生物信息学

手册

郝柏林
张淑誉 编著

上海科学技术出版社

生物信息学手册

郝柏林　张淑誉　编著

上海科学技术出版社

图书在版编目(CIP)数据

生物信息学手册/郝柏林,张淑誉编著.—上海:上海科学技术出版社,2000.10

ISBN 7-5323-5670-1

Ⅰ.生… Ⅱ.(1)郝…(2)张… Ⅲ.生物信息论-基本知识 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字(2000)第 47055 号

上海科学技术出版社出版发行

(上海瑞金二路 450 号 邮政编码 200020)

新华书店上海印刷股份有限公司

新华书店上海发行所经销

2000 年 10 月第 1 版 2000 年 10 月第 1 次印刷

开本 787×1092 小 1/16 印张 17.25 插页 4 字数 264 千

印数 1—3 000 定价:36.00 元

内 容 提 要

目前，从细菌到人类，各种生物数据库的信息量正迅猛增长，生物信息学作为一门崭新的前沿学科也应运而生。生物学不再是“数学等于零”的学科，也不再是仅仅基于观察和实验的科学，理论和计算将发挥日益巨大的作用，网络技术将为生物学研究插上腾飞的翅膀。

生物信息学的飞速发展，很容易使人在浩如烟海的信息中迷失方向。本书列举了近千条网址和引文，基本涵盖了生物学研究的各个方面，堪称生物信息的汪洋大海中的导航图。对生物信息学的服务、软件和算法，本书也作了比较全面的描述。

本书可供广大生命科学工作者以及由物理学、数学和计算机科学转入生命科学领域的研究教学人员参阅。

前　言

20世纪的数理科学对无生命物质的结构和运动的研究，从微观到宇观，可谓既深且远。生命物质和生命现象必定是21世纪数理科学研究的重要对象。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，生物信息学应运而生。它使生物学研究者如虎添翼。它也是数理科学工作者进入生命研究领域的自然插入点。

从细菌到人类，众多物种的基因和蛋白质数据正在以科学史上从未有过的高速度增长。目前已测定出30多种细菌，以及一些比细菌更高等的物种如酵母、线虫和果蝇的完全基因组序列。人类基因组，即一个典型的“人”的全部基因，也将提前在2001年完全测定。到2000年4月中旬，基因数据总量的增长速度达到每8个月翻一番。同时，每个月还至少测出160种蛋白质的三维结构。人本身当然是研究的核心。没有两个人的基因组完全相同。人类基因组计划的完成，只是更为细致的人群乃至个体的正常和病理基因及其表达产物的研究出发点。预计10年内，如何利用生物信息库和生物计算手段，即将成为广大临床医师和农林畜牧工作者基本训练的一部分。生物信息对未来军事和国防的影响也不容忽视。

这种情况不仅反映了科学知识的深化和研究方式的转变，在短短几年内必将影响生物、医学、农业乃至军事的众多领域。生物学不再是恩格斯所说“数学等于零”的学科，也不再是仅仅基于观察和实验的科学。理论和计算将发挥日益巨大的作用，数学、物理、计算机科学将越来越多地把生物学问题作为当然的研究课题。事实上，如果没有跨学科的发展，仅仅靠生物学工作者，不可能充分利用如此迅猛增长的海量数据。

发达国家如美国，目前也面临着生物信息研究跟不上需求，相关人才严重缺乏的局面。1999年6月初美国国家卫生署的一个专家委员会建议，迅速在大学和研究机构中建立5至20个生物计算中心，给予每个中

心可达 800 万美元的年度支持，以便从事研究和培养人才[1]。这一建议可能从 2001 年开始实施。

然而，欧美发达国家在生物信息方面早有积累。手工搜集的蛋白质结构数据库早在 20 世纪 60 年代就在美国开始建立。美国洛斯阿拉莫斯国家实验室 1979 年开始的核酸序列库 GenBank，现在由 1988 年成立的国家生物技术信息中心 (NCBI) 管理维护。欧洲分子生物学实验室的 EMBL 数据库 1982 年开始服务，随后又建立了欧洲分子生物学网 (EMBNet)。EMBL 数据库 1994 年改由当年建在英国剑桥的欧洲生物信息研究所 (EBI) 管理。日本 1984 年着手建立国家级的核酸数据库 DDBJ，1987 年正式服务。目前绝大部分核酸和蛋白质数据由美国、欧洲和日本三家产生。以上三家共同组成了 DDBJ/EMBL/GenBank 国际核酸序列数据库，每天交换数据，同步更新。其他国家如德国、法国、意大利、澳大利亚、瑞士、瑞典、丹麦、加拿大、以色列、南非等，在分享网络资源的同时，还纷纷建立自己的生物信息中心，为本国服务。

自从 1985 年 11 月应邀参加中国科学院生物科学部常务委员会关于“生物学发展战略”的扩大会议以来，我们一直在学习生物学的基本知识，为从非线性科学向理论生命科学的战略进军作准备。1993 年中国科学院理论物理研究所的局域网与国际互联网接通之后，各种生物数据库和信息网页就成为学习和研究的必需条件。近几年来目睹生物信息学成为一个活跃的新兴领域，深感所谓生物信息学其实就是信息和计算机网络时代的新生物学。我国的描述生物学根底雄厚，但生物信息学方面与国际前沿差距甚大。我国学者特别是年轻一代必须迅速赶上。因此，我们把自己这几年为入门而积累的工作笔记整理出来，供初学者参考。将来，国家级的生物医学信息中心成立和新一代专家成长之后，著书育人乃是他们的责任，这本小册子也就完成了历史任务。

有几件事应当说明：

第一，全书取材和表述颇不均匀。我们稍为知晓或记录较多的事情写得详细一些，重要而不熟悉的方面只给出一些引文和网址，当然还有众多疏漏。我们希望这本书能部分地起到参考手册的作用。实际上，全书也是以“手册体”写成。

[1] 请参看网址：http://www.nih.gov/welcome/director/060399.htm .

第二，语言和名词：这本中文书里夹杂着许多英文和少数拉丁字，这其实增加了确切性，并可免去读者费心猜测。没有公认译名的术语我们或试为命名或直用原文。有些法定译名似颇欠妥，如因特网 (Internet) 我们仍译为国际互联网或互联网。书末的索引，既可借以查找数据库或软件，也是英汉译名对照表。应当指出，像生物信息学这样在欧美国家迅猛发展的领域，目前不通晓英文就无法工作。

第三，引文和索引：全书有大量期刊论文、书籍和网址的引用。每项引用有一个通贯全书的统一编号，例如 [R-30] 就是第 8 页上 R. F. Doolittle 所编《大分子序列分析的计算机方法》一书，读者不难顺统一编号查到。因此，书末只有一个索引，不再列举文献。读者可以借助目录、索引和这些统一编号查找所需的内容。我们希望大家觉得这种组织方式是方便的。另一方面，网址的引用有些重复。这是为了减少前后翻查。

第四，数据库是一切生物信息学工作的基础。本书主要篇幅用于扼要介绍一批生物医学数据库，首先是《核酸研究》1999 年和 2000 年第 1 期和法国生物信息中心的 DBcat[R-207] 所列举的那些库。然而，也有一些它们未反映的库。另外，少数已经停止发展的库也偶尔提到，以便读者在文献中见到时，可以查明出处。

第五，学习方法：计算机、生物学和两者结合产生的生物信息学都是千头万绪、盘根错节的领域。有效的学习方法是“全局在胸、单刀直入”。这本小书力图勾画全局，并给出可援以攀登的一些线索。应当特别说明，本书不是计算机入门，不讲如何用鼠标点菜单之类的操作。

第六，针对我国学术界经济贫困的现实情况，我们着重介绍国际互联网上的免费生物信息资源，对商业性的软件只偶有提及。应当指出，知识共享是国际生物信息学界的突出特点。然而，随着生物信息容量、成本和重要性的上升，免费使用数据库的情况已经开始改变。近两年，瑞士蛋白质数据库 SWISS-PROT [R-401]、德国转录因子数据库 TRANSFAC [R-219]、美国的 RepBase [R-223] 等数据库都已对商业性用户收取费用，但对学术性用户仍继续免费。我国学者应当恪守学术道德，为发展科学而分享资源，并尽可能有所贡献，切不可以学术名义谋取经济利益。在事涉商业时，应主动与资源所有人联系并达成协议。

在计算机网络时代，书本的地位和作用也正在发生变化。一个理想

的、每天自动更新的服务性网页应当比任何书本更方便。不过，从一个网页出发，有成百上千种链接，每个链接导致新的网页和链接；即使在同一个网点内，信息组织的层次也可能很“深”，要正确发掘才能到达所需位置。这种情景很容易使人在信息的汪洋大海中迷失方向。一本篇幅有限、组织适宜的手册，可以起一点导航作用，提高工作效率。然而，国际互联网上的信息每时每刻都在更新和重组，记录在纸张上的情况在随时老化。我们奉劝读者在自己浏览器的书签 (Bookmark) 中，保持几个重要国际生物信息中心的网址，例如美国国家生物技术信息中心 NCBI [R-134]、欧洲生物信息研究所 EBI [R-131] 和北京大学生物信息中心 CBI [R-166] 的网址，经常浏览以关心最新进展。

我们曾经从许多学者的学术报告或面谈交流中受益，这里只能提到一部分：中国科学院上海生物化学研究所徐京华、美国 Oracle 公司郑强、美国南加州大学医学院朱钦士、台北阳明大学医学院杨永正、中国科学院生物物理研究所陈润生、北京大学生命科学院顾孝诚和罗静初、天津大学生命科学院张春霆、中国科学技术大学生命科学院施蕴渝、内蒙古大学物理系罗辽复、清华大学生物系孙之荣、美国国家生物技术信息中心万宏辉、中国科学院理论物理研究所郑伟谋、美国《科学》周刊中国代表郝欣等。特别是北京大学顾孝诚、胡美浩和罗静初，阅读了此书手稿，提出宝贵建议。本书由作者使用中国科学院计算数学与科学工程计算研究所张林波等编制的科技排版软件 LaTeX中文 CCT [R-77] 接口排版。理论物理研究所程希有和陈国义，以及上海科学技术出版社潘友星和叶剑在排版方面给予指导。我们向所有这些同仁致谢。当然，书中一切不确和失误之处概由我们自己负责，并恳请读者赐教。

目　录

第 1 章　什么是生物信息学

生物信息学是一个词典里还没有的英文新词 bioinformatics 的直接翻译。这是计算机和网络大发展、各种生物数据库迅猛增长形势下如何组织数据、并从数据中提取生物学新知识的一门学问。生物信息学的突飞猛进正在引发生物学研究方式的一场革命，它必将影响到 21 世纪的农林医药和人类生产与生活的许多方面。

为了说明这种变化，可以考察图 1.1 中画出的三条曲线。缓慢上升，似乎趋近饱和的曲线是 1966 年以来美国国家医学图书馆 (National Library of Medicine，简称 NLM) 所提供的在线检索服务 MEDLINE [R-599] 所收录的文章中的一大类，即“分子生物学和遗传学”论文数目的增长情况。MEDLINE 的选用范围超出医学而囊括几乎全部重要的生物学期刊。这条曲线大致反映了人类消化理解实验事实和数据，使之上升为科学知识的过程。从 20 世纪 80 年代初迅速抬头的曲线是美国核酸序列数据库 GenBank [R-212] 中核酸序列数目的增长情况。这条线清楚地表明，数据增长越来越快，传统的研究方式已经来不及迅速消化新数据，把后者及时提升为科学知识。

所幸有一条跨越以上两条曲线、由 8 个数据点构成的的第三条线，它反映出大规模集成电路单个 CPU 芯片上的三极管数目的增长速率。正是这一技术进步提供了解决问题的关键手段。目前一个典型的基因测序中心，每年可以产生 10^{14} 字节即 100 000GB 原始数据[2]。数据的产生、搜集和分析，都必须依靠计算机和网络，都必须发展数据库、算法和程序。这就是生物信息学的使命。

[2] 见 *Science* **284** (1999) 1742。

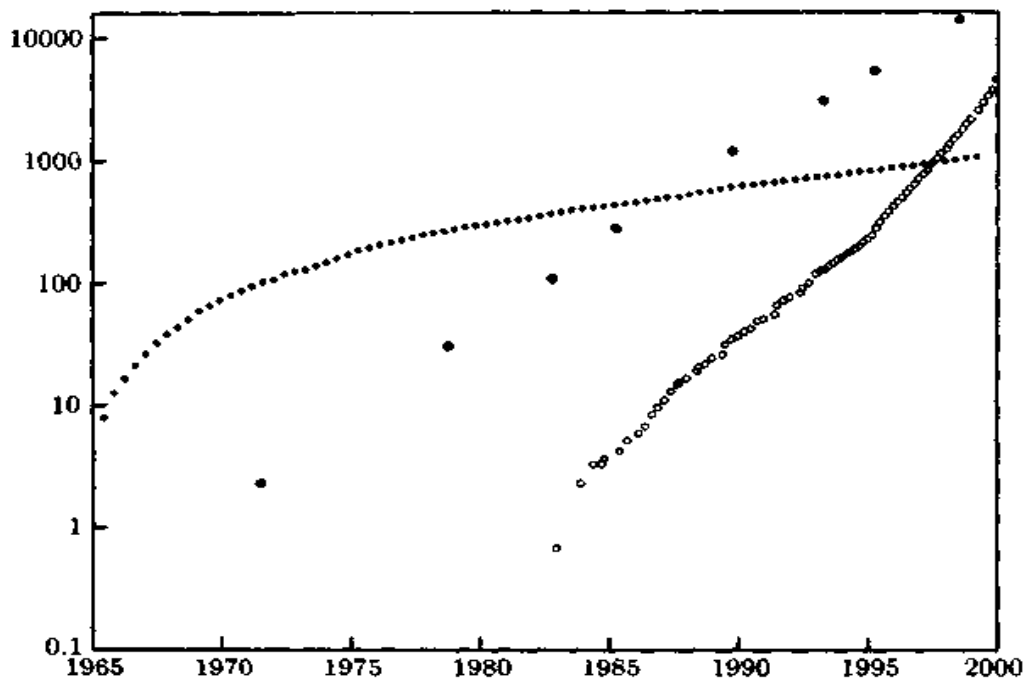

图 1.1 生物信息的增长情况

1980 年以来 GenBank 中核酸序列总碱基数目（空心圆点、高度须再乘以 10^6）、1966 年以来 MEDLINE 收录分子生物学和遗传学论文数目（实心圆点、高度须再乘以 10^3）、以及 1971 年以来 Intel 公司每个 CPU 芯片上三极管数目（大实心圆点，高度须再乘以 10^3）的增长情况。GenBank [R-212] 的数据取自每版的说明文件 gbrel.txt，关于 MEDLINE 请参看 [R-599]。Intel 的数据只有 8 个点，取自 http://www.physics.udel.edu/wwwusers/watson/scen103/ 本图构思参考了 [R-4]。

§1.1 生物数据与生物计算

20 世纪后半叶分子生物学的长足进展，把生命活动的物质基础追溯到核酸和蛋白质两大类生物大分子的序列，它们构成了生物数据的主要部分。关于这些生物大分子的结构、相互作用和生物功能的研究，也产生着大量数据。直到不久前，人类科学实践产生数据量最大的领域，还是高能物理实验和脑神经活动成象，两者都达到每年 10^{15} 字节。现在生物数据的产生率已经达到同样水平，而且很快就要超出前两者。

生物信息学与计算生物学或生物计算有密切关系，但又不尽相同。目前归入生物信息学领域的大致有以下几个方面：

第一，各种生物数据库的建立和管理。这是一切生物信息学工作的基础，通常要有计算机科学背景的专业人员与生物学者密切合作。本书不讨论数据库的建立和管理，但要列举大量现成数据库的网址，对其中一些略加说明。

第二，数据库接口和检索工具的研制。数据库的内容来自万千生物学者的日积月累，最终又为生物学所用，但不能要求一般生物学工作者具有高深的计算机和网络训练。因此，必须发展查询数据库和向库里提供数据的方便接口。这是专业人员才能胜任的工作，通常在生物信息中心里进行。本书不关心接口和检索工具的编写，但将简要介绍某些接口和工具的使用。

第三，人类基因组计划的实施，配合大规模的 DNA 自动测序，对信息的采集和处理提出了空前的要求。从各种图谱的分析、大量序列片段的联配、计算机克隆、寻找基因和预测结构与功能，到数据和研究结果的视象化，无不需要高效率的算法和程序。研究新算法，发展方便适用的程序，是生物信息学的日常任务。

第四，生物信息学最重要的任务，是从海量数据中提取新知识。这首先要从 DNA 序列中识别编码蛋白质的基因，以及调控基因表达的各种信号。其次，从基因组编码序列翻译出的蛋白质序列的数目急剧增加，根本不可能用实验方法一一确定它们的结构和功能。从已经积累的数据和知识出发，预测蛋白质的结构和功能，成为常规的研究课题。

第五，DNA 芯片和微阵列的发展，把一定组织或生物体内万千基因时空表达的研究提上日程。研究基因表达过程中的聚群关系，从中提取调控网络和代谢途径的知识，进而从整体上掌握细胞内的全部互相耦合的生化反应，这一切都要求发展新的算法。这是生物信息学刚刚掀开的新篇章。

当然，任何新兴领域的开拓，都不应从学科定义出发，事先限制自己的研究方向。生物信息学的内容也在不断扩展，而不局限于上面列举的几条。

§1.2 生物信息学与生物实验

生物信息学的发展，将造就一批不直接做实验而每天坐在计算机终端前的生命科学工作者。“生物学是实验科学”这类曾经完全正确、但已不十分符合当今科学实践的提法，如果不正确理解，就会在一定时期里挫伤有志于生物信息学的年轻人的积极性，妨碍他们获得必要的经费支持和晋升。因此，我们要专门讲一下生物信息学与生物实验的关系。

首先，作为生物信息学基础和出发点的核酸与蛋白质序列都来自实验。即使是高产出的自动测序机，也都基于以往的实验成就。同时，这也表明以往艰苦卓绝的某些实验技术已经发展成现代化生产线。

其次，在全球每天产生以千万碱基对计数的核酸序列，从中翻译出成百的可能的蛋白质序列的时代，已经根本不可能用实验办法去逐一确定它们的结构和功能。只有根据以往积累的数据和经验，对大量新序列进行分析筛选，才能突出应当由实验去决断的问题，再投入极其宝贵的人力物力。这一决策也得借助计算机完成。

第三，越来越多的物种的基因组将被基本上完全地测定。那种倾毕生精力研究一个基因、一条代谢途径、一种生理周期的时代已经过去。还会有学者这么做，但他们将只代表一种研究风格，而不再是学术主流。人们正在阐明细胞内的全部互相耦合的调控网络和代谢网络，细胞间的全部信号转导过程，从受精卵到成体的全部生理和病理的基因表达的变化，等等。这一切都超出了手工分析的可能性。

因发明了一种 DNA 快速测序方法而同 F. Sanger 分享 1980 年诺贝尔化学奖 (见本书 3.6.5 小节) 的 W. Gilbert， 1991 年在英国《自然》周刊撰写短文[3]，针对生物学的研究范式的变化指出，“正在兴起的新的范式在于，所有的‘基因’将被知晓 (在可用电子方式从数据库里读取的意义上)，今后生物学研究项目的起点将是理论的。一位科学家将从理论猜测开始，然后才转向实验去继续或检验该假设。”这一观点正在被越来越多的生物学工作者所认同。

[3]Walter Gilbert, “Towards a paradigm shift in biology” , *Nature* **349** (1991) 99.

从根本上说，实验始终起着决定性的作用．然而，这并不表明事事取决于实验，而是指那些精心设计的、决定性的新实验，否则就是忽视体现在数据库中的以往的大量实验成果．考虑到数据库中不可避免的错误和测序误差，盲目依靠数据库去对新序列进行注释，早晚会导致“辗转注释灾难” (transitive annotation catastrophe)[4] ．科学的态度当然不是因噎废食，而是发展正确的生物信息学方法，在“噪声背景”中提取信号．回顾物理学的发展，在 19 世纪曾是实验科学， 20 世纪上半叶发展成理论和实验密切结合的科学， 20 世纪下半叶成为鼎立在实验、理论和计算三足之上的成熟的发达学科．生物也是物．生物学的发展也会从物理学得到启示．

§1.3 期刊和会议

这里先列举一些与生物信息学有关的期刊和早期的会议文集，目的在于说明生物信息学的渊源．近来已经有一些经常举行的生物计算和生物信息学会议，如 PSB 太平洋生物计算研讨会 [R–825] 、ISMB 分子生物学中的智能系统会议 [R–826] 、 RECOMB 计算分子生物学年会 [R–827] 等，请访问本书第 5 章列举的网址．

据说，出生在马来西亚的美籍学者林华安 (Hwa A. Lim) 首先创造和使用了 bioinformatics 这个词，见 [R–2] ．

R–1 O. Hatase, and J. H. Wang, eds. *Bioinformatics: Information Transduction and Processing Systems from Cell to Whole Body*, 1989. 此会议文集虽用了 bioinformatics 一词，但涵义与目前用法不同．

R–2 C. R. Cantor, and H. A. Lim, eds. *Electrophoresis, Supercomputing and the Human Genome*, World Scientific Publishing Co., 1991.

R–3 H. A. Lim, J. W. Fickett, C. R. Cantor, and R. J. Robbins, eds. *The Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, World Scientific Publishing Co., 1993. 云南大学曾翻译了文集中 12 篇文章，油印成册．

R–4 *Trends Guide to Bioinformatics*, 这是 1998 年 11 月号的 *Trends in*

[4] 见 *Nature Genetics* 杂志 1999 年 10 月号的社论．

Genetics 的附刊，其中有 11 篇关于生物信息学各个方面的综述文章。

R-5 期刊 *Computer Applications in Bioscience*，简称 *CABIOS*，是生物信息和计算方面的重要期刊。这是一家不收取版面费的刊物，但它所发表的程序必须在两年内提供给学术界自由使用。从 1998 年第 14 卷起，该刊顺乎潮流、因实正名，改称 *Bioinformatics*。正式订户才能在网上阅读电子版，非订户可请求用电子邮件通知每期目录。网址：

`http://bioinformatics.oupjournals.org/subscriptions/etoc.shtml`

R-6 *Nucleic Acids Research*(《核酸研究》杂志)，每年第一期是生物数据库专集，并不限于核酸数据库。它平时也发表一些算法文章。其网页是：

`http://www.nar.oupjournals.org/`

R-7 自 1998 年开始出版的新刊 *In Silico Biology*，强调生物学研究从体内 (*in vivo*) 和体外 (*in vitro*) 的实验观察，发展到靠硅芯片 (*in silico*) 的处理和运算。这个刊物创刊号的文章可以免费下载，截至 2000 年 5 月其他新文章也还是免费的。网址是：

`http://www.bioinfo.de/isb/`

R-8 *Bioinformer* 是欧洲生物信息研究所 EBI [R-131] 的电子通信，每季度一期，自由读取。网址：

`http://bioinformer.ebi.ac.uk/`

R-9 从 2000 年开始，在 EMBNet 支持下出版新的季刊 *Briefings in Bioinformatics*(《生物信息学简报》，ISSN: 1467-5463)，由 Henry Stewart Publications 发行。出版社的网址：

`http://www.henrystewart.com/publications/bib/`

§1.4 生物信息学参考书

下面列举的书籍分成两类。第一类与生物信息学有直接关系，均系近几年的新书。

R-10 H. A. Lim, and C. R. Cantor, eds. *Bioinformatics and Genome Re-*

search, World Scientific Publishing Co., 1995.

R–11 S. K. Swindell, K. R. Miller, and G. S. A. Myers, eds. *Internet for the Molecular Biologist*, Horizon Scientific Press, 1996.

R–12 L. F. Peruski Jr., and A. Harwood Peruski, *The Internet and the New Biology. Tools for Genomic and Molecular Research*, American Soceity for Microbiology, 1997, xi + 314.

