分析和蛋白质三维结构仍是生物信息学研究的热点。

生物信息学作为生命科学研究所必需的研究工具，在生命科学实践中越来越显示出它的重要作用，特别是在实验设计、结果分析上，离不开生物信息学的指导。

1.1.2 生物信息学的发展历程及方向

生物信息学的发展大致经历了 3 个阶段。

第一个阶段是前基因组时代。这一阶段主要是以各种算法法则的建立、生物数据库的建立以及 DNA 和蛋白质序列分析为主要工作。在这一阶段，著名的 Needleman-Wunsch 和 Smith-Waterman 序列比对算法先后发表；国际上的三大核酸序列数据库（EMBL、GenBank、DDBJ）相继建立并提供序列服务。

第二个阶段是基因组时代。这一阶段以各种基因组计划测序、网络数据库系统的建

立和基因寻找为主要工作。以人类基因组计划和各种模式生物基因组测序为主要工作，大规模测序全面铺开。

第三个阶段是后基因组时代。这一阶段的主要工作是进行大规模基因组分析、蛋白质组分析以及其他各种基因组学研究。随着人类基因组计划和各种基因组计划测序的完成，以及新基因的发现，系统了解基因组内所有基因的生物功能成为后基因组时代的研究重点。生物信息学进入了功能基因组时代。

人类基因组计划的完成标志着后基因组时代的开始。蕴藏着人类生命遗传奥秘的遗传语言由 30 亿个碱基对组成，破译遗传语言的人类基因组计划被誉为生命科学的“登月”计划，它耗时 13 年，于 2003 年完成。各国分别承担工作比例为美国 54%，英国 33%，日本 7%，法国 2.8%，德国 2.2%，中国 1%。许多模式生物基因组计划完成了测序工作，更多的生物基因组正在被列入测序计划中。人类掌握了极大量的遗传数据，期待揭示其中的生命奥秘。这也标志着基因组学进入了揭示基因功能的阶段，即功能基因组时代。功能基因组研究基因组的组成和功能，认识基因与疾病之间的关系，掌握基因的产物及其在生命活动中的作用。功能基因组从基因组整体水平上研究生命活动规律，特别是在研究基因与疾病的关系上，改变了生物学研究的传统方法。传统上研究基因和蛋白质的功能只能通过假设的方法，经过反复实验对候选基因进行逐一验证，才能确定所选基因与其表型相关。这种传统的研究方法周期长、花费大，而且依赖于能观测到的表型，不能全面了解所有相关基因及其功能。而生物信息学新的研究方法从生物基因组这个整体水平上，通过从基因到表型的途径，直接对其生物功能进行研究。这种新的实验设计思路缩小了研究对象的范围，实现了功能预测、实验证实，极大地推进了研究速度。

生物信息学步入后基因组时代，其发展方向主要有以下这些方面。

1）各种生物基因组测序及新基因的发现

人类基因组和许多模式生物的基因组测序已经完成，接下来的工作是对更多生物基因组的测序，获取更多物种的全部基因。这是基因组研究的首要工作。大规模测序工业化，自动基因测序技术日趋成熟，使得基因组的测序速度得到极大提高。目前以这样的设备和技术完成一个细菌基因组的测序只需一周时间，完成一个平均大小的高等真核生物基因组也缩短至 2 年内。测序基因组过程是这样的：将基因组 DNA 切断成小片段，分别测序，再将它们拼接起来。这一看似简单的过程，实践起来却是相当复杂的。这样在大规模测序中，小片段序列的拼接和填补序列是保证序列测定准确性和高效性的关键。全基因组鸟枪法（whole genome shotgun）这一测序战略在基因测序上的成功应用，使得测序速度倍增。而更重要的是在测序过程中适当算法和软件的应用，以解决其中具有高度重复序列的海量数据。目前广泛用于大规模测序和拼接的软件包有美国华盛顿大学研发的 Phred-Phrap 软件包和加利福尼亚大学研发的 GigAssembler 软件包。

从得到的基因组序列中发现新基因是生物信息学研究的热点之一。对于基因组较小的原核生物和一些真核生物，通过基因组学的理论方法预测其中新基因是可行的。例如，酿酒酵母所包含的 6000 个基因，大约 60%是通过数据分析得到的。对于从人类基因组这类复杂的基因组中发现新基因，可以利用 EST 和比较基因组学方法进行研究。