R–13 B. A. Gaëta, *ANGIS Bioinformatics Handbook*: vol. 1 Interfaces, vol. 2 Basic Bioinformatics Techniques; vol. 3 Applications; vol. 4 Specialized Databases, University of Sydney, 1997. 这是澳大利亚国家基因信息服务中心为澳生物学家准备的手册，可以全套 100 美元特价自 amazon.com 购买，但内容局限于 ANGIS 的软件环境。

R–14 夏云主编，《Internet 实用技术与生物医学应用》，军事医学科学出版社，1997, 1998，xi + 429．此书用相当大篇幅讲述网络和计算机的基本知识，包括如何用鼠标点菜单，因而直接涉及生物信息学的内容相对较少。

R–15 L. Alphey, *DNA Sequencing. From Experimental Methods to Bioinformatics*, Springer, 1997, xvi + 206.

R–16 M. J. Bishop, ed. *Guide to Human Genome Computing*, 2nd ed., Academic Press, 1993, 1998, xiv + 306.

R–17 Andreas D. Baxevanis, and B. F. Francis Ouellette, eds. *Bioinformatics. A practical Guide to the Analysis of Genes and Proteins*, Wiley-Interscience, 1998, xiv + 370. 清华大学李衍达、孙之荣的汉译本，将由清华大学出版社出版。

R–18 P. Baldi, and S. Brunak, *Bioinformatics. The Machine Learning Approach*, MIT Press, 1998, xviii + 351. 由 P. Baldi 编写的 HMMPro 程序，学术性用户可免费下载，请参看 [R–740]．

R–19 M. Bishop, ed. *Genetics Databases*, Academic Press, 1999, xiv + 295.

R–20 T. K. Attwood, and D. J. Parry-Smith, *Introduction to Bioinformatics*, AWL Press, 1999, xx + 218. 北京大学罗静初的汉译本，将由北京大学出版社出版。

R–21 Stanley I. Letovsky, ed. *Bioinformatics: Databases and Systems*, Kluwer Academic Publishers, 1999, viii + 304.

R-22 H. H. Rashidi, and L. K. Buchler, *Bioinformatics Basics Applications in Biological Science and Medicine*, 1999.

R-23 S. Misener, and S. A. Krawetz, eds. *Bioinformatics. Methods and Protocols, Methods in Molecular Biology* **132**, Humana Press, 2000, xi + 500. 这是 [R-27] 的更新版本。

第二类书籍侧重序列分析的理论和算法。这类书早有出版，例如 [R 24]。下面主要列举一些较新者：

R-24 D. Sankoff, and J. B. Kruskal, *Time Wraps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, 1983. 这是生物信息启蒙时期很受欢迎的著作。当它问世时，GenBank [R-212] 中只有 606 个 DNA 序列。

R-25 M. S. Waterman, ed. *Mathematical Methods for DNA Sequences*, CRC Press, 1989.

R-26 Russell F. Doolittle, ed. *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Method in Enzymology* **183**, Academic Press, 1990. 此书主要内容已被 [R-30] 取代。

R-27 A. M. Griffin, and H. G. Griffin, eds. *Computer Analysis of Sequence Data*, Part I and II, *Methods in Molecular Biology* **25** , Humana Press, 1994. 此书已被 [R-23] 取代。

R-28 C. A. Pickover, ed. *Visualization of Biological Information*, World Scientific Publishing Co., 1995. 这是 15 篇简短综述的文集，包括我国学者张春霆关于核酸序列的一种形象表示 (Z 曲线) 的介绍。

R-29 Michael S. Waterman, *Introduction to Computational Biology. Maps, sequences and genomes*, Chapman & Hall, 1995, xv + 431.

R-30 Russell F. Doolittle, ed. *Computer Methods for Macromolecular Sequence Analysis, Method in Enzymology* **266**, Academic Press, 1996.

R-31 M. J. Bishop, and C. J. Rawlings, eds. *DNA and Protein Sequence Analysis*, Oxford University Press, 1996.

R-32 S. Schulze-Kremer, *Molecular Bioinformatics: Algorithms and Applications*, Walter de Gruyter, 1996, xv + 300.

R-33 Dan Gusfield, *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology*, Cambridge University Press, 1997,

xviii + 534.

R-34 João Setubal, and João Meidanis, *Introduction to Computational Molecular Biology*, PWS Publishing Company, 1997, xiii + 296.

R-35 S. R. Swindell, ed. *Sequence Data Analysis Guidebook*, Humana Press, 1997.

R-36 R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Methods of Proteins and Nucleic Acids*, Cambridge University Press, 1998, xi + 356.

R-37 Ben-Hui Liu, *Statistical Genomics. Linkage, Mappings, and QTL Analysis*, CRC Press, 1998, xxix + 611.

R-38 S. L. Salzberg, D. B. Searls, and S. Kasif, eds. *Computational Methods in Molecular Biology*, Elsevier Science, 1998.

R-39 J. T. L. Wang, B. A. Shapiro, and D. Shasha, eds. *Pattern Discovery in Biomolecular Data: Tools, Techniques, and Applications*, Oxford University Press, 1999.

本书取材，许多来自以上书刊和互联网上的各种网页，一般不再处处指明. 顺便指出，国内期刊上近几年也有许多介绍生物信息学的文章。例如：

R-40 丁达夫、梁卫平、陈洁，“生物信息学”，《科学》 **50** (1998) 第 2 期，20 - 23.

R-41 罗静初、江涛、李兵、陈新、李笑难、潘卫、唐汶、顾红雅、张兴华、顾孝诚，“分子生物信息镜象系统和数据库”，《高技术通讯》 1998 年 10 月，61 - 62 。

R-42 黄弋、罗静初、顾孝诚，“生物信息资源的应用和二次开发”，《高技术通讯》 1999 年 1 月，60 - 62 。

R-43 姬新颖、郭华章、王海涛、杨为松，“Internet 与生物信息学”，《科学美国人》 (中文版) 1999 年第 1 期，71 - 72 。

R-44 李伟章、恽榴红，“生物信息学与新药研究”，《科学》 **51** (1999) 第 2 期，17 - 20 。

R-45 欧阳曙光、贺福初，“生物信息学：生物实验数据和计算技术结合的新领域”，《科学通报》 **44** (1999) 1457 - 1468 。

R-46 郝柏林，“建议尽快组建国家级生物医学信息中心”，《中国科学院院刊》 15 (2000) 133 - 134；中国科学院，《2000 科学发展报告》，科学出版社，2000，243 - 246。

R-47 郝柏林，“生物信息学”，《中国科学院院刊》 15 (2000) 260 - 264。

第 2 章　计算机和互联网

我们假定读者所使用的计算机已经直接或以电话拨号方式联接到互联网上，因而不必关心联网问题。完全没有使用过计算机的人，应当先补课再看本书。熟悉计算机和互联网的读者，可以跳过这一章，从第 3 章或第 4 章继续阅读。所谓“熟悉”主要指：

第一，对自己使用的计算机 (PC 机或工作站) 和操作系统 (如微软视窗系统或 UNIX) 有基本了解，知道主要的命令或鼠标 / 菜单操作。

第二，知道常用的文件类型和处理相应文件的软件，例如用编辑程序产生纯文本 (text) 文件，用 GhostScript 显示 PostScript 文件，用 Acrobat Reader 显示和打印 PDF 文件，等等。

第三，可以顺利收发英文电子邮件 (E-mail)，知道如何显示和保存电子邮件的附件 (attachment)。

第四，会用 Netscape Communicator 或 Microsoft Internet Explorer 这类网络“浏览器”阅读已知地址的网页，并且借助浏览器“下载”文件。不知道网址时会借助各种搜索工具查找。

第五，会用 ftp 从已经知道地址的远程计算机的公用子目录 (/pub) 中读取文件。

第六，会用 telnet 命令登录到已经开好帐号的远程计算机上去运行作业。

如果对上面某一条没有把握，可以只参阅相应的小节。还有一个 PC 机和 UNIX 机之间，以及多台 UNIX 机的互相联接问题，没有列在上面，但有一小节叙述，必要时可参考。然而，本章并不是计算机入门。感到叙述过于简略的读者应当参阅有关书籍，例如 [R-14]。其实，最有效的学习方法是上网实践，并且请教有经验的同事或学生。

§2.1 计算机和操作系统

由使用者直接操作计算机的时代早已过去。现代用户只需要、也只能够通过由厂家提供的“操作系统”去使用计算机。其实，“操作系统”也早已退到幕后。多数用户看见的只是“窗口”和“菜单”，用鼠标轻点菜单来做自己的事情。由硬件体现的计算机，加上由操作系统总管的软硬件资源，现在统称为“平台” (platform)。目前多数生物信息资源和计算软件，在同一类平台上是完全兼容的，跨越平台随遇而安的软件也越来越多。

我国常见的计算机平台主要有两大类，即运行微软公司视窗系统的“个人”计算机 (本书中简称为 PC 机) 和运行 UNIX 操作系统的工作站 (workstation)。适用于 PC 机的一种免费 UNIX 系统称为 Linux。运行 Linux 的 PC 机逐渐增多；它们的使用与工作站几乎相同。同一台 PC 机上可以安装微软视窗和 Linux 两套系统，安装和切换的办法最好去请教本单位的系统管理人员，这里不讲。Linux 系统的另一个好处是它带有大量国际上流行的、质量相当高的免费 UNIX 应用程序，省去用户在网上寻觅和安装之劳。可以免费下载 Linux 软件和有关文件的网点很多，我们只给出下面几个：

R-48 Linux 软件网址：

`http://www.linux.org/`

R-49 Linux 软件说明书网址：

`http://www.linuxdoc.org/`

`http://www.unc.edu/LDP/`

`http://metalab.unc.edu/`

R-50 汉化的 Linux 系统可参看：

`http://www.linuxchina.org/`

`http://turbolinux.com.cn/`

历史上还有两类平台，即从 VAX 计算机及其操作系统演变而来的 VMS 系统和其实最早使用视窗的苹果机和 Macintosh。它们在国外仍然占有一定市场，在生物信息文献和网上资源中常被提及。由于这类平台在

表 2.1 重要 UNIX 和 DOS 命令简表

操作	UNIX 命令	DOS 命令	补充说明
列出当前目录	ls -l	DIR	长清单
内的文件清单	ls	DIR/W	短清单
复制文件	cp	copy	
比较文件	cmp	COMP	
删除文件	rm	DEL	
建立子目录	mkdir	MKDIR 或 MD	
删除子目录	rmdir	RMDIR 或 RD	
屏幕显示文件	cat, more	TYPE	全文
寻找字符串	grep	FIND	
打印文件	lpr	COPY PRN:	打印
询问当前目录路径	pwd		
回到根目录	cd		用户根目录
		cd \	盘主目录
回到上层目录	cd ..	CD ..	
转到子目录	cd subdirname	CD subdirname	
查手册、帮助	man	HELP	

国内并不普及，我们一律略去。

微软视窗系统的菜单背后，其实是原来磁盘操作系统 (DOS) 的命令。现在的视窗系统仍然允许用户回到 DOS 方式工作。DOS 命令的设计曾受 UNIX 系统的影响。表 2.1 列举了一些重要的 DOS 和 UNIX 命令，并作简单说明。UNIX 系统区分大小写字母，命令一般用小写。DOS 系统不区分大小写，我们在表 2.1 中也两者混用。

如果想核实某条 DOS 命令的可选参数，可在 MS DOS 窗口中发这条命令并带参数 /? 。这时系统返回一段简短说明。例如，要了解磁盘格式化命令的用法，可发命令 FORMAT /? 。

UNIX 命令极其丰富，有些命令自己就是一种语言。UNIX 系统通常带有大量联机文件。如果临时忘记某条命令的使用细节，可以发 man 命令，调阅《使用手册》中关于该命令的说明。读者不妨一试命令 man man 的效果。另一个方便的 UNIX 命令是 apropos 后随一个关键字 (不限于系统命令)，系统就会从随机文件中把包含该关键字的各行都显示出来。例如，发命令 apropos graphics ，会显示出所有联机文件中含 graphics 一

字的行. 如果输出行数太多, 还须进一步缩小查询范围.

§2.2 语言和软件

基于大规模集成电路的计算机硬件, 一般用户已无用武之处. 用户面临的是名目繁多的软件上层建筑. 即使自己不动手编写程序, 也宜对语言和软件有一些粗线条的概念.

20 世纪 50 年代后期兴起的 FORTRAN 语言, 带有强烈的美国实用主义色彩. 它的有限的数据结构和种种副作用, 注定了终将被取代的前景. 然而多年大量投资的积累, 使它成为美国的一种包袱, 尾大不掉, 死而不僵. 从 FORTRAN 77 到 FORTRAN 90, 越改越像 C 语言. C 语言和 UNIX 本是孪生兄弟, 在数据结构和运行效率上都比 FORTRAN 略胜一筹. 本书作者之一曾是我国第一本 FORTRAN 语言教科书[5] 的撰写人, 早从 1987 年起就改用 C 语言编写研究工作中的各种程序.

然而, 这一切都属于过去的发展阶段, 即面向过程的程序设计 (Procedure-Oriented Programming, 简称 POP). 所谓过程, 即相对独立的函数和子程序. 一个完整的程序由大量过程组成. 数据结构都是具体的. 不阅读整个程序就无法知晓如何加工程序前面定义的数据. 各个过程的调用顺序虽能依照执行中的情况和参数变化有所变动, 但不能超出事先设计好的总框架.

作为对比, 请思考一下现在大家都很熟悉的窗口系统, 无论是 PC 上的微软视窗, 还是 UNIX 工作站上的 X 窗口, 本身都是由程序实现的. 然而, 一位用户会打开多少窗口、打开哪些窗口、按何种顺序打开和关闭, 在窗口里做些什么, 这一切都无法事先规定. 用传统的 POP 思想编写实现窗口系统的软件, 会遇到不少困难. 面向对象的程序设计 (Object-Oriented Programming, 简称 OOP) 应运而生. 概括起来, OOP 就是三件事:

第一, 数据结构抽象化: 特定的数据结构加上允许在此结构上执行的操作, 称为对象. 对象是按类分层次定义的. 下层对象继承上层的性

[5] 郝柏林, 《FORTRAN 程序设计讲义》, 《计算机参考资料》 1977 年第 1/2 期, 四机部第 15 研究所出版; 《FORTRAN 77 程序设计》, 人民邮电出版社, 1980, 1987.

质，还可添加新性质。所有对象要向一个调度程序登记。

第二，信号驱动：有一个接受来自键盘、鼠标等设备信号输入的模块，它判断信号来源和指向那个对象，并通知调度程序。

第三，调度程序：它是始终在运行的程序，无信号时空转，有信号时激活相应的对象。

从 20 世纪 90 年代初， OOP 成为程序设计的主流思想。简单情况和基本文献可以参看：

R-51 张淑誉、郝柏林，“面向对象的程序设计”，《计算物理》 9 (1992) 343 - 345 。

从具体实现看， C 语言的 struct 数据结构，加上操作 (有时也叫方法) 就成为对象。因此，从 C 语言发展出来的 C++ 就成为主要的 OOP 语言之一。 C++ 的编译程序可以处理普通的 C 语言程序。 GNUWare [R-62] 中的 g++ 是人们广泛使用的免费编译程序。

窗口系统与图形技术密切相关。几乎所有的 UNIX 工作站都使用了一套称为 X 窗口系统 (X Windows Systems) 、简称 X 系统的基本图形软件。麻省理工学院 1987 年颁布的第 11 版，即 X11 ，早就是工作站行业的工业标准。现在常见的是 X11R6.4 ，即第 11 版第 6.4 次发行。 X 系统的庞大、初级、但面面俱到的函数库 Xlib ，很难直接用来编写最终的用户程序。于是又有基于 Xlib 的 Xtoolkit 上层建筑。 Xlib 和 Xtoolkit 都属于可自由下载的公开软件。请访问以下网址：

R-52 X 系统协会 (X Consortium) 的网址 (请注意其 /R6doc/ 子目录和 Release Notes) ：

`http://www.x.org/download.htm` 和 `/resources.htm`

`ftp://ftp.x.org/`

在每台 UNIX 工作站的系统文件目录中，都可以找到 Xlib 和 Xtoolkit 所在的子目录。实际上，在 Xtoolkit 和用户所看到的窗口、菜单、按钮之间，还有一层界面。目前最常用的界面都是商业性的，如 Sun 工作站的 Openwin ，或 SGI 工作站的 Motif 界面。它们规定了各自窗口系统的感观 (look and feel) 。麻省理工学院的 X 系统软件中带有一个功能不及 Openwin 或 Motif 强大的免费界面，称为 Athens ，可在网址 [R 52] 或工

作站的系统文件中找到。互联网上还有一些“土造”的免费界面，例如模拟 Motif 功能和感观的 LessTif，可下载一试：

R-53 **LessTif** 是免费的 Motif 界面的模拟程序。网址：

http://www.lesstif.org/

http://www.hungry.com:8000/

一般用户不必了解 X 系统及其界面。但是，如果要把自己的某个在工作站上行之有效的程序，发展成供公众使用的软件系统，或是要建立整个实验室的生物信息环境，这就成为应当有所规划的事情。

OOP 概念影响了软件技术的各个方面，出现了一批前面冠以 OO 的新缩写，包括数据库技术。生物数据库的发展，最初多是把原始数据组织在约定格式的纯文本文件中，并没有引用很多数据库技术。数据量急剧增大之后，人们不得不注意把生物数据库的进一步发展纳入现代数据库技术的主流，例如采用关系数据库和符合标准查询语言 (Standard Query Language，简称 SQL) 的协议。大的生物信息中心，如 GSDB [R-214] 使用 Sybase 公司的技术，而 EBI [R-131] 则引用 Oracle 数据库技术。专门数据库也开始设计成 OO 型的，转录因子数据库 OOTFD [R-222] 就是一例。最重要的例子，当推最早为秀丽线虫 (*C. elegans*) 基因组计划发展的 ACeDB [R-851]，可以免费下载的 ACeDB 数据库，现在已应用于许多其他基因组计划，包括在 Sanger [R-299] 中心组织每条人类染色体的序列数据。我们讲了这些似乎与一般用户无关的话，是要提醒大家注意，在生物信息学工作起步甚晚的中国，首先要建立国际上各种重要数据库的镜象，当然只能照用人家现成的技术框架。然而，一旦自己动手研制新数据库，就应当从最先进处着手，采用 OOP 概念。ACeDB 是值得借鉴的良好起点。

在网络环境下跨越平台、随遇而安的软件日益增加。这里必须提到体现 OOP 概念的 Java 语言。Java 语言不依赖平台的原因，在于它先把程序编译成普适的“字节码” (bytecode)，再由各平台上的解释程序去执行。解释执行使速度下降，这曾是对 Java 语言的主要批评。然而，随着 CPU 速度的提高，速度限制将不再突出。Java 语言在保证安全的前提下，大为增加了网络上的动态交互作用。从发展看，即使自己不用它编写程序，从网上下载 Java 软件的事也会成为家常便饭。Java 软件分为

Java Applet 和 Java Application 两大类。前者在远程的服务器上运行，用户看起来像是在自己的浏览器里运行；后者要下载到本地计算机上运行，这时要求本地计算机已安装支持 Java 的软件系统。如果本地计算机上还没有相应的系统，可以免费下载：

R-54 **Java** 网址：

`ftp://java.sun.com (/pub/)`

`ftp://www.blackdown.org (/pub/Java/)`

R-55 Sun 公司免费提供一套 Java 开发工具 (Java Developer's Kit，简称 JDK)，其中包括了 Java 编译程序 javac、解释程序 java、查错程序 jdb 和 appletviewer 等许多工具。用户可从网上下载适合于自己所用平台的 JDK 版本：

`ftp://ftp.javasoft.com (/pub/)`

为编写各种应用程序和互联网之间的接口，人们常用免费的 Perl 或 Python 语言。有一些网点专门交流用于生物信息学和生物计算的相应语言的程序文本：

R-56 **BioPerl** 组织，专门交流用于生物信息学、遗传学和生命科学研究的 Perl 工具。1999 年还召开过 BioPerl99 国际会议。请参看网址：

`http://www.bioperl.org/`

R-57 顺便指出，在华盛顿大学 Lindberg 的个人网页上有一些生物信息学用的 Perl Scripts，可以下载。网址：

`http://www.id.wustl.edu/~lindberg/docs/programs/`

R-58 **BioJava** 组织。网址：

`http://www.biojava.org/`

R-59 **BioPython** 组织。网址：

`http://www.biopython.org/`

R-60 **BioXml** 组织。XML 是在网络环境下描述数据的一种标准语言，目前虽在生物信息学中用得不多，但 BioXML 组织正在探讨其发展前景。请参看网址：

`http://www.bioxml.org/`

R-61 **BioCORBA** 组织。关于 CORBA 请参看 [R-850]。网址：

`http://biocorba.org/`

此外，还可以注意有关会议 [R-828]。

这里应当特别提一下免费软件。首先，知识产权并不意味着事事收费，更重要的是尊重作者的首创和署名。应当说，知识共享是人类社会发展的主流。国际上一直有一派软件工作者主张，软件，包括源程序应当自由免费交流，特别是以 Richard Stallman 和他在 1984 年建立的 Free Software Foundation (FSF) 为代表的群体。英文 free 一词有免费和自由双重含义，汉语没有简单确切的对应。我们在书中交替使用“自由”和“免费”二词，读者最好作两者兼顾的理解。自由软件的提倡者们的主要办法是编写了许多高质量的免费程序，这些软件通称 GNU 或 GNUWare：

R-62 **GNU 或 GNUWare**，包括著名的编辑程序 Emacs、C 和 C++ 语言的编译程序 g++、C 和 C++ 的函数库、绘图软件 Gnuplot，显示 PostScript 文件的 GhostScript 和 GhostView、文件压缩和解压程序 gzip，以及 GNU/Linux 系统等。

FSF 明显起到的作用，一是促进了商业性软件的质量和服务，二是为学术界提供了一种高尚的标准。我们在欣赏和享用自由软件时，应当尊重作者的劳动，遵守 FSF 的自由软件许可协议和传播他们的主张。详细的软件目录和 FSF 的方针，可以参看：

R-63 **FSF 自由软件基金会的网页：**

`http://prep.ai.mit.edu/`

R-64 全世界有许多可以下载 GNU 软件的服务器，这里略举数例：

`ftp://ftp.uu.net (/systemsgnu/)`

`ftp://utsun.s.u-tokyo.ac.jp`

`ftp://cair-archive.kaist.ac.kr`

`ftp://ftp.cs.columbia.edu`

GNU 自由软件最初目标之一是提供一套高效、优质的替代 UNIX 的命令 (据说 GNU 的意思乃是 GNU is Not Unix)。因此，GNU 程序多数针对 UNIX 平台。

至于 PC 机和微软视窗，互联网上也有一些免费软件。例如，读者可以访问：

R-65 **tucows** 公司以及它在世界各地区包括我国的镜象点，列出了大量商业软件、共享软件和自由软件，并且有所评价，标有 5 头“牛”的质

量最好。总公司网址：

`http://www.tucows.com/`

共享软件 (shareware) 可以自由下载，使用满意后再付费，通常比商业软件便宜。许多商业软件也有免费试用期。

§2.3 互联网和浏览器

当今世界上绝大多数计算机都已联接成网。没有计算机网络，就谈不上生物信息学。

国际互联网的物质基础，当然是由各种有线 (光缆、电缆、电话线) 和无线 (微波、卫星) 通信线路联接起来的计算机资源。一旦联接成网，它就可以支撑各种各样的由软件实现的“上层建筑”。对用户说来，上层建筑中最重要的软件是网络浏览器。

2.3.1 TCP/IP 和 IP 地址

网络上有各式各样的计算机平台，局域网之间也可能有差异。为了正确交换信息，必须遵守共同的网络协议。目前使用较多的 TCP/IP 实际上是两个层次的协议：数据被分解并包装成“数据包”，TCP (Transmission Control Protocol) 控制数据包的传输，而 IP (Internet Protocol) 负责为数据包寻找传送途径。

网络上每一台计算机都有一个唯一的 IP 地址。例如，北京大学生物信息中心 [R-166] 服务器的 IP 地址是 202.112.7.9 。每个 IP 地址还有一个用字母拼写的“域名”。北大 CBI 服务器与 IP 地址等价的域名是：`cbi.pku.edu.cn` 。这个地址从右往左，表示由大到小、由整体到局域的网络区域名称或“域名”：`.cn` 是中国，`edu.cn` 是中国教育网，`pku.edu.cn` 是中国教育网的北京大学局域网，`cbi.pku.edu.cn` 是北京大学生物信息中心。网络和 IP 地址都是分层管理的。即使是一台 PC 机首次联网，也要自动获取或由局域网管理员分配一个新的 IP 地址。每个网络层次都有“域名服务器” (Domain Name Server，简称 DNS)。当用户要求使用某个 IP 地址交换信息时，DNS 会逐级往上查找，直至找到或返回查找不到域名的通知。网络上重要的服务器通常保持 IP 地址的缩写名字不变，而

实际使用的计算机可能因更新换代或重新组织而变换由数字构成的“绝对” IP 地址。因此，本书中尽可能使用与 IP 地址对应的缩写域名。

在互联网上按用户名和 IP 地址传送信息的最简单办法是收发电子邮件 (E-mail)。用 telnet 远程登录到其他计算机去运行作业或用 ftp 同远处的计算机交换文件，也都要用到 IP 地址。这些在后面还有专节介绍。这几种手段的局限性在于用户必须事先知道对方的 IP 地址。

2.3.2 gopher 服务器

历史上第一个协助用户在互联网的汪洋大海中搜寻所需信息的工具是 gopher 服务器。所有的 gopher 网点，逻辑上联接成树状结构。用户可在 gopher 协助下沿树枝树干搜索所有 gopher 服务器，查找所需的信息，而不必关心信息所在的实际地址。虽然 gopher 很快就被功能更为强大的 WWW 及其浏览器超过，但并未被完全代替。对于受硬件图形功能限制，只能依靠纯文本文件的用户，gopher 仍不失为一种方便的网络界面。因此，多数网点保留了原有的 gopher 服务器。我们不再介绍 gopher，但给出可以免费下载 gopher 软件的网址：

R-66 **gopher** 软件网址：

`ftp://boombox.micro.umn.edu (/pub/gopher)`

2.3.3 WWW 和 HTML

WWW 是 World Wide Web 的缩写，有时也写作 www 或简称 Web(万维网)。与 gopher 的树状结构不同，WWW 的每个结点在逻辑上都与任何其他结点保持联系，“透明”地交换信息。这就从信息组织和显示两方面提出新的要求。首先，从术语讲，凡是与跨越网络“透明”交换有关的文件、链接等均冠以“超” (hyper) 字头，如超文本 (hypertext)、超链接 (hyperlink)、超文本标注语言 (HyperText Markup Language，简称 HTML 或 html)、超文本传输协议 (HyperText Transfer Protocol，缩写为 HTTP 或 http) 等。任何一个结点上准备提供给 WWW 上其他用户共享的信息，必须用 HTML 语言加以标注。其所以叫标注 (markup)，是因为在最简单情况下，只须把纯文本文件头尾和其中段落前后加一些标签，它就成为超文本了。

由于生物信息工作每天要同网页打交道，最好知道一点 HTML 的基本概念。作为最简单的超文本文件实例，我们为自己制作一个朴实无华的网页。用任何熟悉的编辑程序，输入如下的文件：

```
<HTML>
<H1>Welcome to Bai-lin HAO's Homepage!</H1>
<BODY>
<P>Brief <A HREF="vitae.html">vitae</A>.</P>
<P>Fields and recent reserach <A HREF="interest.html">
            interests</A>.
<P>Selected title of recent <A HREF="shortlst.html">
            publications</A>.
<P>My favorite biolink is <A HREF="www.cbi.pku.edu.cn">
            CBI</A> at Peking University.
</BODY>
</HTML>
```

给这个文件起名字 index.html (UNIX 系统) 或 index.htm (PC 系统)，并且把它放到本单位网络管理员规定的公开子目录中。例如，在 UNIX 工作站上，这可能是名为 /public_html 的子目录。以上简短文件清楚说明了 HTML 语言的风格。它是用一批成对的“标签”组织起来的。标签的种类很多。表 2.2 给出几对常见的标签。

表 2.2 最常见的几对 HTML 标签

标签	说明
⟨HTML⟩ ··· ···⟨/HTML⟩	中间是超文本文件
⟨BODY⟩ ·· ···⟨/BODY⟩	中间是文件主体
⟨P⟩ ··· ···⟨/P⟩	中间是一节，⟨/P⟩ 可省略
⟨A ...⟩ ··· ···⟨/A⟩	中间是一个超链接
⟨H1⟩ ··· ···⟨/H1⟩	用 1 号大字作标题
⟨i⟩ ··· ···⟨/i⟩	用意大利体即斜体
⟨b⟩ ··· ···⟨/b⟩	用黑体

最重要并且应当特别说明的，是形成超链接的标签

⟨A HREF="www.cbi.pku.edu.cn"⟩CBI⟨/A⟩

其中只有 CBI 会被浏览器用特殊的方式显示出来，或用蓝色或下面划线。至于 ⟨P⟩ 后面、⟨A ...⟩ ··· ··· ⟨/A⟩ 外面的文字，浏览器照原样显示。

用鼠标点击 CBI，浏览器就自动去访问“超引用” HREF= 指出的 IP 地址，即北京大学生物信息中心的网页。超链接也可以指向本地文件。前面例子里，简历 vitae.html 、领域 interest.html 和近作 shortlst.html 三个超文本文件都同 index.html 在一个子目录里，在点击 vitae 、 interests 和 publications 三个字之一时，被分别访问。

我们看到，HTML 语言并不用来写任何文件主体。不管是纯文本文件，还是图形、动画、声音、电影等“多媒体”文件，只要恰当地加上标签，就成为有声有色、可以被浏览器跨越网络访问的超文本文件。这个例子只是说明，HTML 语言入门并不难，深造也是办得到的。这里只点一本参考书：

R-67 Chuck Musciano, and Bill Kennedy, *HTML. The Definitive Guide*, 2nd ed., O'Reily, 1997.

实际上，专门学习 HTML 语言的必要性不大。目前有许多为一般人用的工具，例如，微软公司的 FrontPage 98 和 MacroMedia 公司的 Dream Weaver，可以帮助人们制作网页和管理网站。

2.3.4 浏览器和 URL

互联网与用户之间最重要的接口软件是浏览器 (browser)。 WWW 上的超文本文件由浏览器访问和显示。目前最常用的浏览器有两种：

R-68 **Netscape Navigator** 和 **Netscape Communicator**，有适用于各种计算机平台的版本，可以免费下载。网址是：

`http://www.netscape.com/`

用户应注意经常更新软件版本，每一新版都随带提供许多可选择下载的应用程序。本书作者比较喜欢使用 Netscape 。我们现在使用 1999 年 10 月发行的 Netscape Communicator 4.7 版本。

R-69 **Internet Explorer**，这是微软公司为其各种视窗系统提供的浏览器，因此没有工作站版本。它通常随视窗系统提供，也可以从微软的网址：

`http://www.microsoft.com/`

免费下载较新的版本。本书作者使用的是 Internet Explorer 5.0 版。

现在的浏览器自动与许多应用软件密切合作。例如，用户下载一个 PDF [R-81] 文件时，浏览器会自动调出 Acrobat Reader 来显示它；如果计算机上没有 Acrobat Reader，浏览器会协助用户从 Adobe 公司的网页下载。这样下载的 Acrobat Reader 留在用户的计算机里，成为可以单独使用的软件。因此，对许多用户而言，浏览器就是其所见所用的整个计算机。现在形形色色的网页往往有大量彩色图形，其中不少是商业广告。线路传输条件不好时，占用极长时间。有些网页允许用户选择图形量较少的方案。如果图形不重要，还可以选用只显示字符信息的浏览器，例如：

R-70 **Lynx**。这个字符信息浏览器，可从网上免费下载：

`http://lynx.browser.org/`

`ftp://ftp2.cc.ukans.edu (/pub/lynx/)`

浏览器使用统一资源定位符 URL(Universal Resource Locator)，来指定按何种信息交换协议、向哪一个网址发送、读取或交换信息。目前常用的 URL 有以下 6 类：

R-71 **http**：按 HTTP 超文本传输协议交换信息，首先是从指定网址读取对方公开的用 HTML 标注的网页。例如，读者可试通北京大学生物信息中心的网页，其 URL 是：

`http://www.cbi.pku.edu.cn/`

由于这是最常用的工作方式，即使省略 http 或 www，许多浏览器都能正确处理。

R-72 **ftp**：按 ftp 文件传输协议从指定网址的 ftp 服务器读取其公开目录中的文件。例如，从北京大学生物信息中心的 ftp 服务器下载文件用：

`ftp://ftp.cbi.pku.edu.cn`

详见 2.7.2 小节关于 ftp 的描述。

R-73 **gopher**：按 2.3.2 小节中介绍的 gopher 协议读取信息。例如：

`gopher://gopher.ebi.ac.uk/`

R-74 **mailto**：按电子邮件协议，不退出浏览器即可收发电子邮件，然后再继续浏览。例如，用：

`mailto:somebody@someuniversity.edu.cn`

给某校某人发电子邮件。注意 mailto: 之后不写 //。

R-75 **news**：按用户网协议 (Usenet Protocol) 阅读一定新闻组的网页。例如，用：

```
news:bionet.software.www
```

访问关于生物软件的网上新闻组 (请参看 [R-807])。注意 news: 之后也不写 //。

R-76 **telnet**：远程登录到指定 IP 地址的计算机上去执行作业。当然，必须事先在远程计算机上开好用户帐号并知道口令。例如，

```
telnet:mycomputer.myuniversity.edu.cn
```

请参看后面第 2.7.1 小节。

还有不少商业性的网络浏览器，例如“美国在线” (American OnLine，简称 AOL) 浏览器和网上的某些性能很好的免费浏览器。但是，多数生物信息网页的设计并未考虑这类浏览器的特点。我们建议读者使用标准的浏览器，如 Netscape [R-68] 或 Internet Explorer [R-69]，好在它们都是可以免费下载的软件。

2.3.5 文件的下载和上载

在网页上标明的可以下载 (download) 的文件，用鼠标左键点击就被调到窗口中供阅读。如果用鼠标左键点击同时按下“上挡” (shift) 键，浏览器就准备把该文件存到盘上。这时会出现关于子目录、文件格式等的对话窗口。用户正确回答后，还会开小窗口显示下载进程。

使用生物信息网页时，最常见的一种“上载” (upload)，是由用户提供一条核酸或蛋白质序列去做数据库查询或联配。这时常常用两种方法之一：

第一，使用视窗系统的剪 (cut)、抄 (copy) 和贴 (paste) 的功能，把序列从用户的一个窗口中涂黑后剪 (抄) 下来，实际上是送入视窗系统的缓冲存储器，再贴到网页中已显示出的输入窗口。对于不太长的序列，这样做很方便。

第二，网页中显示一个上载窗口，用户可用浏览目录的方法找到需上载的序列文件，把文件名字填写进去，再按提交 (submit) 钮即可。用上载法可以提交较长的序列。

表 2.3 一些网上“搜索器”的地址

名称	地址	特色
Yahoo	www.yahoo.com	可按分类搜索
AltaVista	altavista.com	还允许访问新闻组
Lycos	lycos.cs.cmu.edu	可按关键字搜索
Infoseek	www.infoseek.com	有网址点击百分比
Excite	www.excite.com	可按概念搜索
About	About.com	可解释名词或概念

2.3.6 网上“搜索器”

互联网上有五花八门的海量信息，查找真正有用的网点并非易事。现在网上有各式各样的“搜索器” (searching engine)，免费帮助用户搜寻所需信息。表 2.3 开列了一些搜索器的网址。通常只要按浏览器的“搜索” (Search) 按钮，就会显示出一批搜索器的名字供选用。

使用这些搜索器时，应注意几件事。

第一，要迅速缩小主题范围，从 Science 到 Biology 到 Molecular Biology 到 RNA，不可泛泛查找。

第二，要用专业而非一般的关键字。有些搜索器允许若干关键字的逻辑组合，要恰当利用才不会适得其反。有时可加上引号，如“ tandem repeats ”表示只找两字的此种固定组合，排除单个出现的情形。

第三，搜索结果中往往有许多重复出现的网址。例如，用 Yahoo 查找 bioinformatics，返回的一百多个网点中近一半重复。

第四，这些搜索器公司，主要靠广告费支持。因此，它们的网页上有各种五彩缤纷的诱人广告，用户要慎重自持，切勿落入“陷阱”，浪费光阴。

§2.4 常见的文件类型

信息通常以文件形式保存和传输。文件所保存的不只是可读的文字信息，图形、相片、动画、音乐、影片等等，都可成为文件。许多文件类型可由通用的文件名后缀识别。例如，最简单的“纯文本” (plain text)

文件，后缀是 .txt 。它不含任何字体、字号之类的格式信息。文字处理软件，如微软的 Word ，产生带种种格式的文件，通常后缀是 .doc 。然而，在 UNIX 系统中带后缀 .doc 的往往是各种说明书，通常只是纯文本文件。

特定的软件系统接受和产生一定类型的、带特定后缀的文件。例如，常用的排版软件 TEX和 LATEX要求带 .tex 后缀的纯文本输入，产生后缀为 .dvi 的“与设备无关”的输出文件，还会产生 .aux 和 .log 类型的辅助文件。顺便指出，本书是由作者们用中文 LATEX系统自己排版，按手稿印刷的。所用软件见：

R-77 郭力、张林波等，《CCT 中外文科技激光照排系统》，海洋出版社，1993 。网址：

ftp://ftp.cc.ac.cn (/pub/cct/msdos/)

又如 PostScript 最初是 Adobe 公司设计的一种描述由文字和黑白或彩色图形组成的页面的语言，现在几乎已经成为一切页面输出设备如打印机的工业标准。这种文件通常由特定的软件建立，例如许多绘图软件或生物计算的 GCG [R-792] 软件包，允许用户把输出“设备”选取为 PostScript ，相应文件后缀为 .ps 。还有一种包装起来的 (encapsuled) PostScript 文件，可以作为整体放大、缩小、旋转、变形或插入其他文件，其后缀为 .eps 。

其实，PostScript 本身是一种程序设计语言，花一点功夫就可以掌握。最主要的参考书是：

R-78 Adobe Systems Inc. *PostScript Language. Tutorial and Cookbook*, Addison-Wesley.

R-79 Adobe Systems Inc. *PostScript Language Reference Manual*, 2nd Ed. Addison-Wesley.

.ps 和 .eps 都是由 ASCII 字符组成的纯文本文件，知道 PostScript 语言的人不难读懂。然而，对于一般用户，只有打印或显示出来才能看到效果。最方便的显示程序是名为 GhostScript 的免费软件：

R-80 **GhostScript** 是与 PostScript 等价的解释语言。**GSView** 是它的显示程序，也具有打印功能，可从以下网址获取：

http://www.cs.wisc.edu/~ghost/

ftp://ftp.cs.wisc.edu (/ghost/rjl/gsview*.zip)

Adobe 公司的另一个贡献是所谓 PDF 文件:

R-81 **PDF** 即可移植文件格式 (Portable Document Format)，常用后缀为 .pdf，产生 PDF 文件的工具，如 Acrobat PDFWriter 或把 PostScript 文件转变成 PDF 文件的 Acrobat Distiller 都是商业软件，但阅读 PDF 文件的工具 Acrobat Reader 可免费由 Adobe 公司的网页下载:
http://www.adobe.com/

我们目前使用的是 Acrobat Reader 4.0。初次下载 PDF 文件时，如果浏览器发现本地计算机上还没有安装 Acrobat Reader，它会协助用户下载.

在网上会遇见与图象信息有关的一些文件格式，如 .png 、 .gif 和 .jpeg，这里只简单提一下。 JPEG 来自 Joint Photographic Experts Group。相片扫描进计算机后，通常以 .jpeg 格式保存，它要对信息压缩并导致一些失真。 GIF (Graphics Interchange Format) 是最简单的图象格式，只有 256 种颜色。 GIF 软件涉及商业产权，因此又发展出意在取代 GIF 的自由的 PNG (Portable Netwrok Graphics) 格式。关于后者，可以参看:

R-82 http://www.w3.org/TR/REC-png-multi.html

还有一批与超文本文件 HTML 有关的格式，读者通常不需要去产生它们。知道一些名字，可减少生疏感。有些与 HTML 接口的应用程序，例如在网页上填表提交序列和参数，用到 CGI(Common Gateway Interface) 机制。相应的命令文件可以用任何语言编写，最常用的是 Perl 语言。这类文件通常带后缀 .cgi 或 .pl。有些动态产生网页的 HTML 文件，从服务器本身调用某些插入文件 (Server Side Includes，简称 SSI)，它们的后缀是 .shtml (UNIX) 或 .stm (PC)。

§2.5 文件的压缩和解压

在传输或保存数据时，为了减少数据量，大的文件通常以压缩形式保存。特别对于图形文件，压缩尤其必要。在 UNIX 系统上，标准的压缩命令是 compress myfile，压缩后的文件自动加上后缀 .Z。解压命令是

uncompress myfile.Z .

请注意, UNIX 中的为多个文件建立档案的 tar c 命令没有压缩功能。由 tar 得到的档案带后缀 .tar , 通常比原来的几个文件长一点, 压缩后成为 .tar.Z 文件。使用前要先 uncompress , 再用 tar x 把文件从档案中取出来。详情请看 man tar 和 man compress .

PC 机上的微软视窗系统没有标准的压缩和解压命令, 但有一批广泛使用的压缩程序, 如 pkzip.exe(压缩后缀为 .ZIP) 和 pkunzip , pkpak.exe 和 pkunpak.exe(压缩后缀为 .arc) , 以及 arj.exe(压缩后缀为 .arj) 等. 网上有一些针对 PC 视窗系统的自由或带免费试用期的共享压缩、解压软件, 如 FreeZip 、 WinZip 等, 见:

R-83 http://www.ozemail.com.au/~nulifetv/freezip
R-84 http://www.winzip.com/

还可查阅 [R-65] 等网址.

在 UNIX 系统和 PC 视窗之间双向兼容的压缩软件是 GNUWare [R-62] 中的 gzip . 压缩命令是 gzip filename , 压缩后的文件名带后缀 .gz . 命令中可以使用 * 来分别压缩一批文件. 解压命令是 gzip -d filename.gz 或 gunzip filename.gz , 命令中也可以使用 * 来解压一批文件.

这里顺便讲一下 uuencode 和 uudecode 命令. 早年在 UNIX 系统之间用二进制传送文件时, 为了避免连续空格和特殊代码被错误处理, 事先要把文件用 uuencode 命令编码为 "可读" 的 ASCII 文本, 收到以后再用 uudecode 命令解码复原. 直到现在, 在 EBI [R-131] 的 ftp 服务器的 /pub/software/ 子目录中 (见 [R-612]), 仍有一些用 uuencode 加工过的软件. 特别是, 一个大文件往往被分成几个文件, 后缀为 .uua 、 .uub 等, 要由用户自己解码和拼接.

§2.6 电 子 邮 件

虽然现在多数生物信息服务都可以通过 WWW 网页享用, 电子邮件仍然是重要的提交询问、获取信息和搜索数据库的手段. 特别当所提作业要求较长计算时间或返回信息量较大时, 电子邮件更是不可替代的办

表 2.4 一些生物信息电子邮件服务的地址

电子邮件地址	简短说明
blast@ncbi.nlm.nih.gov	搜索序列数据库 [R-631]
fasta@ebi.ac.uk	搜索序列数据库 [R-641]
blitz@ebi.ac.uk	搜索蛋白质数据库 [R-651]
blocks@howard.fhcrc.org	蛋白质分类和同源性 [R-476]
Q@ornl.gov	GenQuest 多方搜索 [R-652]
query@ncbi.nlm.nih.gov	NCBI 的 Entrez 集成检索服务 [R-200]
retrieve@ncbi.nlm.nih.gov	从单个 NCBI 数据库索取条目 [R-201]
grail@ornl.gov	基因序列中预测外显子 [R-719]
nnpredict@celeste.ucsf.edu	预测二级结构的神经网络 [R-766]
phd@dodo.cpmc.columbia.edu	预测蛋白质二级结构 [R-760]
repeatmasker@ ftp.genome.washington.edu	掩藏序列中的重复和平庸片段 [R-748]
signalp@cbs.dtu.dk	预测蛋白序列中的信号多肽 [R-767]

法. 表 2.4 中开列了一批提供生物信息服务的电子邮件地址, 并附简短说明. 较为详细的用法, 请根据说明中的引用号, 查阅有关段落. 多个生物信息中心可能提供同一项服务, 表中只给出一处地址, 也请看后文. 通常只要按表中地址发一封电子邮件, 正文中只写一个字: HELP, 就可以获得详细的使用说明.

R-85 从瑞士生物信息研究所 [R-141] 的 ftp 服务器可以获取更加详尽的电子邮件服务器地址清单, 但其数据较旧, 有些地址已不复存在. 网址:

ftp://www.expasy.ch (/databases/info/serv_ema.txt)

R-86 另一个重要资料来源是印第安那大学的生物信息档案 [R-611]:

http://iubio.bio.indiana.edu/

使用电子邮件服务时, 请特别注意以下两点:

第一, 必须严格按照规定格式提交作业, 因为这些无人干预的自动服务, 不会正确处理违规来函.