表达序列标签（expressed sequence tag，EST）是从 cDNA 文库中生成的一些很短的序列，长 300～500bp，它们代表在特定组织或发育阶段表达的基因，有时也可代表特定的 cDNA。因此 EST 的研究可以作为发现一种新基因的有效方法。通过比较基因组学的多种分析方法，可以从与已知基因和蛋白质的序列同源性得到证据，区分出基因组上编码蛋白质区域和非编码蛋白质区域，从而确定新基因的编码区。

2）单核苷酸多态性（SNP）分析

单核苷酸多态性（single nucleotide polymorphism，SNP），是指在给定的一个群体中，超过 1%的个体在给定的遗传区域内发生一次核苷酸改变。在群体的分布中，基因的多态性使得生物表型表现出对外界物质的反应各不相同，这些便是基因 SNP 造成的结果。SNP 也就是相同基因在不同个体中存在的单个碱基上的变异所造成的基因差异表现。通常每 100～300 个碱基就会有一个 SNP。SNP 被认为是一个物种中不同个体表型差异的主要遗传来源。随着人类基因组研究的深入，SNP 研究成为生物信息学研究的热点之一。作为研究序列变化与可遗传的表型变化的有力工具，SNP 在基因与疾病方面大有可为。目前 SNP 研究的工作包括制作 SNP 图谱，找出适合的 SNP 作为重要的遗传标记。通过与健康个体 SNP 的比较，找出与遗传疾病相关的 SNP 标记。SNP 在人类基因与疾病的研究中将发挥越来越大的作用。

3）基因组非编码区信息结构与分析

对于生物完整基因组，原核生物与真核生物的非编码区域占整个基因组的比例大不相同。微生物中的原核生物所含非编码区或内含子非常少，如细菌中只占 10%左右；对于高等生物和人的基因组，它却占了很大的比例。人类基因组测序完成后的研究表明，真正编码蛋白质的区域只占 5%左右，95%的区域是非编码区。从生物进化的角度看，这样庞大复杂的非编码区必然包含着与生物进化有关的信息。因此生物体的复杂结构和功能不仅仅是由基因决定的，也不仅仅是由基因组中大量的非编码信息决定的，而是由这些元素在生物体各个层次上复杂、动态的相互作用决定的。

4）比较基因组学和生物进化研究

大量生物基因组计划完成测序任务，产生了海量的生物数据，为科学家研究各种基因组之间的进化关系提供了丰富的材料。比较基因组学正是在基因组水平对各种生物进行比较，发现蛋白质功能，揭示生命起源和进化。它是通过比较人类基因组与其他模式生物基因组，从而为了解人类基因组结构、发现新基因和功能提供依据。在人类基因组计划实施的同时，大肠杆菌、酿酒酵母、线虫、果蝇、小鼠这 5 种模式生物的基因组也在同时进行。这些模式生物的基因组相对于人类基因组研究和分析更方便，为人类基因组的新基因提供了预测的依据。不少人类的基因可以通过对比模式生物的已知基因后发现，这是比较基因组学的应用之一。

比较基因组的重要研究方向之一就是生物进化。达尔文研究生物进化的方法是比较生物形态学特征获得进化信息。而今天进化的研究进入了分子水平，比较序列间碱基或氨基酸的差异，就可以获取有关进化的深层次信息。各种完整基因测序工作的完成，使

得序列比较不再是单独片段的比较，而是多种生物序列整体的比较。构建系统进化树是描述生物进化的重要步骤，可以根据序列同源性分析的结果使用 PYLIP 软件包完成。而同源序列的比较可以由多序列比对程序诸如 CLUSTAL 程序确定。其中同源性序列要首先进行序列相似性比较获得。这里相似只是类似，而同源则是有共同祖先的相似。同人类基因组十分相近的小鼠基因组，基因的数目与人类相近而且大部分同源，但人和鼠的表型差异相去甚远。这种差异是基因组的组织上的差别造成的。存在于鼠 1 号染色体上的基因已分布到人的 1、2、5、6、8、13、18 号 7 个染色体上了。或许正是基因组的组织的差异导致了表型上的显著差异。通过比较基因组学构建系统进化树可以对生命的起源、生物的进化等若干重大生物学问题进行分析研究。同时随着基因组数据的大量增加，对序列差异和进化关系的争论也越来越激烈。由某一种分子序列构建的进化树并不一定能代表物种之间的真正进化关系，同时对“垂直进化”和“水平演化”之间关系的讨论正逐渐受到人们的重视。正是全基因组的比较研究使得生物信息学的研究实现了片面向全面的突破。