第二, 网上服务不是一种当然权利, 而是享用国际同行好心提供的机会, 而且总有某个单位为所用资源付出费用. 因此, 首先不可滥用, 务必节省对方机器时间; 其次, 在论文中要明确引用服务的来源; 再次, 当

有可能时，在自己的网页上提供有益的服务，以回报国际科学界。

上面第二条，也适用于其他一切形式的网上服务。

§2.7 远程计算机

互联网提供了使用远程计算机资源的可能性，允许与许多远程计算机交换文件。所谓远程计算机可能在同一间办公室内，也可能远隔重洋。

2.7.1 telnet — 登录到远程计算机

经互联网登录到另一台计算机上去运行作业，要求事先在那台机器上取得用户名和口令。登录命令是：

`telnet` 远程计算机的 IP 地址或域名

实现联接之后，按对方要求完成登录手续。

2.7.2 ftp — 远程文件传送

所有大型信息中心和许多单位均设有遵从文件传输协议 (file transfer protocol) 的服务器即 ftp server。人们只要知道 IP 地址，就可以使用 ftp 命令以无记名 (anonymous) 方式访问公用目录区，读取文件或下载软件。虽然可以在浏览器里实现 ftp 传输，单独使用 ftp 命令有时仍有好处：有些服务器可以对文件或子目录作实时压缩、只须在命令中把文件名加上 `.Z`、`.gz` 等后缀 (如服务器上文件并未压缩)；整个过程在用户监控之下，效率可能略高，而且线路中断时可设法补救。

表 2.5 中开列了一些内容较为丰富的 ftp 服务器的 URL。更详尽的 ftp 服务器清单，可从瑞士生物信息研究所 [R-141] 的 ftp 服务器下载：

R-87 瑞士生物信息研究所生物信息 ftp 服务器清单：

`ftp://www.expasy.ch (/databases/info/serv_ftp.txt)`

这个清单中许多地址已不复存在。请参考印第安那大学的生物信息档案 [R-611]。此外，现在的网络浏览器中，几乎对每一个 http 地址都可以试用 ftp 命令。只要存在相应的 ftp 服务器，就可以联接上并读取文件。

看一个实例。用美国 NCBI 的 ftp 服务器由 GenBank 读取大肠杆菌全基因组的步骤如下：

表 2.5 一些重要的生物信息 ftp 服务器的 URL

ftp 服务器的 URL	简短说明
ftp.ebi.ac.uk/pub/software	欧洲生物信息研究所 [R-131]
ncbi.nlm.nih.gov	美国国家生物信息中心 [R-134]
ftp.nig.ac.jp/pub/	日本国立遗传所 [R-137]
sanger.ac.uk/pub/	英国 Sanger 中心 [R-299]
ftp.infobiogen.fr/pub/	法国生物信息中心 [R-148]
ftp.expasy.ch/pub/	瑞士蛋白质组专家系统 [R-141]
ftp.embl-heidelberg.de/pub/	欧洲分子生物实验室 [R-133]
iubio.bio.indiana.edu	印第安那大学 [R-161]
ftp.cbi.pku.edu.cn	北京大学生物信息中心 [R-166]

```
ftp  ncbi.nlm.nih.gov
login: anonymous (或 ftp)
password: name@computer.domain (以自己的电子邮件地址作口令)
Guest logged in. Restrictions apply.
bin
cd genbank/genomes/bacteria/Ecoli
get ecoli.tar.Z
quit (下载完毕之后退出)
```

在上面的对话记录中，普通字体是用户在本地计算机上输入的命令，黑体字是远程计算机的反应。此处省去了从远程服务器返回来的一些文字信息。在 ftp 过程中可以使用一批类似 UNIX 的命令，如表 2.6 所示。本书中为了节省篇幅，ftp 联接成功后需用 cd 转入的子目录路径放在括号中，用空格隔开，置于 ftp 地址之后，前面 [R-77] 中已经使用过这种记法。

有些命令普通用户无权在远程服务器上使用，例如删除文件 (rm)、建立子目录 (mkdir) 或删除子目录 (rmdir)，未在表中列出。远距离传输文件时，一定要置二进制 (bin)，才不会因为不同平台文件系统的差异而出错。用 mget 命令时，文件名中可含 * 而指定多个特定名称的文件。例如，取一批以 .ps 作后缀的 PostScript 文件，发命令 mget *.ps。远程服务器在给出每个具体文件名时，都要询问是否读取。用户回答 no 即跳过，回答 yes 才取回。使用 prompt 命令，可以取消此种对话操作，连

表 2.6 常用的 ftp 命令

命令	解释
bin	用二进制传输
asc	用 ASCII 码传输
prompt	取消对话操作
cd	在远程服务器上改变子目录
lcd	在本地计算机上改变子目录
pwd	显示远程服务器当前子目录名
ls	显示远程子目录中文件名
get	点名取一个文件
mget	取一批文件
help	取得简短帮助信息
quit	退出 ftp

续读取。公共文件通常在服务器的 /pub/ 子目录中 (上面从 GenBank 取大肠杆菌基因组恰巧是个例外)。有的服务器设有 /incoming/ 子目录，供用户投稿或提交数据，这时可用 put 或 mput 命令。

从远程服务器用 ftp 读取大文件往往需很长时间，线路不佳时还会中断。除了选择周末深夜，网络较为空闲的时间，还可以安装一个免费的下载管理软件。例如，可从

http://www.gozilla.com/ 或

http://www.tucows.com/

取来名为 go!zilla 的程序，它可以自动重新联接到服务器，从上次中断处继续读取文件，还可在读完文件后自动挂断电话线甚至关闭计算机。

§2.8 多种平台共存的工作环境

一个工作单位内多种平台共存，目前已是普遍情形。如何保证硬件联接和相容，是系统管理员的责任，这里不讲。下面只提几件经常遇到的事情。

第一，不同平台间文件传输：从 PC 机或 UNIX 机向另一台 UNIX 工作站进行文件传输和远程登录，要求在工作站上开有帐号和知道口令 (用无记名 ftp 访问工作站的公开目录不受此限，已在前面讲过)。从 UNIX

或 PC 机向另一台 PC 机进行文件传输和远程登录，则要求后者被置成服务器模式 (通常要运行 Windows NT 或 Windows 2000 系统)。事实上，互联网上有许多 PC 机服务器。不过，一般个人用户并无必要把自己置成服务器，让别人访问。

第二，DOS 和 UNIX 文件格式转换：由于 DOS 和 UNIX 系统处理文本文件的“回车”、“换行”方式不同，从 DOS 用二进制 (bin)ftp 文件到 UNIX 之后，每行末尾会出现 ^*M* 符号。许多软件知道如何对待这些符号，但有些程序则会出问题。解决办法有三：改用 ASCII 模式 (asc) 传输；使用 dos2unix 或 unix2dos 程序进行转换；最后，靠编辑程序修正。我们借第三种办法，简单介绍 UNIX 的屏幕编辑程序。

任何 UNIX 系统都有标准的屏幕编辑命令 vi，它是由显示器尚不普及时的打字机编辑程序 ed 演变而来的。读者可发 man vi 或 man ed 命令，了解它们的详细用法。为了在 UNIX 下删去每行末尾的 ^*M*，只须调用 vi filename，进入编辑程序后输入冒号“:”。冒号“:”出现在屏幕下方后继续输入命令：

```
:1,$s/VM//
```

注意：这里 VM 不是两个大写字母，而是在同时按下“上挡”(shift) 和“控制”(ctrl) 两键时再按 V 和 M，其效果是 ^*M*。上面这个编辑命令说：从第 1 行到最末行 ($) 把 /^*M*/ 置换 (s=substitute) 成 // 之间的空无一物，即取消。

第三，用本地机器为远程 UNIX 平台作 X 终端：人们常常在自己的 UNIX 工作站上开窗口，远程登录到另一台速度较快的 UNIX 工作站去运行作业，包括运行图形显示程序，却想在自己的工作站上观看图形。设本地计算机的域名地址为 station1.myuniv.edu.cn，而远程计算机的域名地址为 station2.myuniv.edu.cn。这时应在 station2 上置环境变量：

```
setenv DISPLAY station1.myuniv.edu.cn:0.0
```

要求把 station1 的主显示器，即 0.0 显示器设置为当前显示器。同时，要在 station1 上发命令：

```
xhost station2.myuniv.edu.cn
```

说明 station2 是运行 X 系统的主机。偶尔会遇到的问题，是 station2 抱怨“颜色不够分配”而拒绝显示。这有时是因为在 station1 上开了太

多使用颜色的窗口，例如 Netscape 。把它们关闭之后，应能正常显示。

从 PC 机远程登录到 UNIX 工作站，进行与上面类似的作业，实际就是把 PC 机作为工作站的 X 终端。这时要在 PC 上运行专门的 X 窗口模拟软件，例如：

R-88 **X-WinPro**，这是 Labtam Finland 公司发展的、可以多次免费试用的共享软件。下载网址：

`http://www.labf.com/`

第 3 章　生物学引论

这一章将极为扼要地介绍现代生物学、主要是分子生物学的基本知识，为后面讲述数据库和算法准备一些背景概念。生物学者们可以跳过这一章，从第 4 章继续阅读。

§3.1　地球上的自然史

地球上的生物是自然界在特定条件下演化的结果，而且仍继续处于变化之中。对于从物理科学和数学转而关心生物的学者，这更是要始终牢记的事实。简单回顾一下地球上的自然史是颇有教益的。

人类目前观察所及的宇宙大约产生于 120 亿年前发生的一次“大爆炸”[6]。大约 49 亿年前诞生了太阳系。迄今为止，在浩瀚宇宙之中，我们只知道这个小小地球在有限的历史时期里产生了奇妙的生命现象。36 ～ 38 亿年以前在地球海洋里出现了似藻类的原始生物。靠光合作用维持生活的蓝藻类的繁殖，使大气中氧的含量逐渐增加。臭氧层虽开始形成，但是还不足以保护生命免于宇宙线的辐射杀伤。因此，最初的生命活动只能存在于海洋里。大约到 7 亿年前，已经演化出各种多细胞生物，包括许多无脊椎动物。

大约 5.3 亿年前，有过一次“寒武纪大爆发”：在约 1500 万年的短短期间里，海洋里突然出现了极其众多的物种。最早的实例是在加拿大发现的 Burgess 动物群。1984 年以后，在我国云南澄江县境内发现了时间略早而蕴藏更丰富的澄江动物群[7]。近年在贵州省瓮安等地更发现了软组织依稀可见的化石[8]。

---

[6] 可参看 *Science* **279**(1998) 981 。

[7] 可参看陈均远等著，《澄江动物群，寒武纪大爆发的见证》，台中自然博物馆，1996 。

[8] S. Xiao 等，*Nature***391**(1998) 553; C. W. Li 等，*Science* **279**(1998) 879.

大约 4.3 亿年前发生了“志留纪大爆发”。那时大气中的氧已达到现代含量的 10%，地球上形成的臭氧层开始发挥对生命的保护作用。因此，这次物种大爆发的特点，就是生命活动从海洋扩展到陆地。物种的大爆发和消亡，在漫长的地质史中至少有五次记录。但寒武纪以来的主要趋势是物种减少。

恐龙的大量灭亡大致是 6500 万年前的事，虽然 4500 万年前的恐龙化石也曾有所发现。鸟类的起源大概与恐龙消灭同时，因而有一种观点是鸟类源于翼龙。古脊椎动物中猩猩科 (Pongidae) 与人科 (Hominidae) 的分离不过是 350 万年前的事。 50 万年前生活在今日周口店地区的“北京人”属于直立人 (*Homo erectus*)。 75 000~35 000 年前广泛生活在西欧和中亚一带的尼安德特人 (*Homo neanderthalensis*) 曾被定为一个单独的属[9]，现在认为是早期智人[10]。我们自己的生物学学名是智人 (*Homo sapiens*)。两万年前在周口店地区生活过的山顶洞人与我们同是智人。地球上现在生活着的不同肤色的人类，都是同种智人。

相对于在地球上生活了 30 多亿年的细菌，哺乳动物是十分年轻的物种。个体生命周期的差异，使这种对比更为悬殊。

§3.2 生物的分类

生物分类体系是瑞典博物学家林奈 (Carolus Linnaeus，1707 – 1788) 建立的。他把一切生物分成界 (kingdom)、门 (phyla, 单数 phylum)、纲 (class)、目 (order)、科 (family)、属 (genera, 单数 genus)、种 (species) 七级，每级还可再冠以前缀超 (super) 或亚 (sub)，分出新的层次。一个具体物种的学名由属名和种名两个拉丁字组成，后面还可以标注首次发现的地名和发现者的名字。例如，动物界 (Animalia) 脊索动物门 (Chordata) 哺乳纲 (Mammalia) 食肉目 (Carnivora) 猫科 (Felidae) 豹属 (*Panthera*) 的金钱豹，学名是 *Panthera pardus*。只有种名小写，其他都用大写开头。

使用已经作古的拉丁文，是为了物种的统一命名不因民族语言而分歧。林奈甚至把自己的姓名也用拉丁文拼写。由 *Linnaeus carolus* 拟定的

[9] 可参看 *Science* **241** (1979) 118 – 133。

[10] 吴新智，“寻找人类祖先的足迹：智人”，《科学》(双月刊)，**52** (2000) 18 – 20。

学名通常缀以 L。

所有的生物首先分成原核生物 (prokaryote) 和真核生物 (eukaryote)。原核生物多为单细胞或聚居成丝状。它们的 DNA 没有用膜包裹起来形成细胞核，而是聚在称为拟核的区域里。它们没有微管蛋白、肌动蛋白和组蛋白，细胞里面也没有线粒体或叶绿体这类细胞器。这些特征使它们明确有别于真核生物。从单细胞的酵母到人都属于真核生物。真核生物的 DNA 借助组蛋白形成多个染色体，染色体再由双层磷脂膜包在细胞核里面。细胞核的膜上开有用蛋白质镶嵌好的孔洞。从 DNA 转录出来的信使 RNA，经过加工之后由核孔送到细胞质去。真核生物又区分成原生生物、真菌、植物、动物等“界”。

目前在地球上栖息的生物，尽管形态和生活方式千差万别，但遗传密码的统一性和基本生物化学过程的一致性，使人们相信它们都是由一个共同的祖先演化而来。根据生物形态学作出的分类，同时也给出了追溯演化过程的参考。辅以古生物化石的研究，可以粗线条地构建物种的亲缘关系或亲缘树。分子生物学的进展，特别是大量核酸和蛋白质数据的积累，使得人们能够从分子水平追溯亲缘关系，构建亲缘树或演化树。这也是生物信息学的一项重要内容。

生物“界”的划分，在 20 世纪 70 年代末发生的一次重大变化，就来自分子水平的对比研究。Carl Woese 等人发现，原核生物事实上分成两大集团，即古细菌 (archaea) 和 真细菌 (eubacteria)。古细菌其实更“新”一些，真核生物是从中分出来的。Woese 等建议把原核生物再分成两个界。并不是所有的学者都赞成他们的意见。因此，现在生物分界，有三界、五界、六界、八界之争[11]。下面这本普及书，是了解演化和分类的好参考，尽管作者们也不同意 Woese 的主张。

R-89 Lynn Margulis, and Karlene V. Schwartz, *Five Kindoms. An Illustrated Guide to the Phyla of Life on Earth*, W. H. Freeman and Co., 1982, 1988, 1998.

[11] 参看 N. A. Campbell, *Biology*, 4th ed., Benjamin/Cummings, 1996，第 495 页。

§3.3 模 式 生 物

对地球上现存物种的总数有不同估计，一般认为有 500 万到 3000 万种。科学家们当然不可能对如此多样的物种逐一研究。通过集中研究一些典型的模式生物，人们获取了丰富的知识。从简单到复杂，研究得最多的模式生物有：

R-90 噬菌体 (bacteriophage)。这是细菌的病毒，例如 $\Phi\chi174$、λ、T4、T7 噬菌体等等。噬菌体并不总是坑害细菌，它们有时候也把自己接到细菌的 DNA 里，请细菌帮助繁殖。这时称为前噬菌体 (prophase)。

R-91 病毒，如猿猴病毒 SV40，人艾滋病毒 HIV 等。病毒和噬菌体是高度发展了的寄生生物。它们除了作为遗传物质的 DNA 或 RNA 外，只保留了极少蛋白质来帮助保护自己和入侵宿主。它们的 DNA 与宿主有较多关系，有些就是演化过程中从宿主那里偷来的。因此，在分类上把它们单作一群，不好嫁接到演化树的枝杈上。

R-92 大肠杆菌 (*Escherichia coli*)。这是研究得最为详尽的一个模式生物，分子生物学的许多重要发现都是用大肠杆菌做出来的。这种只有 1.6 微米长的、可以迅速繁殖的单细胞生物，已经成为实验室和基因工程的重要工具。有关大肠杆菌的数据库很多，如 K12 菌株的基因组数据库 [R-346]、ECDC [R-347]、EcoGene [R-348]、RegulonDB [R-349]、EcoCyc [R-552]、MetaCyc [R-552] 等等。

R-93 酿酒酵母 (*Saccharomyces cerevisiae*)。英文俗名叫 baker's yeast 或 budding yeast 或简称 yeast，我们在本书中就叫它酵母。这个属于真菌界的单细胞真核生物，有 16 个染色体，在某些方面与人已经有不少共同之处。它的完全基因组已在 1996 年测定。与酵母有关的数据库有 SGD [R-358]、LISTA [R-359]、MIPS [R-139]、YIDB [R-361]、YPD [R-499] 等。

R-94 秀丽线虫 (*Caenorhabitidis elegans*)。英文又叫 nematode，文献中有时直称 worm。这种透明的、生活在海滩泥沙中的小虫是细胞数目一定的动物。它在发育过程中细胞数目超过 1 000，但成虫只有 959 个细胞，其中包括 302 个神经元。发育过程中自动调控的细胞凋亡 (apoptosis)，是近来热门研究课题之一。线虫的 6 个染色体中 9 700

万核苷酸的排列顺序，已经在 1998 年底基本上测定。美国《科学》周刊特别为此发了专集 [12]。

R-95 果蝇 (*Drosophila melanogaster*)。这种繁殖很快、容易诱发变异的小昆虫，已经为遗传学带来了许多知识。果蝇的总长达 1.8 亿核苷酸的基因组的主要部分已在 2000 年初发表，见 [R-369]。与果蝇有关的数据库有 FlyBase [R-371]、FlyNets [R-372]、GIF-DB [R-373]、Flyview [R-501]、Flybrain [R-502] 等。

R-96 拟南芥 (*Arabidopsis thaliana*)。这种个体生活周期只有 6 周的十字花科小草，是一种理想的模式植物。与拟南芥有关的数据库见 MATDB [R-391]、AtDB [R-392]、DAtA [R-394]、AGR [R-396] 和 TIGR-AT [R-397] 等。

R-97 水稻 (*Oryza sativa*)。作为亚洲人民主要食物的水稻，其基因组计划是中国和日本的研究重点。水稻基因组是小麦的 1/37。1997 年包括中国在内的 10 个国家或地区开始实行国际水稻基因组计划 IRGSP。2000 年 4 月初，孟山都公司宣布完成了水稻全部 12 个染色体 DNA 的“工作草图”，并将把它提交给 IRGSP 继续研究。日本的水稻基因组数据库 INE 见 [R-568]。中国的水稻基因组计划的进展见中国科学院国家基因组研究中心的网页 [R-175]。

R-98 非洲爪蟾 (*Xenopus lavias*)。它的一粒受精卵在 24 小时内就分裂到各种器官初具雏形的程度，因而很便于研究。参看 Axeldb[R-506]。

R-99 斑马鱼 (*Danio rerio*) 英文俗名 Zebra fish。这是一种通体透明的小鱼，生活周期约三个月，是研究脊椎动物发育过程的良好对象。美国国家卫生署 1997 年即建立了斑马鱼网页 [R-376]。斑马鱼基因组数据库见 ZFIN [R-377]。

R-100 家鼠 (*Mus musculus*)，它的基因组大小同人类相近，有约 30 亿个核苷酸对，组织在 19 对染色体里。家鼠的完全基因组原来预计在 2008 年全部测定，很有可能提前。与家鼠有关的数据库见 [R-379]、MGD [R-380]、MTB[R-536] 等。

R-101 当然，人 (*Homo sapiens*) 自己是重点研究的典型物种。事实上，在 GenBank [R-212] 等数据库中绝大多数序列来自人。

---

[12] *Science* **282**，1998 年 12 月 11 日。

§3.4 构成生物的四类分子

从化学成分看，生物体内除了水、无机盐类和离子，主要有四类分子，其中三类可以形成大分子。见表 3.1 。

表 3.1 构成生物的四类分子

小分子	大分子
单糖、双糖	多糖、淀粉、糖原、纤维素
脂肪酸	
核苷酸	核糖核酸 (RNA) 和脱氧核糖核酸 (DNA)
氨基酸	蛋白质

3.4.1 单糖、双糖和多糖

糖类是碳水化合物。单糖包括葡萄糖、果糖和半乳糖等。双糖中麦芽糖由两个葡萄糖组成，蔗糖由葡萄糖和果糖组成，乳糖由葡萄糖和半乳糖组成。单糖可以聚合成线性或分支的多糖大分子。同核酸或蛋白质比，多糖分子并不很大，也不包含许多信息，但在分子识别和免疫方面有一定作用。它们主要用于储存能量或作结构材料。植物用于储存的直链淀粉 (amylose) 是线性分子，而支链淀粉 (amylopectin) 是分支大分子；动物储存能量用的糖原 (glycogen) 是分支更多的大分子，是具有 α 构型的葡萄糖的聚合物。作为结构材料的纤维素 (cellulose)，是具有 β 构型的葡萄糖的聚合物。植物细胞外面的壁主要由纤维素构成，昆虫的壳多糖 (chitin) 则是其外骨骼的主要成分。

3.4.2 脂肪酸

脂肪酸不能形成很大的聚合分子。它是脂肪 (fats)、油 (oils) 和磷脂 (phospholipids) 的结构成分。一端亲水、一端疏水的磷脂分子是组成生物膜的主要材料。类固醇 (steroids) 是不含脂肪酸的脂类，它们作为激素和维生素在体内起着重要作用。

3.4.3 核苷酸和核酸

核酸大分子是由四种单体聚合成的一维高分子链。遗传信息就编码在这些单体的不同排列次序上。每个单体由三部分组成：一个五碳糖，其五个碳原子从 1′ 编号到 5′；一个接在 5′ 碳上的磷酸根；一个连在 1′ 碳上的碱基，总称为一个核苷酸。共有五种碱基和相应的五种核苷酸。三种碱基是含氮的六元杂环，即胞嘧啶、胸腺嘧啶和尿嘧啶，形成胞苷酸 (C)、胸苷酸 (T) 和尿苷酸 (U)。两种碱基是五元环配六元环的杂环，即腺嘌呤和鸟嘌呤，形成腺苷酸 (A) 和鸟苷酸 (G)。碱基杂环中的碳和氮原子统一编号，数码上不加′。如果五碳糖的 2′ 和 3′ 位上都是羟基 OH，就是核糖 (ribose)。如果 2′ 位上脱氧，只剩下 H，那就成为脱氧核糖 (deoxyribose，缩写 d)。当 3′ 位上也脱氧，就成为双脱氧核糖 (dideoxyribose，缩写 dd)。5′ 位上的磷酸根有单、二、三磷酸之分，分别以字母 M、D、T 表示。这些可能性组合起来，就成为生物化学文献中常见的种种缩写。例如，ATP、dGDP 和 ddCTP 分别表示三磷酸腺苷 (腺三磷)、脱氧二磷酸鸟苷和双脱氧三磷酸胞苷。有时写 NTP、dNTP 和 ddNTP，用 N 代表 A、C、G、T 或 U 中任何一个。

核苷酸单体聚合成大分子时，前一个核糖 3′ 位上的羟基和下一个核糖 5′ 位上的磷酸根脱水形成磷酸二酯键。因此，核酸大分子是有方向的一维链，通常从 5′ 端看到 3′ 端。脱氧核糖核酸 DNA 由 A、C、G、T 聚合而成，核糖核酸 RNA 由 A、C、G、U 聚合而成。在演化历史上，RNA 可能早于 DNA 出现，这是“RNA 世界”的观点：

R-102 W. Gilbert, *Nature* **319** (1986) 618.

R-103 R. F. Gresteland, and J. F. Atkins, *The RNA World*, Cold Spring Habor Laboratory Press, 1993.

聚合过程中如果遇到 ddNTP，就无法继续。这一事实后来启发了一种 DNA 测序方法 (见 3.6.5)。两条“互补”(或叫“共轭”) 的 DNA 链靠氢键维系。A 与 T 间有两个氢键，称弱耦合；G 和 C 间有三个氢键，为强耦合。DNA 双链进一步形成螺旋结构。从携带信息的角度看，DNA 双螺旋中的一条已经含有全部信息，但两条链并不是等价的。后面讨论 DNA 复制和基因在链上分布时，都会看到这一不等价性。RNA 通常是

表 3.2 碱基 (核苷酸) 的标准符号

符号	意义	符号	意义
A	腺嘌呤	M	A 或 C (氨基)
C	胞嘧啶	S	G 或 C (强耦合)
G	鸟嘌呤	W	A 或 T (弱耦合)
T	胸腺嘧啶	B	G 或 T 或 C (非 A)
U	尿嘧啶	D	G 或 A 或 T (非 C)
R	G 或 A，嘌呤	H	A 或 C 或 T (非 G)
Y	T 或 C，嘧啶	V	G 或 C 或 A (非 T)
K	G 或 T，酮	N	A 或 G 或 C 或 T，即任意
X	未知	-	不定长度的空隙

单链，但可借助不同部位上的互补或反序互补片段的耦合，形成一些二级结构。这类二级结构往往对基因的表达起调控作用。由 RNA 的一维字母序列预测二级结构，是生物信息学的课题之一。

表 3.2 给出由国际生物学联合会 (IUB) 和国际纯粹和应用化学联合会 (IUPAC) 共同制定的核苷酸的标准符号[13]。核苷酸虽然只有 5 种 (DNA 中的 T 在 RNA 中换成 U)，它们的各种组合都有一定的符号代表，即所谓“多义” (ambiguity) 符号。许多软件能识别这些符号。

3.4.4 氨基酸和蛋白质

蛋白质是由氨基酸聚合而成的生物大分子，单体数目从数十到数千不等。很短的氨基酸链不能独立地折叠成特定的三维结构，通常叫做多肽，不称为蛋白质。氨基酸是比核苷酸略小的有机分子。它的中心碳原子，特称为 α 碳 (C_α)。C_α 的四个化学键，一个接羧基 (COOH)、一个接氨基 (NH_2)、一个简单地连氢 (H)；只有第四个键上的侧链 R，从一个 H 到接近 30 个原子的基团，共有 20 种组合，导致 20 种氨基酸。自然界和实验室里合成的氨基酸不止此数，但所有的蛋白质只由这 20 种氨基酸组成。R=H 的甘氨酸，左右对称，不具有光学活性。其他 19 种氨基酸都有左、右之分，具有光学活性。氨基酸聚合成大分子时，相邻的氨基

[13]IUPAC-IUB Commissions on Biochemical Nomenclature, "Abbreviations and symbols for nucleic acids, polynucleotides and their constituents", *Eur. J. Biochem.* **15** (1970) 203 – 208.