5）蛋白质结构和功能的研究

随着人类基因组计划和各种生物基因组全序列测序的完成，了解这些基因编码的蛋白质有什么功能成为备受关注的问题。而且在细胞合成蛋白质之后，这些蛋白质往往还要经历翻译后的加工修饰。也就是说，一个基因对应的不是一种蛋白质而可能是几种甚至是数十种蛋白质。包容了数千甚至数万种蛋白质的细胞是如何运转的？或者说这些蛋白质在细胞内是怎样工作、如何相互作用、相互协调的？这些问题远不是基因组研究所能回答得了的。正是在此背景下，蛋白质组学（proteomics）应运而生。

蛋白质组学以研究一种细胞、组织或完整生物体所拥有的全部蛋白质为特征。它的研究目标是分析蛋白质间相互作用和蛋白质的功能。蛋白质组研究的对象不是单一少数蛋白质，它更强调从全面性和整体性角度来揭示和阐明生命活动的规律。

要了解基因编码的蛋白质的功能，只有氨基酸排列顺序是远远不够的。蛋白质功能的实现是依靠其空间三维结构执行的。了解蛋白质的三维结构是当务之急。在核酸序列数据增长的同时，蛋白质序列数据也在迅速增长，目前已知的蛋白质序列已过百万，而被测定的蛋白质结构只有 2 万多，其原因在于以 X 射线晶体学技术、多维核磁共振（NMR）波谱学技术为主要方法的蛋白质空间结构测定手段以每天只能得到几个生物大分子空间结构的速度前进。依靠这样的实验方法很难满足蛋白质三维结构研究进度的需要。因此蛋白质空间结构预测成为生物信息学研究的焦点之一。所谓蛋白质空间结构预测是指从蛋白质的氨基酸序列预测出其三维空间结构。这就需要发展一种预测蛋白质结构的新方法。这种新方法以我们已知结构的蛋白质为基石，采用计算的方法找到或预测与已知结构相似的结构。进行蛋白质结构预测首先要将待测蛋白质与数据库中的同源蛋白质进行序列比对，然后根据计算比对，依次进行二级结构、三级结构的预测。这实际上是利用同源相似性，推测待测同源蛋白质的结构。

蛋白质结构预测对于理解蛋白质结构与功能的关系，加强蛋白质工程研究及基于结构的药物分子设计具有十分重要的意义。蛋白质工程是以蛋白质结构和功能的关系为基础，利用基因工程技术，设计和定向改造蛋白质，构建出性能更符合人类需要的新蛋白

质。以蛋白质结构为设计基础的药物分子设计，其作用的靶标部位具有特定的空间结构。只有对靶标部位空间结构有了充分了解，才能设计出有效的药物分子。这些都离不开对蛋白质空间结构的深入认识。

1.2 生物信息学的应用

1.2.1 基因组分析

人类基因组计划和其他各种基因组计划完成后产生的最直接的数据就是大量的基因组 DNA 序列。作为最主要的生物信息基本数据，从这些 DNA 序列中寻找新基因成为生物信息学最主要的应用。从编码区域推导出基因的结构及其对应的蛋白质序列。发现新基因也就是寻找可以编码蛋白质的序列。对于原核生物来说，因为其中不含内含子，找到具有起始密码的可读框，预测出功能基因相对容易些。而真核生物基因有内含子和外显子并进行选择性转录，相对复杂许多。现阶段主要应用的方法有从已知 cDNA 及表达序列标签（EST）序列比对得到证据，从已知基因的蛋白质序列同源性得到证据，以及从相近物种间基因比对得到证据。此外，还有采用隐马尔可夫模型（HMM）和神经网络（neural network）在内的学习的方法来识别剪切位点、密码子使用偏爱及外显子和内含子长度。而且现在有许多软件用于预测识别真核基因的编码区，诸如 Genpaser、GenScan、GeneFinder 等。这些方法从识别基因组的外显子和内含子及剪切位点上提供了预测，但同时多基因的不同组合问题仍是预测外显子的难点。

基因组分析除了发现新基因外，分析非编码区的工作也同时在进行。主要工作集中在基因表达调控及基因转录调控元件等的分析研究。通过对基因组编码区和非编码区的深入研究，必将对基因组结构信息组织的规律有更全面的认识。