和羧基缩水形成相当强的肽键. 因此, 蛋白质也是有方向的一维链, 带氨基的一头称为 N 端或记为 N′, 另一头带羧基称为 C 端, 常用 C′ 表示.

氨基酸有三字母和单字母两套符号, 前者便于记忆, 后者便于计算机处理. 表 3.3 给出由 IUB 和 IUPAC 共同制定的氨基酸标准符号. 许多通用的软件如 BLAST(见 [R-631]) 都接受表 3.2 和表 3.3 中的符号.

表 3.3 氨基酸标准符号

符号	意义	符号	意义
A(Ala)	丙氨酸	P(Pro)	脯氨酸
B	天冬氨酸或天冬酰胺	Q(Gln)	谷氨酰胺
C(Cys)	半胱氨酸	R(Arg)	精氨酸
D(Asp)	天冬氨酸	S(Ser)	丝氨酸
E(Glu)	谷氨酸	T(Thr)	苏氨酸
F(Phe)	苯丙氨酸	U	硒代半胱氨酸
G(Gly)	甘氨酸	V(Val)	缬氨酸
H(His)	组氨酸	W(Trp)	色氨酸
I(Ile)	异亮氨酸	Y(Tyr)	酪氨酸
K(Lys)	赖氨酸	Z	谷氨酸或谷氨酰胺
L(Leu)	亮氨酸	X	任意
M(Met)	甲硫氨酸	*	翻译终止
N(Asn)	天冬酰胺	-	不定长度的空隙

核酸是遗传信息的携带者, 而蛋白质是信息转化成生物结构和功能的表达者. 蛋白质按照外形和在生物组织中的位置和作用, 粗略地分成三大类:

第一, 纤维蛋白 (fibrous protein). 筋骨中的胶原 (collagen), 毛发中的角蛋白 (keratin), 皮肤羽毛中的表皮素 (epidermin) 等.

第二, 跨过或部分镶嵌在磷脂膜中的膜蛋白. 它们的功能是实现膜内外的信息交换或物质传递.

第三, 大致为球形的球蛋白. 它们的种类最多, 其中一大部分是各种生物化学反应的催化剂, 即酶. 这也是最重要、最多样化的一类蛋白质. 许多生化反应如果没有相应的酶协助, 反应速率甚至会降低到原来的百万分之一以下, 以致事实上停止进行.

3.4.5 遗传密码

在 DNA 序列的编码区，每三个核苷酸翻译成蛋白质中一个特定的氨基酸。表 3.4 按转录后的 mRNA 给出通用的三联体密码，即 T 已换成 U 。表中没有再写氨基酸的中文名字，这是因为考察数据库里的蛋白质序列时，必须熟记这些字母。

表 3.4 通用遗传密码表

第一字母	第二字母				第三字母
	U	C	A	G	
U	F(Phe)	S(Ser)	Y(Tyr)	C(Cys)	U
	F(Phe)	S(Ser)	Y(Tyr)	C(Cys)	C
	L(Leu)	S(Ser)	终止	终止 *	A
	L(Leu)	S(Ser)	终止	W(Trp)	G
C	L(Leu)	P(Pro)	H(His)	R(Arg)	U
	L(Leu)	P(Pro)	H(His)	R(Arg)	C
	L(Leu)	P(Pro)	Q(Gln)	R(Arg)	A
	L(Leu)	P(Pro)	Q(Gln)	R(Arg)	G
A	I(Ile)	T(Thr)	N(Asn)	S(Ser)	U
	I(Ile)	T(Thr)	N(Asn)	S(Ser)	C
	I(Ile)***	T(Thr)	K(Lys)	R(Arg)**	A
	M(Met)	T(Thr)	K(Lys)	R(Arg)**	G
G	V(Val)	A(Ala)	D(Asp)	G(Gly)	U
	V(Val)	A(Ala)	D(Asp)	G(Gly)	C
	V(Val)	A(Ala)	E(Glu)	G(Gly)	A
	V(Val)	A(Ala)	E(Glu)	G(Gly)	G

请注意，64 个密码子 (codon) 中有三个终止密码子 UAA 、UAG 和 UGA ，其余 61 个密码子编码 20 种氨基酸，因此有些氨基酸有多种编码 (简并)。具体而言：

三种氨基酸有 6 重简并编码：亮氨酸 Leu 、丝氨酸 Ser 和精氨酸 Arg 。五种氨基酸有 4 重简并编码：缬氨酸 Val 、脯氨酸 Pro 、丙氨酸 Ala 、甘氨酸 Gly 和苏氨酸 Thr 。有 3 重简并编码的是异亮氨酸 Ile 和终止密码子。九种氨基酸有 2 重简并编码：苯丙氨酸 Phe 、酪氨酸 Tyr 、组氨酸 His 、谷氨酰胺 Gln 、天冬酰胺 Asn 、赖氨酸 Lys 、天冬氨酸 Asp 、

谷氨酸 Glu 和半胱氨酸 Cys。只有两种氨基酸具有单重编码：甲硫氨酸 Met 和色氨酸 Trp。

还应指出，表 3.4 中用几个星号 * 给出的是脊椎动物线粒体遗传密码与通用密码的差别：* 编码色氨酸 Trp；** 编码终止符号；*** 编码甲硫氨酸 Met。因此，线粒体中没有单重编码的氨基酸。线粒体的 DNA 系自我复制，但有一部分蛋白质要在细胞质中合成，再输送到线粒体里。因此，有一种观点认为线粒体本是独立生活的微生物，后来被俘获形成内共生关系。从 EBI [R-131] 或 NCBI [R-134] 的网页可以查到多种遗传密码的例外情形：

R-104 EBI 的遗传密码一览表 (Genetic Code Viewer)。网址：

`http://www2.ebi.ac.uk/`

R-105 NCBI 的 **Genetic Codes** 表，可以通过其网页

`http://ncbi.nlm.nih.gov/`

的 Taxonomy 选项或 ORF Finder [R-710] 进入。这些表的编号就是 GenBank [R-212] 数据库条目中给出的 translation table 号。

顺便提一下，植物叶绿体也有自己的 DNA，甚至比线粒体 DNA 还大。它们也可能是演化过程中形成的共生物。细菌除了基本的 DNA 链或环之外，还可能含有数量会变化的 DNA 小环或链，称为质粒 (plasmid)。例如，细菌的抗药性往往就编码在质粒中。野生型细菌常有多种质粒，实验室培养若干代之后，质粒的品种数量会减少。这些都是染色体之外的遗传信息携带者。

§3.5 分子生物学的中心法则

DNA 双螺旋结构的发现者之一 F. H. C. Crick 在事实尚不充分的 1957 年，把分子生物学中的主要关系概括为一项“中心教条” (central dogma)[14]，它后来被众多实验事实所证实和补充，成为图 3.1 所示的中心法则。简单地说，DNA 双螺旋是遗传信息的携带者，它在一定条件下可以准确地自我复制。遗传信息只能通过最终的蛋白质产物体现或“表

[14] F. H. C. Crick, "On protein synthesis", *Symp. Soc. Exp. Biol.* **12** (1957) 138–163; "Central dogma of molecular biology", *Nature* **227** (1970) 561–563.

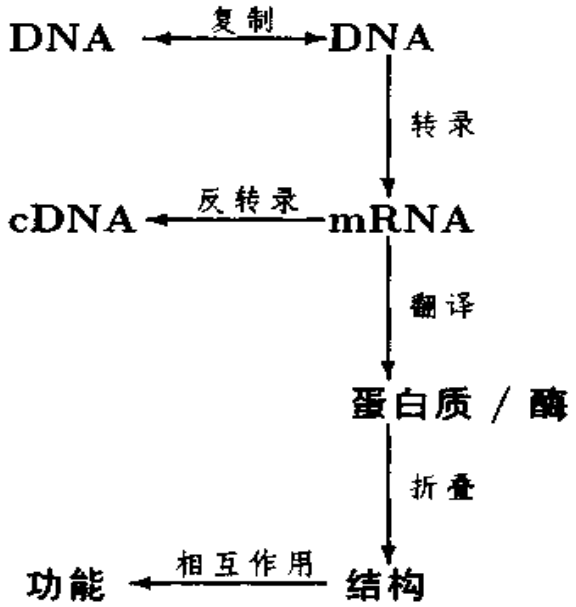

图 3.1 分子生物学的中心法则

达”出来. 为此要先把信息“转录”到单股的信使 RNA，即 mRNA 链上. 后者与前者的差别，仅在于把 DNA 序列中的 T 换成 U、然后再有所剪裁. 细胞液中有大量核糖体，它们是根据 mRNA 上的信息制造蛋白质的生物化学工厂. 新生的蛋白质要折叠成特定的三维形状，才能有生物活性，在生命过程中发挥功能.

在 DNA 序列中的“基因”一般用斜体字母命名，而基因表达的产物蛋白质，用相应的正体字母表示，且首字母大写. 例如，一个名为 swallow (*swa*) 的基因，表达后的蛋白质记为 Swa .

下面分成小节，简略说明中心法则提到的几件事.

3.5.1 DNA 的复制

DNA 的自我复制是细胞周期中的重要事件. 一旦复制开始，细胞当然不能分裂；而 DNA 复制的结束就会触发细胞的分裂. 复制过程靠许多种酶帮助，其中最重要的是 DNA 聚合酶[15] . DNA 聚合酶有两个特性对

[15] DNA 聚合酶有许多类，这里主要讲最重要的第 III 类，即 DNA polymerase III .

理解复制过程十分重要.

第一，它作用的方向，只能从 5′ 端往 3′ 端发展.

第二，它不会凭空促进聚合作用，而必须以一条 DNA 单链作模板，模板的 3′ 端要先有小小一段耦合好的双链引物 (primer) . 这些引物是引物酶 (primase) 协助合成的小段 RNA . 引物的 3′ 端悬在那里， DNA 聚合酶就从那里开始根据模板要求，把适当的核苷酸按 5′ → 3′ 方向聚合上去，形成双链.

仔细思考一下，就发现复制过程并不简单. 首先， DNA 双螺旋要在复制起点解旋，暴露出两条单链作聚合模板. 复制起点是特定的，例如在大肠杆菌中是称为 *Ori*C 的位点. 复制起点两侧，形成两个“复制叉”，在电子显微镜下像一只眼睛 (“复制眼”) . 两个复制叉或向相反方向发展，或一个固定、一个前进. 在复制叉后面的两条单链，一条链从 3′ 到 5′ ，其引物方向正好允许 DNA 聚合酶顺利地按模板聚合出相应的共轭链；这是复制的先导链. 另一条单链从 5′ 到 3′ 方向，其上 RNA 引物在 DNA 聚合酶帮助下只能向着与复制过程相反方向延长，长出一个小片段. 这时复制叉后面的空白单链上已经形成又一小段引物，往反方向聚合新的双链片段. 这些分别形成的冈崎片段 (Okazaki fragments) ，要在特定的酶协助下修补、连接，同时第 I 类 DNA 聚合酶把引物 RNA 变成 DNA ，最终产生完整的共轭链. 这条手续繁杂的链，成为复制过程中的滞后链.

无论是原核生物还是真核生物，都按上述模式复制. 只是真核生物可能有多个复制起点，参与复制过程的酶数目和品种有所不同. 为分子克隆制备“载体”时，都要恰当地包含复制起点，外源基因才能在宿主细胞内复制增殖，见 3.6.2 小节.

3.5.2 DNA 到 mRNA 的转录

双股 DNA 螺旋的每一股上，都散布着长短不等的包含遗传信息的片段，即“基因”. 观察一个基因，从 5′ 端往 3′ 看，首先是一段并不翻译成蛋白质的区域，即所谓 5′UTR 区 (UTR=UnTranslated Region) . 这一区域内有一些起控制作用的“字”，通常是某种蛋白质 (酶、因子) 的结合位点，例如启动子 (promotor) 、增强子 (enhancer) 等. 然后是编码区的起始密码子，最常见的是编码甲硫氨酸的 ATG ，也有不少例外. 编码区

以三个终止密码子之一结束。然后是 3'UTR 区，这里主要是一些有关结束转录过程的信息。

启动子是 RNA 聚合酶的结合点，由此开始转录。启动子前后还有若干其他起控制作用的 DNA 片段。特别是在真核生物中，这些控制片段更为多样。典型的启动子常包含 TATA 片段、CAAT 片段等，但没有这些片段的启动子也不少。各种转录因子帮助 RNA 聚合酶结合到控制片段上，启动和完成 RNA 的转录。目前基因组测序速度远远超过实验验证的可能性。因此，人们不得不越来越多地依赖计算机寻找基因。有关启动子和转录因子的数据库和软件很多，如 TRANSFAC [R-219]、TRRD [R-221]、COMPEL [R-227] 等，还有一些专门识别调控片段的程序，如 GeneExpress [R-720]、Promotor Scan [R-727]、Signal Scan [R 728]、TFSearch [R-729]、PatSearch [R-730]、PromFD [R-734] 等。

真核生物刚从 DNA 转录出来的 mRNA 前体 (pre-mRNA) 还要继续加工，才能作为成熟的 mRNA，经过核孔送到细胞质中的核糖体去翻译成蛋白质。加工的主要内容是剪去不表达的内含子 (intron)，把将来要表达的外显子 (exon) 连接起来。内含子和外显子的长短多寡不一。酵母有一个长 76 个碱基的 tRNA 基因，被长度为 14 个碱基的内含子断开。人的甲状腺球蛋白基因总长约 10 万个碱基，被 40 多段内含子隔开，真正编码蛋白质的序列只有约 8500 个碱基。预测内含子和外显子的剪接点，是生物信息学的一项重要课题。现在已经有多个内含子、外显子和剪接的数据库，如 ASDB [R-242]、IDB [R-244]、ExInt [R-246] 和 Intronerator [R-243] 等。

3.5.3 mRNA 翻译为蛋白质

核糖体是根据 mRNA 上的编码信息制造蛋白质的生物化学工厂。一个细菌细胞里大约有两万个核糖体，而真核细胞里则多达百万。它们的结构大同小异，都是由相当复杂的 rRNA 骨架和许多蛋白质组成的复合体，由大小两个亚基组成。大肠杆菌的核糖体大亚基由一个 23S rRNA 和一个 5S rRNA 作骨架，上面结合了 31 个蛋白质，而小亚基的 16S rRNA 骨架上结合了 21 个蛋白质 (关于 S 这个特别单位，请参阅 3.6.4 小节)。最近发表了细菌核糖体小亚基的 5.5Å(0.55nm) 分辨率的结构[16]。哺乳动

[16] W. M. Clemons, *Nature* **400** (1999) 833 - 840.

物核糖体大亚基有 28S 、 5.8S 和 5S 三个 rRNA 以及 49 个蛋白质，小亚基有一个 18S rRNA 和 33 个蛋白质。

下面以原核生物为例，考察一下根据 mRNA 所携带的信息制造蛋白质的“翻译”过程。

首先，有一批酶协助 mRNA 和核糖体完成蛋白质的生产：起始因子帮助 mRNA 先和空闲的小亚基结合，找到携带第一个甲硫氨酸的 tRNA ，再把它们和大亚基拼接到一起，开始翻译；延长因子使翻译过程一直继续下去，新生的肽链不断延长；最后达到终止密码子时，由结束因子终止翻译过程；新生的蛋白质肽链和 mRNA 离开核糖体，大小亚基分开，等待下一轮合成任务。起始因子，延长因子和结束因子都有多种，各司其职，分工合作。

真正的“翻译”是由一大类 tRNA 完成的。每个 tRNA 有一只由三个相连的“反”密码子组成的“脚”，和一只抓住与相应密码子对应的氨基酸的“手”。当它的脚踩住 mRNA 上恰好合适的密码子时，那个氨基酸就被带来接到新生肽链的末端。一般说来，同一时间在核糖体里，有三个 tRNA ：一个携带着下一步需要的氨基酸；一个带着已经合成的肽链，准备把新的氨基酸接上去；第三个 tRNA 在上一步里已经把接好氨基酸的肽链转交给带来氨基酸的那个 tRNA 、现在空着手准备离开核糖体到细胞质里去寻找合适的氨基酸，继续执行运输任务。由于遗传密码的“简并”， tRNA 有许多种。即使是密码唯一的甲硫氨酸，对应于起始密码 AUG 和延长用的 AUG ，其 tRNA 也有差别。据说，细胞质中至少要有 31 种 tRNA 和相应的氨基酸，翻译过程才能不断进行。

原核生物细胞中，许多核糖体可以“骑”在一条 mRNA 上复制蛋白质，一条 mRNA 可以多次参与翻译过程。各种酶和 mRNA 、 tRNA 、 rRNA 等，既是翻译过程的执行者，又是翻译的产物。它们都寿命有限，最终被其他的酶降解。只要生命在继续，就要不断地合成蛋白质。人体内蛋白质的平均寿命约为两周。

3.5.4 mRNA 的反转录与 cDNA

最初人们曾经以为，遗传信息只能从 DNA 传到 mRNA，再从 mRNA 翻译成蛋白质，由蛋白质来“表达”这些信息，即体现为各种生物功能。然而， 1970 年 D. Baltimore 和 W. H. Temin 等人同时发现[17]，有些 RNA 病毒会把 RNA 反转录成 DNA，并且找到了促成这一过程的反转录酶。人们扩展了对中心法则的认识 (参看脚注所引同一期《自然》周刊，1198 – 1199 页)。更重要的事实是，反转录酶可以在试管里把 mRNA 反转录成 DNA，这样的 DNA 里没有内含子，特称为互补 DNA (complementary DNA，简称 cDNA)。

真核生物每个细胞核里都有全套染色体和遗传信息。然而，在不同的组织和环境中，只有一部分基因被表达为蛋白质。所有要表达的基因，都有相应的 mRNA 被转录和加工。原则上可以提取一定组织如肝脏细胞中的全部 mRNA，把它们反转录成稳定而便于保存的 cDNA，形成 cDNA 库 (注意，这不是“文”库或数据库，而是存放在容器中的实物)。目前可以从外国基因工程公司，购买现成的一定组织器官的 cDNA 库，从中发现未知的基因。

3.5.5 蛋白质的剪接

20 世纪 90 年代初发现，有些新生肽链要剪去中间一段，把两边连接起来，才变为成熟的功能蛋白。这称为蛋白质的剪接。与内含子 (intron) 和外显子 (exon) 类比，被剪切掉的肽链称为“内质” (intein) 或“蛋白质内含子”，而保留下来的部分称为“外质” (extein)。内质序列的 N 端大约有 100 个氨基酸，C 端大约有 50 个氨基酸，构成剪接区。这两个剪接区各自有一些保守的模体 (motifs)。详细情况可参看 InBase [R–436] 数据库及库中文献单。

[17] D. Baltimore 和 W. H. Temin and S. Mizutani 的两篇文章发表在同一期《自然》周刊上，见 *Nature* **226** (1970) 209 – 213。

3.5.6 蛋白质的折叠

新生的肽链必须折叠成唯一的、特定的三维结构，才能发挥生物活性，成为真正的蛋白质. C. B. Anfinsen 的早期实验，证明折叠所需信息完全包含在氨基酸排列成的一维链中，他因此荣获 1972 年诺贝尔化学奖。氨基酸或代表它们的 20 个字母的排列顺序，称为蛋白质的一级结构. 二级结构是由氢键维系的 α 螺旋和 β 片 [18]。三级结构是完全折叠好的蛋白质的空间结构. 四级结构是多个蛋白质亚基组成蛋白质复合体的结构。

目前， X 射线晶体衍射分析和核磁共振 (NMR) 是测定三级结构的主要手段。做 X 衍射要求事先把蛋白质结晶，而这远非易事. NMR 虽可在溶液中做而不要求结晶，但目前只能分析较小的蛋白质。蛋白质三维结构的测定，虽然从 20 世纪 50 年代的几年测一个结构，发展到现在每个月平均测定 160 个以上，但仍然远远落后于核酸序列的测定速度. 蛋白质结构数据库如 PDB [R-441]，在 1999 年底收有详细三维原子坐标的蛋白质虽已超过 1 万种，但从基因序列翻译出的蛋白质序列，增长速度月以千计，完全不可能依靠实验手段一一测定他们的结构和功能. 因此，蛋白质结构和功能的预测成为生物信息学的重要任务. 分析已知蛋白质序列和结构，为从氨基酸序列预测蛋白质的结构与功能，提供愈益增加的根据。

近年来人们注意到，在二级结构和三级结构之间，由 α 螺旋和 β 片组装成的紧凑折叠起来的单元，对于蛋白质结构的分类和预测有重要作用，称为“折叠单元”或简称折叠 (fold). 尽管蛋白质序列数目以百万计，折叠的种类却极为有限，很可能不超过 1000 种.

另一方面，蛋白质的氨基酸序列中有一些在演化过程中最为保守的单元，称为结构域 (domain)。一个结构域不能再划分为更小的结构域. 有的蛋白质只包含一个结构域，有些蛋白由多个结构域串起来组成. 结构域通常对应二级结构的某种紧致排列，可以相对独立地进行折叠，并有疏水核心. 具有序列相似性的结构域，对应相同的折叠单元；这是同源性的佐证. 具有相同折叠的结构域，不一定同源. 蛋白质同源性的分析，宜着眼结构域层次，而不从整个氨基酸序列入手. 关于结构域的早期讨论，可参看：

[18] 为避免与 fold 混淆，我们不用 β “折叠” 或 “折叠片” 的说法.

R-106 J. Janin, and C. Chothia, *Meth. Enzymol.* **115** (1985) 420 430.

在比结构域更小的层次上，序列上有相似性的一些区域称为 motif，建议音译为“模体”[19]。一些非球蛋白的跨膜螺旋、卷曲螺旋 (coiled coil)、乃至信号肽链，有时也被称为模体。关于模体的较近讨论，请看

R-107 P. Bork, and E. V. Koonin, *Curr. Op. Struct. Biol.* **6** (1996) 366 - 367.

人们有时也在比折叠单元稍小的层次使用模体一词，指由 α 螺旋和 β 片形成的、出现在许多彼此无关的蛋白质折叠中的较为固定的组合。

在本书第 4 章所介绍的数据库中，不少涉及结构域、模体、模式 (pattern)、轮廓 (profile)、折叠等，它们并不都有彼此一致的定义和用法。读者切不可从字面取意，而应参考相应数据库的详细说明。

蛋白质折叠问题有两个层面。一是解释为什么如此众多的氨基酸序列只导致数目有限的折叠方式，二是预测给定序列的具体三维结构。前者是一个真正的物理问题，而后者是蛋白质设计或基因工程关心的焦点；正如说明物质三态和气液相变是物理学的任务，而测量或计算出酒精的确切沸点是材料科学的课题。为了说明前一问题，有人引入了蛋白质折叠的最简单模型，即把 20 种氨基酸归并为疏水 (hydrophobic 即 H) 和极性 (polar 即 P) 两类，只考虑疏水核心和亲水外围位置导致的能量差别。这就是 HP 模型：

R-108 K. A. Dill, *Biochemistry* **24** (1985) 1501; H. S. Chan, and K. A. Dill, *Macromolecules* **22** (1989) 4559.

近来李浩、汤超等穷举了某些有限格点 HP 模型的结构和序列，引入结构的可设计性 (designability) 概念，说明确实有大量氨基酸序列对应可设计性高的少数结构：

R-109 H. Li, R. Helling, C. Tang, and N. S. Wingreen, *Science* **273** (1996) 666; *Proc. Natl. Acad. Sci.* USA **95** (1998) 4987.

对蛋白质数据库如 PDB [R-441] 和 Dali [R-467] 中的实际序列作 HP 约化，可以看出可设计性与折叠早期迅速形成 α 螺旋有关：

R-110 C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao(郝柏林), C. H. Hsieh, and

[19] 有人把 motif 译为“基序”、“结构花色”等.

H. C. Lee, "The HP model, designability and alpha-helices in protein structures", *Phys. Rev. Lett.* **84** (2000) 386 - 389.

将 20 种氨基酸归并为 H 和 P 两类，虽能抓住折叠初期的某些特点，终属过度约化。有实验表明，归结为 5 类氨基酸的约化方式，可能足以反映折叠过程的实质：

R-111 D. S. Riddle 等 7 位作者， *Nature Struct. Biol.* **4** (1997), 805 - 809.

试图从"第一原理"出发，预测蛋白质三维结构的尝试，迄今收效不大。前几年有人估计，按当前计算速度的发展外推，可能到 2030 年左右可以实现。参看：

R-112 H. S. Chan, and K. A. Dill, "The protein folding problem", *Physics Today*, February 1993, 24 - 32.

§3.6 基因工程技术简介

分子生物学的许多发现，可以用来加工特定的 DNA 片段，用生物方法大量产生某些基因或基因产物。这就导致了基因工程和全新的生物技术。基因工程所用到的许多技术，也是实验室中不可或缺的手段。我们极其简要地叙述一些基本概念，主要是为了以后介绍生物数据库和软件算法时，知道一点背景。

3.6.1 限制性内切酶

20 世纪 60 年代末，在大肠杆菌中首先发现了一种酶，它会准确识别外来的 DNA，并且在特定的位点把后者切断。这就是限制性内切酶。为了保护自己的 DNA 不被误切，大肠杆菌还生产甲基化酶，把本身的 DNA 按一定规律甲基化，即把某些特定位置上的氢 (H) 换成甲基 (CH_3)。若干内切酶和甲基化酶组成一个微生物的一种防御系统。现在已经在各种微生物中发现 3000 种以上限制性内切酶，它们的识别位点也超过 300 种。许多限制性内切酶的识别位点是 4 ～ 8 个字母的"回文" (palindrome)，如 cgcg、ccgg、cctagg 等，它们在 DNA 双螺旋的两股上等同。当剪切点在两股 DNA 上错开，形成"粘端"，就更有利于加工后重新联接，因而在基因工程中有广泛应用。甲基化酶也已知多种，也有特定的甲基化位

点。请参看内切酶和甲基化酶的数据库 REBASE [R-424] 。限制性内切酶是遗传研究和基因工程的重要工具。把染色体 DNA 用两种识别位点不同的内切酶先后处理两次，测量出所得各片段的大小，原则上就可以恢复出这些识别位点在原来序列中的排列顺序。这叫做酶切图谱 (restriction map) 。

3.6.2 分子克隆

克隆 是一个极不成功但又约定俗成的译名，意思是用某种无性繁殖手段，再生产出生物分子、细胞、乃至个体。这里只讲分子克隆，即用生物方法而不是化学合成来复制生物大分子。下面省略许多技术细节，只讲基本概念。假定已经从某种染色体里分离出一段 DNA，要把它增殖到较多数量，才便于研究。我们请大肠杆菌来帮助。先选择一种载体 (vector)，通常是 DNA 序列已经清楚的质粒 (例如长度为 4 361 碱基对的 pBR322) 或噬菌体 (例如长度为 48 502 碱基对的 λ 噬菌体)。这些载体的酶切图谱可以在许多手册中查到。选取适当的内切酶把 DNA 的两端加工好，并把载体在设计好的位点切开。利用聚合酶和连接酶把 DNA 和相应的遗传标记连接进去。然后把这些带有所需 DNA 的载体引入大肠杆菌内。大肠杆菌并不能察觉质粒带有异物，噬菌体也会继续在大肠杆菌中繁殖。若干代之后，再利用原设的标记把载体分离，用内切酶割出所要的 DNA 片段，它们的数量已经大为增加。

分子克隆能复制的 DNA 大小，依赖于所用载体。普通质粒可容纳几千碱基对。选用包含名为 *cos* 的粘端的一段 λ 噬菌体接入质粒载体，可以克隆长达 45 000 碱基对的 DNA，特称为 粘粒 (cosmid)。

用质粒或噬菌体作载体，只能增殖较小的 DNA 片段。后来发现，可以制备酵母人工染色体 (Yeast Artificial Chromosome，简称 YAC)，令其随酵母的有丝分裂而增殖。

一个 YAC 载体是由以下几部分组成的双链 DNA 环：一个酵母染色体的着丝粒及将来分离用的遗传标记，着丝粒用来增加 YAC 的稳定性；一段带复制起点和遗传标记的大肠杆菌 DNA 序列，它还含有一个克隆位点；一对来自四膜虫 (*Tetrahymena*) 的染色体端粒。两个端粒用一段将来要切除的 DNA 连起来，成为闭环。环状 YAC 载体可像普通质粒一样在

大肠杆菌中增殖。把载体分离出来之后，用限制性内切酶切开克隆位点，把特定的 DNA 片段连接进去。用另一种内切酶把端粒间的 DNA 切除，形成线性的人工染色体。这样的 YAC 可在酵母细胞中增殖。用 YAC 可以克隆长达百万碱基对的 DNA 序列。关于 YAC 方法的最初描述，请参看：

R-113 D. T. Burke, G. F. Carle, and M. Y. Olson, "Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors", *Science* **236** (1987) 806 – 812.

许多细菌的 DNA 比酵母的单个染色体还大，也可用以制备细菌人工染色体 (Bacterial Artificial Chromosome，简称 BAC)，来克隆长的 DNA 片段。请参看：

R-114 H. Shizuya 等 7 位作者，"Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector", *Proc. Natl. Acad. Sci.* USA **89** (1992) 8794 – 8797.

大肠杆菌 P1 噬菌体 也可以容纳较长的 DNA 片段，例如平均长度达 85000 碱基对的序列。这样的环状载体连同接进去的 DNA 序列，尺寸可比原来 P1 的 DNA 大很多，有时叫做 PAC。

BAC、YAC 和 PAC 都广泛应用于各种基因组的测序。

3.6.3 聚合酶链反应 (PCR)

20 世纪 80 年代初，在 Cetus 公司工作的 Kary Mullis 发明了一种可使 DNA 片段增殖百万倍的聚合酶链反应技术 (Polymerase Chain Reaction，简称 PCR)。现在已经发展出 PCR 的许多变种，我们只介绍最基本的做法。实现 PCR 需要以下条件：微量待增殖的双链 DNA 片段，耐热的 DNA 聚合酶，恰当的引物，足够的 dNTP 单体，以及促进酶活性的镁离子等。先把上述混合物加热到 94°C 保温 5 min，双链分离成单链 DNA。降温到 30°C ～ 65°C，保持 30 min，引物结合到单链 DNA 的左、右端。在 65°C ～ 75°C 保持 2~5 min，DNA 聚合酶根据单链模板把引物延长成双链。这时所要的 DNA 数量已翻番。再升温至 94°C 重复以上过程。理想情况下，20~30 次循环就可以增殖 $2^{20} \sim 2^{30}$ 倍。PCR 所用的 DNA 聚

合酶不应因高温处理而降低活性。因此，从海底火山附近的嗜热细菌分离出来的 DNA 聚合酶对 PCR 的发展起了重要作用。关于 PCR 的早期描述，请参看：

R-115 K. Mullis 等 6 位作者，"Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction"，*Cold Spring Harbor Symposia on Quantitative Biology* **51** (1986), 263 - 273.

K. Mullis 于 1993 年获得诺贝尔化学奖。顺便指出，美国《科学》周刊曾在 1986 年拒发 Mullis 的文章，而 1989 年该刊把 PCR 选为当年的大事。

3.6.4 超速离心、凝胶电泳和印迹法

把混合在一起的种种大分子和细胞器等按分子量大小分离开，早在基因工程兴起之前就是生物化学实验室的日常要求。人们使用超速离心机、质谱分析、凝胶电泳等各种手段来实现分离。

超速离心机中大小分子集团沉降速度不同，带来了一个并不准确但已不能摆脱的计量单位，即沉降系数 S 或称 Svedberg 单位 。质量为 m 的分子集团受到的离心力是 $m(1-\alpha\rho)\omega^2 r$，这里 α 是分子集团的比容，ρ 是水溶液密度，ω 是旋转角速度，r 是距旋转轴的距离。离心力与摩擦力 kv 平衡时，沉降速度 v 可以算出来。通常取沉降速度和角加速度之比 $v/\omega^2 r$ 作尺度，称为若干 S 。S 的真正量纲是秒。$1\mathrm{S}=10^{-13}\mathrm{s}$。如果所有分子集团的比容 α 都一样，S 就比例于 m。但生物大分子和细胞器等恰恰不是这样。因此，3.5.3 小节中提到的 23S rRNA 确实比 16S rRNA 分子量大，但并非成简单比例。

凝胶电泳的思想很简单。在铺平的凝胶表面上，梳出若干规整平行的小槽。小槽一头的"井"中放置要分离的混合液体，其中一个"井"里是分子量分布已知的标准混合体。加上电场后，液体中的大小分子集团沿小槽向另一端扩散，轻者快，重者慢 (运动速度与质量的对数成反比)，隔一定时间后就分成许多条纹。与标准样品对比，可知每个条纹对应的分子量。

1975 年 E. M. Southern 把"跑"完的凝胶板上的 DNA 先变性成单链，再覆以硝化纤维素薄膜，上面盖上若干层试纸。水分往试纸扩散，把

凝胶条纹转移到硝化纤维素薄膜上。再同放射性 ^{32}P 标记的已知的互补 DNA 杂交，就可以把特定的 DNA 片段鉴定和分离出来。这套手续发展成强有力的实验方法，称为 DNA 印迹法 (Southern blotting)。

1977 年有人把 DNA 印迹法推广到不如 DNA 稳定的 RNA，称为 RNA 印迹法 (Northern blotting)。后来又推广到蛋白质，称为蛋白质印迹法 (Western blotting)。Northern 和 Western 都不是人的名字。

用聚丙烯酰胺凝胶电泳技术，每次可分辨几十种蛋白质。为了提高分辨率，1975 年发明了二维聚丙烯酰胺凝胶电泳 (2D-PAGE)。一个蛋白质由于含有各种带电基团，在溶液中表现出电荷。在特定的酸碱度 (pH 值) 下呈中性，这个 pH 值称为该蛋白质的等电点 (isoelectric point，简称 Ip)。先把蛋白质混合物在恒定的 pH 梯度下跑一次凝胶，不同的蛋白质按等电点分布到一条线上，再在垂直方向用老办法做电泳，把等电点相同的蛋白质按分子量分开，得到分布在二维平面中的斑点。这样一次实验，可以分辨上千种蛋白质。国际互联网上有大量二维凝胶电泳的文字数据和斑点图象，帮助实验工作者辨认蛋白质，请参看 SWISS-2DPAGE [R-419] 等数据库和 Flicker [R-775] 等网上服务。

3.6.5 DNA 测序方法

可以毫不夸张地说，生物信息学的迅速进步，受到 DNA 自动测序技术的猛烈推动。从原理看，有两种测序方法。一是令聚合过程停止在特定的字母 (核苷酸) 上；二是把聚合到一定长度的 DNA 在特定字母处“咬断”。

终止聚合过程的双脱氧法，即 Sanger 方法，利用 3.4.3 小节中讲过的双脱氧的 ddNTP，使聚合过程停止在一定的字母上，因此至少要用四组测序反应来测定不同的碱基。

化学降解法，或称 Maxam-Gilbert 法，利用一些针对 DNA 长链上具体核苷酸的特异性反应，例如 pH 值 8.0 的二甲基硫 (dimethyl sulfate) 专门使链中的鸟苷酸甲基化。再用 90°C 的热哌啶 (piperidine) 处理，即可在已甲基化的 G 处切断。

两种方法成功实施后，都得到长短不等的 DNA 片段的混合物，要进一步分离。假定每个被测 DNA 片段长 500 碱基对，就要求 1/500 的分辨

率。这大致是目前能保证做到的精度。因此，任何长 DNA 链都必须分割成大量有一定重叠的小片段，克隆增殖，再进行测序，然后再把测序结果拼接起来。这里涉及的算法和程序，本书第 5 章会提到一些。

以上两种测序方法都有克隆、标记、分离、显示等许多技术细节，这里一概略去。L. Alphey 的小书 [R-15] 是一本简明的参考。于军和汪建为贺林主编的 [R-286] 一书撰写的第 5 章，是大规模测序的经验之谈，值得一读。

上述两种测序方法都是在 20 世纪 70 年代中期发展起来的，原始文章发表在同一卷杂志上：

R-116 A. M. Maxam, and W. Gilbert, *Proc. Natl. Acad. Sci.* USA **74** (1977), 560 - 569.

R-117 F. Sanger, S. Nickelen, and A. R. Coulson, *Proc. Natl. Acad. Sci.* USA **74** (1977), 5463 - 5467.

F. Sanger 和 W. Gilbert 两人分享了 1980 年的半个诺贝尔化学奖。

直到不久之前，基因组大规模测序的基本策略还是先完成遗传图谱、物理图谱等基础工作，再把按一定密度确定了遗传标记的 DNA 切割开，用 YAC、cosmid 和 BAC 等载体逐步增殖。然后以不同 BAC 为对象，逐个测定其两端序列，确定 BAC 之间的覆盖连接关系，再把每个 BAC 打碎测序并进行拼接。这样做，可以部分地回避重复序列所导致的组装困难，而且便于许多单位分工合作、平行作业。人类基因组计划就是按此“分而治之”的策略实行。关于 BAC 的情况，可参看 I.M.A.G.E[R-314] 和 ATCC [R-315] 的网页。不久前，J. C. Venter 等人建议了另一种策略：把长 DNA 随机地打碎，用 BAC 增殖并直接测序，大量测序后再用计算机进行拼接组装，参见：

R-118 J. C. Venter, H. O. Smith, and L. Hood, “A new strategy for genome sequencing”, *Nature* **381** (1996) 364 - 366.

R-119 并请参看随即发表的评论：

P. Little, “Genomic analysis”, *Nature* **382** (1996) 408.

Venter 等建议中提到其方法可以避免 YAC 克隆的不稳定性，而 Little 指出，BAC 克隆的适用性并未证明。

自从 1995 年以来，这种“霰弹法”已经成功地用于许多细菌基因组的测定，最近更用以完成了果蝇基因组中常染色质部分的基本测定，但 [R-369] 的作者们承认接近着丝粒和端粒的 DNA 极难用 BAC 克隆，真核生物基因组测序可能永远不“完全”。不过，以 Venter 为首的 Celera [R-798] 公司，坚持霰弹法可用于人类基因组的测定，正在同国际人类基因组计划竞争。

§3.7 进一步阅读书籍

本章的简短叙述，不可能覆盖生物学的全貌。下面列举一些生物化学和分子生物学的参考书，主要是大学和研究生教材。由于整个领域发展甚快，请一定寻求最新版本。

R-120 J. D. Watson, N. H. Hopkins, J. W. Roberts, J. A. Steitz, and A. M. Weiner, *Molecular Biology of the Gene*, 4th ed. , 1987.

R-121 J. D. Watson, M. Gilman, J. Witkowski, and M. Zoller, *Recombinant DNA*, Scientific American Books, distributed by W. H. Freeman, 2nd ed. 1992, xiv + 626.

R-122 B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 3rd ed. 1995, xliii + 1361.

R-123 B. Lewin, *Gene VI*, Cambridge University Press, 1997.

R-124 朱玉贤、李毅，《现代分子生物学》，高等教育出版社，1997、1998。

R-125 B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Molecular Biology: An Introduction to the Molecular Biology of the Cell*, Garland Publishing Co. 1998, xxii + 740.

对于从数理科学转而关心生物信息学的读者，我们再开列几本较为通俗的书籍。

R-126 Erwin Schrödinger, *What is Life? The Physical Aspect of the Living Cell*, 1944, 1945, 1948, 1951, 1955, 1962, and *Mind and Matter*, 1958, 1959, Combined 1967, 1969, 1974, 1977, 1979, 1980, 1983, 1985, 1986,

Cambridge University Press.

R-127 M. V. Volkenstein, *Physics and Biology*, Academic Press, 1982, viii + 165.

R-128 Freeman Dyson, *Origin of Life*, Cambridge University Press, 1985.

R-129 Michael P. Murphy, and Luke A. J. O'Neill, eds. *What is Life? The Next Fifty Years*, Cambridge University Press, 1995, 1997.

R-130 郝柏林、刘寄星主编，《理论物理和生命科学》，上海科学技术出版社，1997, 1999 .

第 4 章　生物信息数据库

数据库是一切生物信息学工作的出发点。大量数据库集中在一些国际或国家的生物信息中心。这些中心一般还提供数据库检索服务、检索工具和各种免费软件。因此，本章 §4.1 节分别介绍一批较为重要的国际和国内的生物信息中心和网点。由于历史原因，许多数据库或软件使用某种特定的数据格式。用户至少应当知道这些格式的名字和实现格式转换的一些工具。因此，§4.2 节将讨论常见的核酸和蛋白质序列格式。在列举一批重要生物数据库之前，还要介绍几种数据库检索工具。这是 §4.3 节的内容。

我们在本书“前言”中已经指出，这里要再次强调，国际生物信息资源和数据库的免费自由使用，是以从事非营利的教育和科学研究为前提的。如果有个别人不与原始数据拥有者协商而利用学术性的免费数据库从事商业活动，就可能在将来妨碍我国整个教育和科学界使用这些资源。这一点要提请读者特别注意。

§4.1　重要生物信息中心简介

以下分国外和国内两部分介绍。所谓“重要”，当然是相对而言。有些学校或研究所的网页包含某方面的有益信息，也就同真正的“中心”一样列入名单。在这样的前提下，疏漏和偶然入选都在所难免。

4.1.1　国外生物信息中心

最重要的几个国际性中心排在前面，以后基本上按国家或地区分组。

R-131 **EBI**，欧洲生物信息学研究所 (European Bioinformatics Institute)，1994 年建立于英国剑桥。它的前身是位于德国海德堡的欧洲分子生物学实验室 [R-133] 的信息服务部门。EBI 接受了原来 EMBL 数据库的管理和维护，并且是欧洲分子生物学网 (EMBnet) [R-132] 的一

个特别节点。EBI 开展了多方面的生物信息服务和研究。网址：

http://www.ebi.ac.uk/ (EBI 主网页，可链接到其他项目)

http://www2.ebi.ac.uk/ (各种数据库和分析工具)

http://www3.ebi.ac.uk/ (EBI 的公众服务网页)

ftp://ftp.ebi.ac.uk

gopher://gopher.ebi.ac.uk

R-132 **EMBnet**，欧洲分子生物学信息网，建立于 1988 年，作为一个国际组织，它在荷兰注册。网址：

http://www.embnet.org/

从 1996 年开始，EMBnet 把成员国范围扩大到欧洲以外，中国在同年加入。EMBnet 的中国节点设在北京大学生物信息中心 PKUCBI [R-166]。目前，EMBnet 有 29 个成员国 (每国一个节点) 和 10 个特别节点。表 4.1 列举 EMBnet 部分节点的网址和所在单位。

R-133 **EMBL**，欧洲分子生物学实验室 (European Molecular Biology Laboratory)，其主实验室设在德国海德堡。除了实验研究，它还提供多种生物计算和数据库服务，以及序列分析方面的服务。详情请参看网址：

http://www.embl-heidelberg.de/Services

http://www.embl-heidelberg.de/~seqanal

EMBL 还在德国汉堡、法国 Grenoble、英国 Hinxton (即 EBI [R-131]) 和意大利 Monterofredo 设有分部，请参看网址：

http://www.embl-hamberg.de/

http://www.embl-grenoble.fr/

R-134 **NCBI**，美国国家生物技术信息中心 (National Center for Biotechnology Information)。网址：

http://ncbi.nlm.nih.gov/

它的前身是美国国家卫生署 (National Institute of Health，简称 NIH) 所属的一个研究所的计算生物学研究室，1988 年独立为 NCBI，形式上属于国家医学图书馆 (National Library of Medicine，简称 NLM)。NCBI 管理着包括 GenBank(详见 [R-212]) 在内的一批数据库，如 UniGene [R-308]、dbSNP [R-310]、COG [R-496]、LocusLink [R-202]、OMIM [R-335] 和 MMDB [R-463] 等。它提供 Entrez(详

表 4.1 EMBnet 成员单位和网址

节点名称	URL	单位
阿根廷 IBBM	sol.biol.unlp.edu.ar	生化和分子生物学研究所
奥地利 VUCC	www.at.embnet.org	维也纳大学计算中心
澳大利亚 ANGIS	www.au.embnet.org	国家基因信息系统
加拿大 CBR-RBCN	www.cbr.nrc.ca	国家科研委海洋生物研究所
比利时 BEN	www.be.embnet.org	
中国 CBI	www.cn.embnet.org	北京大学生物信息中心 [R-166]
古巴 CIGB	www.cu.embnet.org	基因工程与生物技术中心
丹麦 BioBase	www.dk.embnet.org	丹麦生物技术数据库
法国	www.fr.embnet.org	INFOBIOGEN 中心 [R-148]
芬兰 CSC	www.fi.embnet.org	国家科学计算中心
德国 GeniusNet	www.de.embnet.org	德国癌症研究中心
希腊 IMBB	www.imbb.forth.gr	分子生物学与生物技术研究所
匈牙利 HEN	www.hu.embnet.org	
印度 CDFD	www.in.embnet.org	DNA 指纹和预测中心
爱尔兰 INCBI	www.ie.embnet.org	国家生物信息中心
以色列 INN	www.il.embnet.org	魏茨曼科学研究所 [R-164]
意大利 CNR	www.it.embnet.org	Bari 研究园区
挪威 BiO	www.no.embnet.org	奥斯陆生物技术中心
波兰 IBB	www.pl.embnet.org	生物化学与生物物理研究所
葡萄牙 PEN	www.pt.embnet.org	
俄国 GeneBee	www.ru.embnet.org	莫斯科大学生物物化研究所
斯洛伐克	www.sk.embnet.org	科学院分子生物学研究所
南非 SANBI	www.za.embnet.org	国家生物信息研究所 [R-154]
西班牙 CNB	www.es.embnet.org	国家生物技术中心
瑞典 LCB	www.se.embnet.org	林奈生物信息中心 [R-144]
瑞士 SIB	www.ch.embnet.org	瑞士生物信息研究所 [R-141]
荷兰 CMBI	www.nl.embnet.org	分子和生物分子信息中心 [R-147]
土耳其 RIGEB	www.tr.embnet.org	Tubitak-Marmara 研究中心
英国 HGMP	www.uk.embnet.org	人类基因图谱资源中心 [R-140]
EBI	www.ebi.ac.uk	欧洲生物信息研究所 [R-131]
ETI	www.eti.uva.nl	分类学专家鉴定中心 [R-608]
ICGEB [R-152]	www.icgeb.trieste.it	国际遗传工程与生物技术中心
UMBER	www.bioinf.man.ac.uk	曼彻斯特大学
MIPS [R-139]	www.mips.biochem.mpg.de	马普学会生物化学研究所
Pharmacia	www.pnu.com	Pharmacia & Upjohn
Roche	www.roche.com	F. Hoffmann - La Roche
Sanger 中心	www.sanger.ac.uk	Sanger 中心 [R-299]