1.2.2 基因芯片

生物芯片（biochip）是由微电子学、物理学、化学、计算机科学与生命科学交叉综合的高新技术。生物芯片的概念源于计算机芯片的概念。在计算机芯片上排列的是集成电路，而生物芯片上排列的是密集的探针阵列。生物芯片与生物信息学的关系是密不可分的。基因芯片（gene chip）是生物芯片的一种，又称 DNA 微阵列（DNA microarray），是由大量 DNA 或寡核苷酸探针密集排列所形成的探针阵列，它的基本工作原理就是通过杂交检测信息。通过在片合成法或点样法将数万个探针排列在硅片等载体上，利用芯片扫描仪和相关软件可以分析图像，将荧光信号转换成数据，即可以获得有关生物信息。基因芯片为生物信息学提供大量的研究和分析内容，而生物信息学的方法促进了基因芯片的研究与应用。

生物信息学在基因芯片的研究主要体现在确定芯片待检测的目标序列、芯片设计、检测结果分析和数据管理等工作上。首先，根据所要解决的问题，通过查询生物信息数据库和信息分析，确定基因芯片所要检测的目标。然后根据基因芯片具体的功能要求，采用特定的方法进行探针设计和布局，并进行芯片优化。根据芯片设计结果制备芯片，

进行杂交实验。最后采集并处理芯片杂交后的荧光图像，结合数据库中的芯片描述确定基因芯片检测结果，并对检测结果进行可靠性分析。

基因芯片技术作为一种先进的、大规模、高通量检测技术，其优点有以下几个方面：一是高度的灵敏性和准确性；二是快速简便；三是可同时检测多种基因。这些优点使得基因芯片在生命科学的诸多领域得到应用，特别是在遗传研究、疾病诊断和治疗、新药发现和环境保护等方面。利用基因芯片技术研究生命个体不同生长发育阶段的基因表达，了解多基因协同作用的生命过程，比较正常和疾病状态下基因及其表达的差异，研究生物体在进化、发育、遗传过程中的规律。基因芯片技术给基因组研究提供了强有力的研究手段，使生命科学的研究进入了一个崭新的阶段，它必将在后基因组时代发挥更重要的作用。

1.2.3 药物设计及经济价值

人类基因组计划的重要目的之一就是破解人类的遗传密码与人类疾病之间的关系，为人类的健康和疾病治疗提供依据。基于生物大分子结构的新药设计是生物信息学的主要应用领域，是对传统药物寻找的革命性创新，对现代与未来药物学和药理学产生了重大影响。

根据生物信息学研究产生的生物大分子空间结构的信息，选定对疾病防治起决定作用的靶标结构分子，设计和筛选可以对靶标分子有高活性的药物。这将为新药筛选、药靶设计和分子药理学研究，以及疑难病的药物设计和途径选择等提供新的方法论基础。药物研发的一般流程是：首先是靶标分子的鉴定、候选药物的筛选、药物作用机制的研究、药物动力学和毒性研究等。在进行药物靶点研究的同时，应用生物信息学技术和计算机辅助筛选相结合，开辟了新的药物发现途径。在生物信息学研究的基础上，利用获得的蛋白质结构和功能信息，用计算机辅助药物设计，加快了药物发现的速度。现在许多制药公司充分应用药物基因组学及生物信息学其他分支学科的理论知识和技术手段来设计临床实验并模拟和分析理论与实验数据。这将大大减少新药开发成本，缩短开发周期，为患者、医生和健康医疗机构等诸方面带来选择性治疗的革命。

新药设计是生物信息学研究与开发的主要领域，对生物信息的需求也是巨大的。生物信息蕴藏着巨大的经济价值。生物信息产业成为许多国家和公司投资的重点之一。许多制药公司纷纷投巨资建立自己的生物信息数据库，研发自己的专利，储备生物信息资源。由生物信息学产业和带动的相关产业所带来的经济价值不可估量。

1.3 生物信息学与其他学科的关系

1.3.1 生物信息学是多学科的整合

生物信息学是整合了生物学、统计学、应用数学、计算机科学的交叉学科。它以计算机为工具，利用各种由数学和统计学建立的算法和模型，对生物学数据进行检索、处理和分析，从而阐明大量生物数据所包含的生物学意义。通常这些数据与分子生物学密切相关。从狭义上讲生物信息学可以称作分子生物信息学。

与生物信息学相关的概念有计算生物学和 DNA 计算。计算生物学以开发和应用数

据分析及理论的方法、数学建模、计算机仿真技术等为主，侧重于理论研究，应用范围不如生物信息学广。DNA 计算则是以 DNA 为信息储存器，应用 PCR 技术、生物芯片等进行计算，是计算机科学与分子生物学相结合的产物，也是计算机发展的新途径。