见 [R-199]) 数据库检索工具、 BLAST(详见 [R-631]) 数据库序列搜索等服务。关于 NCBI 数据库和软件资源的最近描述可参看:

D. L. Wheeler 等 8 位作者、 *Nucleic Acids Res.* **28** (2000) 10 - 14.

R-135 **NCGR**, 美国国家基因组资源中心 (National Center for Genome Resources)。此中心名称中虽有"国家"字样, 实际上是一个非营利的非政府机构, 主要由国家科学基金会 (NSF)、美国农业部和卡内基研究会等公私单位支持。 NCGR 的重要项目包括疫霉属基因预研究计划 PGI [R-357]、拟南芥信息资源 TAIR [R-395]、 GSDB [R 214] 数据库、 ISYS [R-855] 集成软件界面等。网址:

`http://www.ncgr.org/`

`http://seqsim.ncgr.org/`

它设有专门的服务器, 运行 BLAST [R-631]、Smith - Waterman [R 623] 等算法程序, 学术界可自由提交序列进行数据库搜索和联配。

R-136 **HHMI** 是 Howard Hughes 医学研究所的简称。这是一个基本上没有自己的实验室的特殊机构。它为杰出的生物医学工作者提供定期的高强度资助, 促进其在本单位的研究。 MMHI 主要支持细胞生物学、遗传学、免疫学、神经科学和结构生物学五个领域的研究, 目前有 300 多位成员。它虽然不是一个信息中心, 但通过它的网页可以进入一批最活跃的学者的网址, 迅速了解前沿研究情况。网址:

`http://www.hhmi.org/`

R-137 **NIG**, 日本国立遗传学研究所 (National Institute of Genetics), 维护和管理着日本 DNA 数据库 DDBJ, 详见 [R-213]。其信息服务始于 1984 年, 1987 年 7 月 1 日发行 DDBJ 第一版。网址:

`http://www.ddbj.nig.ac.jp /`

R-138 **JIPID**, 日本国际蛋白质信息数据库 (Japan International Protein Information Database), 是 PIR [R-404] 库的三个协作单位之一。

R-139 **MIPS**, 慕尼黑蛋白质序列信息中心 (Munich Information Center for Protein Sequences), 同时也是德国环境与健康研究中心 (GSF)、以及国际蛋白质信息资源 PIR [R-404] 三个协作单位之一。它设在马普学会的生物化学研究所。网址:

`http://www.mips.biochem.mpg.de/`

R-140 **HGMP**，英国医学研究委员会 (Medical Research Council) 所属人类基因组图谱资源中心 (Human Genome Mapping Project Resource Center)，现为 EMBnet 英国国家节点。它所维护的 GenomeWeb [R-614] 是内容最丰富、更新最及时的网上生物信息目录之一。网址：

`http://www.hgmp.mrc.ac.uk/`

R-141 **SIB**，瑞士生物信息研究所 (Swiss Institute of Bioinformatics)，EMBnet 的瑞士节点。网址：

`http://www.isb-sib.ch/`

这是以 SWISS-PROT [R-401]、TrEMBL [R-402]、PROSITE [R-406]、ENZYME [R-415]、SWISS-2DPAGE [R-419]、CD40LBASE [R-522]、SWISS-3DIMAGE [R-488] 等各种与蛋白质有关的数据库著称的信息中心。1993 年由日内瓦大学及其附属医院建立的以蛋白质为重点的 ExPASy (Expert Protein Analysis System) 分子生物学服务器，现在是 SIB 的蛋白质组学 (proteomics) 服务器 。

R-142 **ExPASy** 服务器的网址值得单独列出：

`http://www.expasy.ch/`

`ftp://ftp.expasy.ch`

ExPASy 服务器的中国镜象点设在北京大学生物信息中心，可通过后者的网页 [R-166] 进入，或直接访问：

`http://expasy.pku.edu.cn/`

R-143 **ISREC**，瑞士实验癌症研究所 (Swiss Institute for Experimental Cancer Research) 的生物信息组，是 SIB [R-141] 的成员。它的特色是扩展的蛋白质 Profile 计划，即对 PROSITE [R-406] 的扩充和 Prosite-Scan [R-407] 服务器。网址：

`http://www.isrec.isb-sib.ch/index.html`

R-144 **BMC**，瑞典 Uppsala 生物医学中心。网址：

`http://www.bmc.uu.se/`

所属的林奈生物信息学中心 (Linnaeus Center for Bioinformatics) 是 EMBnet [R-132] 瑞典节点所在地：

`http://www.linaeus.bmc.uu.se/`

R-145 瑞典卡若琳斯卡医学院 (Karolinska Institute，简称 KI，即负责评选诺贝尔生理学或医学奖的那个单位) 和卡若琳斯卡医院 (Karolinska

Hospital)，在 1997 年建立了基因组研究中心 (Center for Genomics Research，简称 CGR)。它的生物信息组是 KISAC(Karolinska Institute Sequence And Computer)。KISAC 除了为 KI 和全瑞典的生物医学研究服务，还维护着 HGBASE [R-312] 数据库和 Belvu [R-662]、Dotter [R-749]、Blixem [R-647]、MSPcrunch [R-648] 等生物计算程序。网址：

http://www.cgi.ki.se/

R-146 **BioBase**，丹麦生物技术信息中心。网址：

http://biobase.dk

此网页包含丹麦人类基因组研究中心的入口。这里有多种蛋白质二维凝胶图象数据。网址：

http://biobase.dk/cgi-bin/celis/

R-147 **CMBI**，荷兰分子和分子生物学信息中心 (Centre for Molecular and Biomolecular Information)。网址：

http://www.cmbi.kun.nl/

它从 1999 年 11 月起代替了原有的 CAOS/CAMM 信息服务。

R-148 **INFOBIOGEN**，这是法国国民教育、研究和技术部于 1999 年 6 月在原 INFOBIOGEN(建于 1995 年) 基础上成立的法国国家生物信息中心，也是 EMBnet [R-132] 法国节点所在地。它的网页是法文的。网址：

http://www.infobiogen.fr/

它所提供的生物数据库总目录 **DBcat** [R-207] 很值得参考：

http://www.infobiogen.fr/services/dbcat/GEN/

为了解法国的情况，还可以参看巴斯德 (Pasteur) 研究所 [R-149] 和里昂大学生物信息中心 [R-150] 的英文网页。

R-149 巴斯德研究所的网页有相当丰富的内容和通向许多重要数据库的链接。网址：

http://www.pasteur.fr/

它的部分内容的中国镜象点设在广州中山大学 [R-176]。

R-150 **PBIL**，里昂生物信息中心 (Pole Bio-Informatque Lyonnais)。由法国里昂大学生物计量与演化实验室和蛋白质生物学与化学研究所在 1998 年联合建立。这里维护着一批与细菌有关的数据库，如 EMGLib

[R-345]、NRSub [R-350]、HOBACGEN [R-421] 等。网址：

http://pbil.univ-lyon1.fr/

PBIL 的特点在于分子生物学与生态学结合。

R-151 **LGT**，俄国理论遗传学实验室。俄国新西伯利亚细胞和遗传学研究所与新西伯利亚大学联合设立的这个实验室，以基因调控区的研究为特色。在俄国基础研究基金和人类基因组计划支持下，他们建立了一个名为 GeneExpress [R-720] 的集成系统，其中包括 TRRD [R-221]、SELEX_DB [R-241]、ACTIVITY [R-279] 等一批数据库和若干检索、显示工具。网址：

http://srs5.bionet.nsu.ru/

R-152 **ICGEB**，国际遗传工程与生物技术中心 (International Centre for Genetic Engineering and Biotechnology)，由联合国工业与发展组织倡议建立，目前有 43 个成员国。中国是正式成员国，联系人为科技部所属中国生物技术发展中心的赵爱民：

E-mail: zhaoaim@public.east.cn.net

ICGEB 有两个园区。意大利的里亚斯特园区，网址：

http://icgeb.trieste.it/

他们还维护着 SBASE 数据库，详见 [R-473]。

印度新德里园区，网址：

http://icgeb.res.in/

R-153 **APBionet**，亚太生物信息网，目前仍在筹建中。网址：

http://www.apbionet.org/

其中国节点在北京大学生物信息中心 [R-166]。

R-154 **SANBI**，南非国家生物信息研究所 (South African National Bioinformatics Institute)，成立于 1996 年。网址：

http://ziggy.sanbi.ac.za/services/

R-155 **BIMAS**，美国国家卫生署 NIH [R-134] 所属的信息技术中心 (CIT) 下面的生物信息学和分子分析部 (Bioinformatics and Molecular Analysis Section)。这里有一批可以在其网页上运行的程序，如启动子扫描程序 Promoter Scan [R-727]、SignalScan [R-728]、序列格式变换程序 ReadSeq[R-699]，以及白细胞 HLA 肽链结合位点预测程序 [R-747] 等。网址：

http://bimas.dcrt.nih.gov/molbio/

R-156 **TIGR**，美国基因组研究所 (The Institute for Genome Research)，是一个非营利性的基因组研究机构，研究从病毒、细菌到人类的基因组以及基因产物的结构、功能和比较。网址：

http://www.tigr.org/

它维护着 TIGR 基因组数据库 TDB [R-215]，其特点是拥有大量 EST 序列。参看：

http://www.tigr.org/tdb/

R-157 **WI**，Whitehead 生物医学研究所 (Whitehead Institute for Biomedical Research，简称 WI)，是 1982 年建立的一个非营利的、独立的基础研究和教学机构。它在肿瘤和艾滋病、发育生物学、结构生物学、传染病和遗传学等方面有开创性的研究项目。网址：

http://www.wi.mit.edu/

R-158 **WICGR**，是 WI 研究所与麻省理工学院共同建立的基因组研究中心 (WI/MIT Center for Genome Research)，它是国际上重要的基因组测序中心之一。它维护着自己的人类 SNP 数据库、家鼠辐射杂交图谱数据库等。网址：

http://www-genome.wi.mit.edu/

关于 WICGR 在人类基因组测序方面的进展，请看：

http://www-seq.wi.mit.edu/

R-159 **CSHL**，美国冷泉港实验室 (Cold Spring Habor Laboratory)。这个曾多年以诺贝尔奖获得者 J. D. Watson 为主任的研究所，是分子生物学的国际领先单位之一。网址：

http://clio.cshl.org/

ftp://ftp.cshl.org/

它的不断更新的网页有关于会议的报道和教学内容，宜经常访问：

http://nucleus.cshl.org/meetings/

CSHL 自 1933 年开始出版的《冷泉港定量生物学讨论会》文集 (Cold Spring Harbor Symposia on Quantitative Biology)，每年一卷，除 1943 - 1945 年外，从未中断，许多分子生物学发展史的里程碑都记录在其浩繁卷帙之中。

R-160 **CompBio**，美国圣约翰大学的计算生物学组。网址：

http://www.cs.jhu.edu/labs/compbio/

此网页有通向许多重要数据库的链接和多种识别基因的软件，例如 Glimmer [R-716]。

R-161 **IUBio**，美国印第安那大学生物系生物信息学中心，维护着内容丰富的生物计算软件档案 [R-611]、果蝇数据库 FlyBase [R-371]、真核生物基因信息库 euGenes [R-340] 等重要资源。网址：

http://sunflower.bio.indiana.edu/

R-162 **SMI**，美国斯坦福大学医学信息学 (Stanford Medical Informatics) 实验室的 Helix 生物信息学组的网页有一些软件描述、出版物电子版和太平洋生物计算研讨会 [R-825] 的电子文集。网址：

http://www-smi.stanford.edu/projects/helix/

R-163 **BNL**，美国布鲁克海文国家实验室 (Brookhaven National Laboratory) 曾是蛋白质结构数据库 PDB [R-441] 的创始者和维护者。PDB 库交给 RCSB [R-442] 管理以后，这里还有大量的生物研究信息。网址：

http://genome1.bio.bnl.gov/

R-164 以色列魏兹曼科学研究所 (Weizmann Institute of Science) 是一个从事研究并培养研究生的机构。它设有若干系、所和研究中心。例如，生物系的分子遗传学研究中心 (The Leo and Julia Forchheimer Center for Molecular Genetics) 参与国际人类基因组计划。网址：

http://www.weizmann.ac.il/

魏兹曼科学研究所的生物信息学网页也可一顾：

http://bioinformatics.weizmann.ac.il/

R-165 **CBS**，丹麦技术大学生物序列分析中心 (Center for Biological Sequence analysis)。它维护着 O-GlycBase [R-485] 数据库和 PhosphoBase [R-431] 数据库。网址：

http://www.cbs.dtu.dk/

4.1.2 国内的生物信息网点

国内生物信息学工作起步较晚，目前北京大学生物信息中心建立的数据库和服务项目最多，国际联系较为广泛，并且组织过多次国际、国内

的培训活动和会议。许多学校和研究单位近几年开始草创生物信息研究和服务。下面是一些不完全的概况。

R-166 **CBI 或 PKUCBI**，北京大学生物信息中心，成立于 1997 年 3 月，它是 EMBnet [R-132] 的中国节点，也是亚太生物信息网 APBionet [R-153] 的中国节点。他们的网页很值得访问：

`http://www.cbi.pku.edu.cn/`

他们的 ftp 服务器可以通过网页访问，也可用 ftp 命令直取：

`ftp://ftp.cbi.pku.edu.cn/`

北京大学生物信息中心的电子邮件联系地址是：

`mailto: office@cbi.pku.edu.cn`

从 PKUCBI 可以立即进入 EMBnet 的主页和若干重要生物数据库的镜象点。EMBnet 上有 200 多种可以自由访问的数据库，北京大学生物信息中心目前已经建立了 70 多种分子生物信息镜象系统和数据库，有些库已经做到每日更新。许多数据库都可以通过检索工具 SRS[R-203] 查询。

R-167 **PKUBIOS** 服务器 (Peking University Bioinformatics Server)，设在北京大学化学系物理化学研究所。可从物化所的网址进入：

`http://www.ipc.pku.edu.cn/mirror/mirror.html/`

这里有 PDB [R-441]、SCOP [R-454] 等与蛋白质有关数据库的镜象，但更新速度不及 PKUCBI [R-166]。

R-168 **AMMSnic** 是中国军事医学科学院情报研究所网络信息中心的英文缩写。这里有通向许多国际生物信息数据库和生物医学资源系统的链接，其中一部分在我国已有镜象点，可参看 [R-166]、[R-167]、[R-170]、[R-169] 等。这里有一批医学、药学数据库和国外军事医学研究单位的 URL，本手册未提及。网址：

`http://www.bmi.ac.cn/bio/`

R-169 **CMBI/BJMU**，北京大学医学部生物信息网页：

`http://cmbi.bjmu.edu.cn/`

这里有较多关于流行病学和心血管疾病的中文信息。以色列魏茨曼科学研究所的 GeneCards [R-418] 的中国镜象点在此。

R-170 中国科学院微生物研究所的网页：

`http://www.im.ac.cn/`

这里有一个靠关键字检索的日本 DDBJ 数据库 [R-213] 的镜象点，但没有 DDBJ 库的其他检索工具和服务。中国微生物网也设在这里：

http://micronet.im.ac.cn/

此外，筹建中的亚太先进网 (Asia Pacific Advanced Network，简称 APAN) 有一个建立生物信息镜象点的计划。其中国节点设在微生物研究所。网址：

http://bio-mirror.cn.apan.net

目前该网页上只有通向几个主要生物信息数据库的链接。

R-171 中国科学院上海生命科学研究院的网页，有通向各研究所的链接。网址：

http://www.sibs.ac.cn/

R-172 **BioSino** 是中国科学院上海生命科学研究院生物信息中心的网站，它目前除维护我国的核酸序列公共数据库 [R-216] 外，还提供包括各种链接的生物学导航信息：

http://www.biosino.org/

R-173 中国科学院上海生物化学研究所网页的生物信息学部分，有一批指向国际生物数据库和软件服务的链接。网址：

http://dna.sibc.ac.cn/bio/

R-174 中国科学院遗传研究所人类基因组中心，是我国承担的国际人类基因组计划 1% 测序任务的主要测序中心。网址：

http://hgc.igtp.ac.cn/

http://www.genomics.org.cn/

R-175 中国科学院国家基因组中心，成立于 1992 年。其主要任务是承担中国水稻基因组计划。他们的网页简要介绍计划进展情况。从 ftp 服务器可以下载水稻基因组物理图谱数据。网址：

http://www.ncgr.ac.cn/

ftp://ftp.ncgr.ac.cn

R-176 广州中山大学生物信息中心，与法国巴斯德研究所 [R-149] 合作，于 1999 年 9 月开通了“法国巴斯德亚洲信息网”。网址：

http://genome.zsu.edu.cn/

主要内容是巴斯德研究所生物信息网页的镜象点，如枯草芽孢杆菌的 Subtilist 和大肠杆菌的 Collibri 等。

R-177 中山医科大学的网页有到 PubMed[R-600] 的链接，实际检索仍在 NCBI 进行。网址：

http://www.gzsums.edu.cn/

ftp://ftp.gzsums.edu.cn

§4.2 数据库和序列的格式

由于历史原因，各种生物数据库采用了不同的信息格式，许多生物计算软件也要求特定的核酸和蛋白质序列输入格式。这当然是不方便的。因此，我们专门在本节介绍常见的数据库文件和生物序列格式。介绍分成三部分：数据库文件格式、序列格式和多序列联配所涉及的格式。前两者并没有明确界限，许多程序都会从数据库文件中提取序列。多序列格式花样较多，有些只有历史意义或很少用到的格式，只点出名字。万一遇到，可以临时查询。

有一批软件专门处理格式之间的转换，例如 D. Gibert 编写的免费程序 ReadSeq [R-699]，可以处理 18 种格式。GCG [R-792] 程序包中也有几种格式转换模块，把其他格式变成它所要求的 GCG 格式 [R-184]。

4.2.1 数据库格式

多数生物数据库由文字说明和序列两大部分组成，两者都有固定格式，以便计算机读取。例如，GenBank 中大肠杆菌全基因组条目，文件总长接近 15 万行，其中注释占 71 719 行，序列占 77 322 行。各个数据库的具体格式，又有所不同，大致分成 EMBL 和 GenBank 两种风格。

R-178 EMBL 格式 . 欧洲分子生物学 EMBL 数据库的每个条目是一份纯文本文件。每一行最前面是由两个大写字母组成的识别标志，常见的识别标志列举在表 4.2 中。识别标志“特性表” FT 包含一批关键字，它们的定义已经与 GenBank 和 DDBJ 统一，在文件 [R-210] 中有详细说明。下面介绍 GenBank 格式 [R-179] 时再在表 4.3 中列举这些关键字。

欧洲国家的许多数据库如 SWISS-PROT [R-401]、ENZYME [R-415]、TRANSFAC [R-219] 等，都采用与 EMBL 一致的格式。

表 4.2 EMBL 和 GenBank 数据库的行识别标志

EMBL 识别标志	GenBank 识别字	意义
ID	LOCUS	标识字符串及短描述字
AC	ACCESSION	唯一的提取号
DE	DEFINITION	简单的描述
OS	SOURCE	来源生物体
OC	ORGANISM	及其分类谱系
DT		建立日期
KW	KEYWORDS	关键字
RN	REFERENCE	引文编号
RA	AUTHORS	引文作者
RT	TITLE	引文题目
RL	JOURNAL	引文出处
RX		交叉引用
DR	COMMENTS	对其他数据库的引用
	MEDLINE	引文的 MEDLINE 号
XX		为阅读清晰而加的空行
CC	COMMENT	评注
NI	VERSION	可更新的序列版本号 (AC 不能变)
FH	FEATURES	特性表头
FT	FEATURES	特性表
SQ		EMBL 序列开始，后随长度、字母数
	BASE COUNT	GenBank 碱基数目
	ORIGIN	GenBank 序列开始标志，该行空
//	//	序列结束标志，空行

R-179 **GenBank 格式**。每个条目是一份纯文本文件，每行左端或为空格或为识别字。识别字均为完整英文字，不用缩写。为了同 EMBL 对照，一并列在表 4.2 中。请注意，从 1999 年 12 月 15 日的第 115 版开始，GenBank 取消了 Version 关键字下的版本号 NID 和 PID，以及特性表中 /db_xref 对它们的引用，只保留了 GI 版本号及其 /db_xref 引用。一个 GenBank 条目，从 LOCUS 行到 ORIGIN 行是注释部分，注释按识别字分成若干段，从 FEATURES 开始注释的核心部分。它使用一大批与 EMBL 和 DDBJ 数据库统一的关键字，表 4.3 中列举了若干重要的关键字。

表 4.3 GenBank 注释中的关键字

关键字	意义
3′UTR	3′ 非翻译区
5′UTR	5′ 非翻译区
-10_signal	-10 信号
-35_signal	-35 信号
CAAT_signal	CAAT 信号
CDS	编码序列，含终止密码子
enhancer	增强子
exon	外显子
GC_signal	GC 信号
gene	已命名的基因序列
intron	内含子
LTR	长终端重复序列
mat_peptide	翻译后被修饰的序列，不含终止密码子
mis_binding	错结合点
misc_feature	其他性状
misc_RNA	其他 RNA
misc_signal	其他信号
modified_base	修饰过的碱基
mRNA	信使 RNA
mutation	突变
rRNA	核糖体 RNA
tRNA	运输 RNA
polyA_signal	多聚 A 信号
polyA_site	多聚 A 位点
prim_transcript	初始转录码
promotor	启动子
protein_bind	蛋白质结合位点
rep_origin	复制起点
repeat_region	重复区
repeat_unit	重复单元
satellite	卫星片段
sig_peptide	信号肽
TATA_signal	TATA 信号
terminator	终端子

表 4.3 中没有列出的还有 allele、attenuator、C_region、conflict、D-loop、D_segment、iDNA、J_segment、misc_difference、misc_recomb、misc_structure、N_region、old_sequence、precursor_RNA、primer_bind、RBS、S_region、scRNA、snRNA、stem_loop、STS、transit_peptide、unsure、V_region、V_segment、variation、virion、3'clip、5'clip 等等，其详细定义在文件 [R-210] 中给出。从 ORIGIN 行之后的下一行开始，为序列本身。每行最左端 1 - 9 格，是该行第一个符号的序号，向右对齐。第 11 - 75 格含 60 个符号 (最后一行可以不足 60)，每 10 个符号用空格隔开，以利人工阅读。然后以单独的 // 行作为结束标志。

EMBL 和 GenBank 数据库的序列本身，都可延续多行，前面没有标志，每行 60 个字母，每 10 个字母加一空格以利阅读。EMBL 库每行右端为该行最后字母的序号，GenBank 库每行左端为该行首字母的序号。

GenBank 的 genomes 子目录下，每个物种的子目录中为使用方便而提供了多种格式的文件。这些格式列举在表 4.4 中。

表 4.4 GenBank 的基因组文件类型

文件名后缀	说明
.ffn	FASTA 格式的核苷酸编码序列
.fna	FASTA 格式的核苷酸序列
.faa	FASTA 格式的氨基酸序列
.gbk	GenBank 格式的纯文本文件
.ptt	蛋白质表
.tab	序列片段组装情况表
.asn	ASN.1 格式，见 [R-180]
.val	ASN.1 格式二进制文件

R-180 **ASN.1** 格式 (Abstract Syntax Notation 1) 是 NCBI [R-134] 所发展的许多程序如显示蛋白质三维立体结构的 Cn3D([R-779]) 所使用的内部格式。它必须由一个名为 writeSeq 的程序产生，被专门的程序阅读，而不直接供最终用户使用，因此这里不予介绍。

R-181 蛋白质信息资源库 PIR [R-404] 采用与国际科学数据库 CODATA 一致的 PIR/CODATA 格式，请参看 PIR 库有关文件。关于 CODATA，可参看其中国委员会的网页：

http://www.cncodata.ac.cn/

4.2.2 序列文件格式

下面介绍几种常见的序列文件格式.