作为一门正在迅速发展的学科，生物信息学渗透到了生命科学各个领域。随着生命科学研究的深入，将会有更多的相关学科融入到生物信息学中去，它将成为一门集众多学科而成的交叉学科。

1.3.2 生物信息学与计算机科学

1.3.2.1 数据库技术及算法

在生物信息学中，数据库技术是最基本的技术。生物分子信息的存储、管理、查询等功能都是建立在数据库管理系统之上的。

数据库技术产生于 20 世纪 60 年代末，随后得到了迅速发展。其应用领域不断扩大，已经由最初的一般性事务处理发展到情报检索、人工智能、计算机辅助设计等各个领域。尤其是 Internet 的出现以及多种信息技术的交汇，给数据库技术的应用提供了更广阔的空间，从而促进了数据库技术的快速发展，数据库的地位也与日俱增。

数据库系统是指引进数据库技术后的计算机系统。它包括：①以数据为主体的数据库和管理数据库的软件系统，即数据库管理系统；②支持数据库系统的计算机硬件系统和操作系统；③数据库管理员和用户等。其中，数据库等资源存储在外存储器上；数据库管理系统是为数据库的建立、使用和维护而配备的软件；操作系统是数据库管理系统与硬件的接口，用户通过数据库应用系统在逐层系统的支持下使用数据库，而数据库管理员（database administrator）负责对整个系统的管理工作。

计算机数据的管理主要经历了人工管理阶段、文件系统阶段、数据库系统阶段和分布式数据库系统阶段这几个发展阶段。数据库的种类繁多，大的数据库系统如 Oracle、Informix、Sybase 和 DB2 等，小的如 Foxpro、dBase、Access 等。它们各有所长，能分别满足不同层次的需要：Oracle 以稳定性著称，Informix 因先进性闻名，它们适合建立工程、企业等大型数据库；而 Foxpro 简单快速，Access 小巧便捷，能很好地为家庭及中小型数据库服务。

随着计算机应用领域的迅速扩大，新一代数据库的应用变得更加广泛，数据库技术与其他学科的技术内容互相结合。多学科的技术内容与数据库技术的有机结合，使数据库领域中新的技术内容层出不穷。最新的数据库管理系统为：扩展关系数据库系统、面向对象的数据库系统、分布式数据库系统、并行数据库系统，以及专家数据库系统、数据仓库等，也以其先进强大的技术支持和日趋完善的管理功能逐渐影响和渗透到其他各个领域。

生物信息学中涉及大量的计算，像序列比对、基因组信息分析、系统发生分析、蛋白质结构预测、基因表达分析等，不可避免地用到算法的概念，特别是对算法的计算复杂度尤为关注。究其原因，不外乎生物信息学所处理的对象十分复杂而又庞大。为在现有的计算机上完成工作，不得不考究算法的优劣。从计算理论的角度来讲，它们都是难处理的。换句话讲，我们并不知道是否存在有效的算法去解决这些问题。目前的研究集中在设计好的近似算法或概率算法，这些算法虽然并不能对有关问题的每一实例都能求

出好的解，但对绝大多数实例却行之有效。

1.3.2.2 计算机与 Internet

计算机是生物信息学研究必不可少的工具，离开了计算机生物信息学可以说是无从谈起。

从第一台计算机发明到今天，计算机经历了 60 年的发展。从只有政府和企业才买得起的大型机、中型机、小型机，发展到了今天广泛普及的微型计算机。它的性能更是以摩尔定律成倍的增长。计算机技术向着巨型化、微型化、网络化、多媒体化和智能化发展。从微机的硬件发展来看，以 Intel 公司为主流的 CPU 从 286、386、486、586，发展到今天的双核处理器，内存进入了 G 时代，硬盘更是跨入了百 G 的超大容量。无论是速度还是容量都有了巨大的提高。就操作系统来看，微软的视窗操作系统进入了 Vista，功能更加强大，安全性和稳定性日益提高。除了 Windows 操作系统，目前在不同计算机上安装的还有 UNIX、Linux、MacOS 等操作系统。许多生物信息学服务网站目前多采用 UNIX、Linux 或 Windows 系统。与生物信息学研究密切相关的计算机语言，目前应用的有 C 语言、Perl 语言、PHP 语言、JAVA 语言等。

计算机技术与通讯技术相结合产生了计算机网络。Internet 的出现，使人类进入了网络时代。Internet 作为今天最大的网络系统，为资源共享和信息交流提供了广阔的天地。通过互联网，使得各种生物信息资源存取极为方便。网络促进了生物信息学的研究，加快了生物信息数据的更新与整合，推进了生命科学的前进步伐。为了更快地找到所需的信息，基于互联网的搜索引擎发展迅速，百度和谷歌是当今最主要的两个，通过输入关键词，它们可以很快的列出与之匹配的相关链接供查询者选择。互联网提供了丰富的生物信息，在浏览的同时，通过下载所需数据可以在本地计算机上进行使用。电子邮件（E-mail）作为信息时代必不可少的信息交流方式，它方便即时，被广为使用。对于有些耗时长的生物序列检索，待服务器计算完成后可自动将结果返回用户的 E-mail 信箱。

掌握更先进的计算机技术，对于生物信息学的研究者来说是必不可少的，它将为生物信息学的未来提供更有力的支持。

结束语

人类进入了信息时代，正在经历着一场深刻的革命。信息时代带来了挑战，更带来了机遇。生物信息学作为多学科交叉的新兴学科，受到了世界各国的重视。我国虽然起步较晚，但对生物信息学的研究和应用十分重视。北京大学和中国科学院上海生命科学研究院分别成立了两个专业水平较高的生物信息学网站，我国在一些领域的研究上也取得了较好的成果，但与国际上的领先研究水平仍有一定的差距。面对生物信息学发展的强劲势头，我们应抓住这一时机，加强生物数据库的建立，进行相关方法的研究，加快人才的培养和训练，为我国生物信息学的发展奠定良好的基础。

（唐中伟　编）

2 生物数据库介绍

近年来大量生物学实验数据的积累，形成了当前数以百计的生物信息数据库。它们各自按一定的目标收集和整理生物学实验数据，并提供相关的数据查询、数据处理的服务。随着因特网的普及，这些数据库大多可以通过网络或者网络下载来访问。

一般而言，这些生物信息数据库可以分为一级数据库和二级数据库。一级数据库的数据都直接来源于实验获得的原始数据，只经过简单的归类整理和注释，包括基因组数据库、核酸和蛋白质一级结构序列数据库、生物大分子（主要是蛋白质）三维空间结构数据库。国际上著名的一级核酸数据库有 GenBank、EMBL 和 DDBJ 等；蛋白质序列数据库有 SWISS-PROT、PIR 等；蛋白质结构库有 PDB 等。二级数据库是在一级数据库、实验数据和理论分析的基础上针对特定目标衍生而来的，是对生物学知识和信息的进一步整理。二级生物学数据库非常多，它们因针对不同的研究内容和需要而各具特色，如人类基因组图谱库 GDB、转录因子和结合位点库 TRANSFAC、蛋白质结构家族分类库 SCOP 等。

下面将依次简要介绍一些著名的和有特色的生物信息数据库。

2.1 序列数据库

2.1.1 核酸序列数据库

GenBank、EMBL 和 DDBJ 是国际上三大主要核酸序列数据库。GenBank 由美国国家健康研究所（National Institute of Health）于 20 世纪 80 年代初委托 Los Alamos 国家实验室建立，后交给美国国立生物技术信息中心（NCBI）。EMBL 由欧洲分子生物学实验室（Europe Molecular Biology Laboratory）于 1982 年创建，目前由欧洲生物信息学研究所负责管理。该数据库由 Oracle 数据库系统管理维护，查询检索可以通过因特网上的序列提取系统（SRS）服务完成。向 EMBL 核酸序列数据库提交序列可以通过基于 Web 的 WEBIN 工具，也可以用 Sequin 软件来完成。DDBJ（DNA Data Base of Japan）创建于 1986 年，由日本国立遗传研究所负责管理。DDBJ 也是一个全面的核酸序列数据库，与 GenBank 和 EMBL 核酸库合作交换数据。可以使用其主页上提供的 SRS 工具进行数据检索和序列分析，用 Sequin 软件向该数据库提交序列。

1988 年，GenBank、EMBL 和 DDBJ 共同成立国际核酸序列联合数据中心，建立了合作关系。根据协议，三者每天相互交换数据使三个数据库的数据同步。由于 DDBJ 数据库的内容和格式与 GenBank 相同，下面主要以 GenBank 和 EMBL 数据库为例介绍核酸序列数据库的有关情况。

GenBank 数据库包含了所有已知的核酸序列和蛋白质序列，以及与它们相关的文献著作和生物学注释。它是由美国国立生物技术信息中心（NCBI）建立和维护的。它