R-182 **FASTA** 格式, 又称 Pearson 格式, Pearson 是 FASTA 的主要作者. 这是比较简单而使用最多的序列格式. 序列文件的第一行是由大于符号 (>) 打头的任意文字说明, 主要为标记序列用. 从第二行开始是序列本身, 只允许使用表 3.2 中的标准核苷酸符号或表 3.3 中的标准的氨基酸的单字母符号. 通常核苷酸符号大小写均可, 而氨基酸一般用大写字母. 有些程序对大小写有明确要求, 使用时须注意. 文件中的每一行都不要超过 80 个字母. 由于 FASTA 格式没有特殊的序列结束标志, 建议最后多留一个空行. 这是因为有些电子邮件系统会自动在信尾加上发信人的地址、电话等, 如无空行隔离, 可能被程序误认为序列. 下面是 FASTA 格式的一条 DNA 序列实例:

```
> Human (lambda) DNA for immunoglublin light chain D86989
aactgtactcacgtgacagttccctgaatcttcatacagattatctcctacccttatag
tgcatgtttcttatgaaggcctccaacatgctagccatttcctactaaactaactcaact
agcatgatgtcaacaacacagtcaatcaatgggatattttgtggggtgctcagatggcag
aatgctcccacatcaataaaact
```

R-183 **Staden** 格式是 Staden 程序包 [R-690] 所使用的形式上最简单的格式. 它不允许有任何注释, 每行有 60 个字母, 使用表 3.2 和表 3.3 所规定的核苷酸或氨基酸的单字母符号.

R-184 **GCG** 格式. 这是商业性的 GCG 软件包 [R-792] 的专用格式. 它前面可有任意行注释, 直到两个相连的圆点 "..". 包含两个.. 在内的最后一行注释是序列名字、长度、日期, 以及一个检查和 (Check sum). 检查和的设置原来是为了检查输出输入错误, 现在反倒是给用户交替使用 GCG 和非 GCG 程序制造困难, 并且妨碍 "手工" 编辑序列数据.

R-185 **Plain/Raw** 格式, 即未作任何修饰的原始纯文本格式.

4.2.3 多序列格式

各种多序列联配程序往往使用不同的输入 / 输出序列格式. 下面以假设来自三个物种的三条相似的 DNA 序列为例, 介绍几种常见的多序列

格式[20]。

R-186 **FASTA** 格式可以用于多序列联配。不含空格时如下：

```
>Sequence 1
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTAAATTGCACTAATGCGACGTATACTAAT
>Sequence 2
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTATGCACTAATGCGACGTATACTAAT
>Sequence 3
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTAACTAATGCGACGTATACTAAT
```

允许插入空格“-”时，同样三条序列为：

```
>Sequence 1
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTAAATTGCACTAATGCGACGTATACTAAT
>Sequence 2
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTA---TGCACTAATGCGACGTATACTAAT
>Sequence 3
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTA------ACTAATGCGACGTATACTAAT
```

R-187 **Phylip** 格式。这是免费的亲缘关系计算程序 Phylip(见 [R-677]) 所要求的输入格式。Phylip 的老版本，如 3.2 或 3.4，在细节上还略有不同。较为近期的例子如：

```
     3   96
Sequence   1 GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
Sequence   2 GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
Sequence   3 GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA

          ACCCCACTCTGTGTTACTTTAAATTGCACTAATGCGACGTATACTAAT
          ACCCCACTCTGTGTTACTTTA---TGCACTAATGCGACGTATACTAAT
          ACCCCACTCTGTGTTACTTTA------ACTAATGCGACGTATACTAAT
```

R-188 **NEXUS** 格式。这是商业性的亲缘关系计算程序 PAUP [R-678] 所要求的输入格式。它有两种选择。一是把每个序列连续地排列，结束之后再排下一个序列，例如：

```
#NEXUS
BEGIN DATA;
```

[20] 此例参考了 [R-17] 一书附录 2。

```
DIMENSIONS NTAX=3 NCHAR=96;
FORMAT MISSING=? GAP=- DATATYPE=DNA;
MATRIX
Sequence 1
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTAAATTGCACTAATGCGACGTATACTAAT
Sequence 2
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTA---TGCACTAATGCGACGTATACTAAT
Sequence 3
GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
ACCCCACTCTGTGTTACTTTA------ACTAATGCGACGTATACTAAT
```

这种格式不便于目视观察联配情况。因此，第二种选择是把联配好的各个序列放在一起，同时切断和移行：

```
#NEXUS
begin data;
      dimensions ntax=3 nchar=96;
      format datatype=dna GAP=: interleave;
      matrix
Sequence 1 GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
Sequence 2 GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA
Sequence 3 GATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA

Sequence 1 ACCCCACTCTGTGTTACTTTAAATTGCACTAATGCGACGTATACTAAT
Sequence 2 ACCCCACTCTGTGTTACTTTA---TGCACTAATGCGACGTATACTAAT
Sequence 3 ACCCCACTCTGTGTTACTTTA------ACTAATGCGACGTATACTAAT
```

R-189 **NBRF 格式**。这是支持 PIR [R-404] 数据库的美国全国生物医学研究基金会 (NBRF) 采用的多序列格式。

```
>DL;Sequence
Sequence 1, 96 bases, 750EDA48 checksum.
 GATATAATCA GTTTATGGGA TCAAAGCCTA AAGCCATGTG TAAAATTAAC
 CCCACTCTGT GTTACTTTAA ATTGCACTAA TGCGACGTAT ACTAAT*
>DL;Sequence
Sequence 2, 93 bases, D557AE5C checksum.
 GATATAATCA GTTTATGGGA TCAAAGCCTA AAGCCATGTG TAAAATTAAC
 CCCACTCTGT GTTACTTTAT GCACTAATGC GACGTATACT AAT*
>DL;Sequence
Sequence 1, 90 bases, B916B9DB checksum.
 GATATAATCA GTTTATGGGA TCAAAGCCTA AAGCCATGTG TAAAATTAAC
 CCCACTCTGT GTTACTTTAA CTAATGCGAC GTATACTAAT*
```

R-190 **Stockhom** 格式，由 HMMER [R-739] 、 Pfam[R-478] 和 Blixem [R-647] 等使用的一种多序列格式，可参看 KISAC [R-145] 的网页.

R-191 **MSF**(Multiple Sequence Format) 多序列格式，是商业性的 GCG 软件包 [R-792] 专用的联配格式.

4.2.4 其他序列格式

最后，简单提一下使用较少的一些序列格式. 除了 SELEX 格式外，ReadSeq 程序都可以处理。

R-192 **Standard** 格式，又称 **IG** 格式，每个数据文件可以包含多条序列。注释行由分号 ";" 开头，可有多行。每个序列第一行给出序列名称和简单信息，随以若干行不空格的符号序列，最后一行符号以数字 1 或 2 作结束符.

R-193 **DNAStrider** 格式用于某些 Macintosh 程序，故从略.

R-194 **SELEX** 是一种旧的多序列格式. Stokholm 格式仍保持与其兼容，如遇到可参看 [R-190] .

R-195 **Fitch** 格式， Fitch 是 20 世纪 70 年代初最早的亲缘树算法的提出者之一。

R-196 **Zuker** 格式，这是只用于输入的一种格式.

R-197 **Olsen** 格式，这也是只用于输入的一种格式.

R-198 **Pretty** 格式，这是只用于 "漂亮输出" 的一种格式.

§4.3 数据库检索工具

当前许多生物数据库都不仅是罗列数据，一般还配有自己的检索工具，可以按关键字查询，并且链接 到有关网址或文献。下面介绍几种常见的数据库检索工具.

4.3.1 Entrez 检索工具

Entrez 是美国国家生物技术信息中心 NCBI [R-134] 所提供的集成检索工具. 最方便的用法是在网络浏览器中访问，也可以下载到本地计算机上运行.

R-199 **Entrez** 网址：

http://www.ncbi.nlm.nih.gov/Entrez/

Entrez 根据用户的询问，在 5 组数据库之间进行交叉检索。这 5 组数据库是：

1. PubMed 文献库 MEDLINE [R-599]。

2. 核酸序列库如 GenBank [R-212]。

3. 蛋白质序列库。

4. 结构数据库如 MMDB [R-463]、PDB[R-441]。

5. 在 Genomes 总名称下面的各种基因图谱库。

用户可以从任何一个数据库开始，用作者名字、序列索取号、基因或蛋白质名称、酶的 EC 编号等各种各样的关键字搜索，借助直接链接和“邻域”得到大量有关记录，也可在检索过程中补充新的关键字，缩小查询范围。直接链接的意义很明确。例如从作者找到文章摘要，其中提到的基因或蛋白质都有通向相应库中条目的超链接，一“点”即到；如果该蛋白质的三维结构已经定出，就会看到通向有关条目的链接，甚至可以调用 Cn3D [R-779] 观看其立体结构的转动。

“邻域”是 Entrez 特有的概念，我们稍加解释。序列邻域是在 BLAST [R-631] 局域联配意义下相近的序列集合。结构邻域是用 VAST [R-464] 比较得到的相似的结构。文献邻域则是根据标题和摘要中重要关键字及其衔接关系确定的“与本文类似的文章”。

以上介绍的是所谓 Web Entrez 的服务方式。用户只要装有标准的浏览器，如 Netscape 或 Internet Explorer，就可以工作。检索系统实际上在 NCBI 运行。如果嫌这种运行方式效率太低，而所在单位已经具备快速互联网，那就可以从 NCBI 下载 Network Entrez 到本地计算机上运行。前面提到的 Cn3D 三维结构显示程序，是 Network Entrez 的组成部分。如果使用 Web Entrez，就要事先从 NCBI 下载，作为浏览器的插件 (plug-in) 装好。

对于网络条件不佳的用户，可以通过 Query 电子邮件享用 Entrez 服务。如果只需要从单个数据库提取某一条目，可以使用 Retrieve 电子邮件服务。

R-200 **Query** 电子邮件服务。电子邮件必须遵从一定格式，请参看说明文件：

`http://www.ncbi.nlm.nih.gov/Web/Search/query.txt`

电子邮件地址：

`mailto: query@ncbi.nlm.nih.gov`

R-201 **retrieve** 电子邮件服务，详情也请参看说明文件：

`http://www.ncbi.nlm.nih.gov/Web/Search/retrieve.txt`

我们提醒读者设置好回信行数，否则文件超过补缺行数时只能取到部分记录。电子邮件地址：

`mailto: retrieve@ncbi.nlm.nih.gov`

R-202 **LocusLink** 是 NCBI 提供的对经过审读的序列及其遗传位点描述信息的查询系统，请参看：

`http://www.ncbi.nlm.nih.gov/LocusLink/`

4.3.2 SRS 检索工具

R-203 **SRS** 序列查询系统 (Sequence Retrieval System) 是欧洲分子生物学网 EMBnet [R-132] 的主要数据库检索工具，可从 EMBnet 的主页进入。它的最初设计见：

T. Etzold, *CABIOS* (*Bioinformatics*) **9** (1993) 49 – 57.

现在的版本是 SRS6，掌握 SRS 的方法是实际运用。中国用户可从北京大学生物信息中心 [R-166] 的 EMBnet 镜象点的主页进入。日本 DDBJ [R-213] 最近也向用户提供 SRS 界面。

4.3.3 DBGET/LinkDB 检索工具

日本京都大学化学研究所建立的 GenomeNet 数据库服务网页，包含 KEGG [R-554] 和 DBGET/DB 两套主要系统。前者注重代谢途径，后者处理数据库检索。

R-204 **GenomeNet** 数据库服务网页，提供国际上重要生物数据库和一些日本学者建立的库的检索和交叉引用。有些数据库目前已做到每日更新。详情见网址：

`http://www.genome.ad.jp/`

R-205 **DBGET** 检索工具的使用说明见：

http://www.genome.ad.jp/dbget/dbget_manual.html

§4.4 数据库目录

从 1994 年开始，《核酸研究》杂志 (*Neucleic Acids Research* [R-6]) 每年第一期是生物数据库专集。1998 、 1999 和 2000 三年该刊分别介绍了 102 、 105 和 114 种数据库。有的篇目讲了不止一种库，有的库每年重复介绍。因此，上面的数字并不反映数据库的确切种数。但这是获取生物数据库最新情况的一个好起点。特别是从 2000 年开始，出版《核酸研究》的牛津大学出版社设立了一个数据库目录网页 [R-6]，可以按照字母或分类查找，并且立即链接到所需要的数据库。这个网页把数据库分成 18 类。我们有所合并，并把数据库目录、农林牧有关数据库、医学数据库和文献目录库单独列出，分成以下 16 类介绍：

1. 数据库目录。

2. 综合数据库，包括 DNA 序列库： EMBL [R-211] 、 GenBank [R-212] 、 DDBJ [R-213] 、 GSDB [R-214] 、 TDB [R-215] 和 UniGene [R-308] 。

3. DNA 序列数据库，主要是与基因结构和认定有关的数据，如密码子使用频度表 [R-217] 、真核生物启动子库 [R-218] 、内含子和外显子库 [R-246] 等。

4. RNA 序列和核糖体数据库。

5. 基因图谱数据库。

6. 人类基因组数据库。

7. 其他物种基因组数据库。

8. 基因表达数据库。

9. 基因突变、病理和免疫数据库。

10. 蛋白质序列数据库。

11. 蛋白结构数据库。

12. 比较基因组学和蛋白质组学数据库。

13. 代谢途径和细胞调控数据库.

14. 与农林牧有关数据库.

15. 医学数据库.

16. 其他数据库.

一个数据库可能跨越两个以上门类. 例如 DNA 转录就涉及许多因子和它们的结合位点, 前者是蛋白质, 后者是 DNA, 很难唯一地归入哪一类. 下面介绍的数据库并不限于《核酸研究》近两年所列者. 每个库名条目下, 尽可能给一简单说明和网址, 以及一两篇较近期的引文.

R-206 **NAR** 网页, 基于《核酸研究》杂志 [R-6]1999 和 2000 两年第一期所介绍的数据库, 列举了通向 224 个数据库的链接. 网址:

`http://www.oup.co.uk/nar/Volume_28/Issue_01/html/gkd115_gml.html`

R-207 **DBcat**, 法国生物信息中心 INFBIOOGEN [R-148] 维护的, 建于 1994 年的生物数据库目录. 最近的描述参见:

C. Discala, X. Benigni, E. Barillot, and G. Vaysseix, *Nucleic Acids Res.* **28** (2000) 8 – 9.

网址:

`http://www.infobiogen.fr/services/dbcat/`

北京大学生物信息中心 [R-166] 有其镜象. 2000 年 5 月底 DBcat 列举了 513 种数据库, 其分类统计见表 4.5.

表 4.5 DBcat 列举的 513 种数据库的分类统计

分类	数据库数目	分类	数据库数目
DNA	87	RNA	30
蛋白质	94	基因组	58
基因图谱	30	蛋白质结构	18
文献	43	其他	153
总计			513

R-208 **LiMB** 是 1988 年开始建立的生物信息数据库目录. 那时全部数据库名单都印在一篇文章里:

J. R. Lawton, M. A. Martinez, and C. Burks, *Nucl. Acids Res.* **17** (1989) 5885 - 5899.

LiMB 已经停止发展，它的历史档案保存在：

gopher://gopher.nih.gov/11/molbio/other/

现在请参考法国生物信息中心 INFOBIOGEN[R-148] 所维护的 DBcat [R-207]。

§4.5 综合数据库

综合性的通用数据库多为大型的一级核酸序列数据库，也包括翻译出的氨基酸序列。目前主要是日本、欧洲和美国三家各自建立和共同维护的国际核酸序列库 INSD。

R-209 **INSD** 国际核酸序列数据库 (International Nucleotide Sequence Databank)，由日本 DDBJ [R-213]、欧洲 EMBL [R-211] 和美国 GenBank [R-212] 三家各自建立和共同维护。这三个数据库的格式大同小异，每天自动交换数据，保持同步更新。1995 年三家统一了注释部分的特性表 (FT 或 FEATURES) 的定义。对使用者较为重要的是以下文件：

R-210 The DDBJ/EMBL/GenBank Feature Table: Definition，1.08 版，1995 年 12 月 1 日。可从 GenBank 下载：

ftp://ncbi.nlm.nih.gov (/genbank/docs/)

下面首先扼要介绍这三个数据库，然后提及其他几个综合核酸数据库。

R-211 **EMBL** 库，欧洲分子生物学实验室的 DNA 和 RNA 序列库，它通过科学文献、专利申请和直接投送获得数据，每日更新，每年四版。较新描述参见：

W. Baker 等 7 位作者， *Nucleic Acids Res.* **28** (2000) 19 - 23.

网址：

http://www.ebi.ac.uk/embl.html

http://www.ebi.ac.uk/ebi_docs/embl_db/embl_db.html

ftp://ftp.ebi.ac.uk (/pub/databases/embl/release)

通过 EMBnet [R-132]， EMBL 数据库在许多国家有每日更新的镜象。这些镜象点的名单参见：

http://www.ebi.ac.uk/embl/Access/other_sites.html

北京大学生物信息中心 [R-166] 也设有镜象，并可通过检索工具 SRS [R-203] 查询。

R-212 **GenBank** 是 NCBI [R-134] 所维护的供公众自由读取的、带注释的 DNA 序列的总数据库，每天更新，每两个月发行一次新版。2000 年 8 月 15 日发布的第 119.0 版，一共收录了 8 214 339 个 DNA 序列，计 9 545 724 824 个碱基对 (bp)。然而，每个 DNA 序列的平均长度只有 1 162 bp。这是因为早期收录的序列都比较短。特别是 EST 序列条目很多，而每条长度甚短。关于 GenBank 数据库的较新描述请参看：

D. A. Benson 等 6 位作者， *Nucleic Acids Res.* **28** (2000) 15 - 18.

网址：

http://www.ncbi.nlm.nih.gov/Web/Genbank/

ftp://ncbi.nlm.nih.gov (/genbank/)

R-213 **DDBJ**，日本核酸数据库 (DNA Data Bank of Japan)，设在国立遗传研究所 [R-137] 的遗传信息中心。它首先是反映日本所产生的 DNA 数据，同时与 GenBank [R-212] 和 EMBL [R-211] 合作，互通有无，同步更新，每年四版。日本 DDBJ 库采用与 GenBank 一致的格式。此库的较新描述参见：

Y. Tateno, S. Miyazaki, M. Ota, H. Sugawara, and T. Gojobori. *Nucleic Acids Res.* **28** (2000) 24 - 26.

网址：

http://www.ddbj.nig.ac.jp/

ftp://ftp.nig.ac.jp (/pub/db)

ftp://monet.genes.nig.ac.jp (/data/)

中国科学院微生物研究所设有靠关键字检索的 DDBJ 库镜象，可经由其网页 [R-170] 访问。

R-214 **GSDB** 是由 NCGR [R-135] 维护的 DNA 序列关系数据库 (Genome Sequence DataBase)，它搜集核酸序列及有关生物和文献信息，其目标是提供一个集成的功能基因组学数据库。从 1998 年底起 GSDB 不

再接受用户直接提交的序列，库中已有序列的所有权也转到 GenBank [R-212]。此后它的序列每天晚上来自 INSD [R-209]。GSDB 集中力量为研究者免费提供检索和分析服务。它提供的工具有：

Maestero — http://www.ncgr.org/gsdb/maestro/

Ad hoc query — http://www.ncgr.org/gsdb/adhoc/

Excerpt — http://www.ncgr.org/gsdb/excerpt/

Sequence Viewer — http://www.ncgr.gsdb.org/gsdb.sv/

另外，还有纯文本读取工具 (Flatfile Retrieval Tool) 等。较新的描述请参见：

C. Harger 等 9 位作者，*Nucleic Acids Res.* **28** (2000) 31 – 32.

网址：

http://www.ncgr.org/gsdb/

ftp://ftp.ncgr.org/

R-215 **TIGR Database**，TIGR [R-156] 研究所是国际上重要的测序中心之一，它有大量正在测定过程中的基因组数据，特别是 EST 序列。这里的人类基因索引 HGI 也值得注意，请参见：

J. Quackenbush 等 5 位作者，*Nucleic Acids Res.* **28** (2000) 141 – 145.

TIGR 还拥有世界上最大的 cDNA 数据库之一，不过访问有限制。

网址：

http://www.tigr.org/tdb/hcd/overview.html

§4.6 DNA 序列和结构数据库

归入这一类的不仅是单纯的 DNA 序列。有些与 DNA 的复制、转录、修复等有密切关系的蛋白质因子，也和 DNA 放在一起，以利于查询。

R-216 **BioSino** 是中国自主开发的核酸序列公共数据库，它将发表我国各基因研究中心提供的核酸序列，并接受我国核酸序列的注册登记。

网址：

http://www.biosino.org/

R-217 **CUTG**，密码子使用频度表。这是由 GenBank [R-212] 中的 DNA 序列统计出来的密码子使用频度表 (Codon Usage Tabulated from GenBank)，按物种和模式生物给出。1999 年 9 月 CUTG 库中共有来自

257 486 个完整的蛋白质编码序列的 8 792 个物种的密码子使用频度表。把蛋白质氨基酸序列倒译为核苷酸序列时，必须参考此表。请参见：

Y. Nakamura, T. Gojobori, and T. Ikemura, *Nucleic Acids Res.* **28** (2000) 292.

网址：

`http://www.dna.affrc.go.jp/~nakamura/CUTG.html`

`http://www.kazusa.or.jp/codon/`

`ftp://ftp.kazusa.or.jp (/codon/current/)`

`ftp://ftp.nig.ac.jp (/pub/db/codon/current/)`

`ftp://ftp.ebi.ac.uk (/pub/databases/cutg/)`

`ftp://ftp.dna.affrc.go.jp (/pub/codon)`

`http://www.dna.affrc.go.jp/~nakamura/CUTG.html`

各主要生物信息中心均有镜象。北京大学生物信息中心 [R-166] 也有其镜象。如果关心人类基因组中的密码子使用频度，可访问以色列魏兹曼科学研究所生物信息组的数据库：

`http://bioinformatics.weizmann.ac.il/databases/codon`

R 218 **EPD**，真核生物启动子数据库 (Eukaryotic Promotor Database)。它搜集所有转录起点已经由实验确定的第 II 类 DNA 聚合酶的启动子序列，包含对 EMBL [R-211] 核酸序列数据库、SWISS-PROT [R 401] 蛋白质库、TRANSFAC [R-219] 转录因子库，以及文献的交叉引用。这里还有一个名为 TRADAT 的工具和查询界面。原始库由瑞士实验癌症研究所 ISREC [R-143] 的 P. Bucher 维护，描述见：

R. C. Perier, V. Praz, T. Junier, C. Bonnard, and P. Bucher, *Nucleic Acids Res.* **28** (2000) 302 – 303.

网址：

`http://www.epd.isb-sib.ch/`

`ftp://ftp.ebi.ac.uk (/pub/databases/epd/)`

`ftp://ftp.infobiogen.fr (/pub/db/epd/)`

R-219 **TRANSFAC**，真核生物基因表达调控因子的数据库。它是由 Edgar Wingender 一个人在 1988 年搜集当时仅有的几个转录因子开始建立

的[21]，现在发展成包括从酵母到人的最重要的调控因子库。事实上，TRRD [R-221] 和 COMPEL [R-227] 等库也是由 TRANSFAC 衍生出来的。从 1999 年第 3.5 版起，商业性用户需事先取得许可。详见：

E. Wingender 等 10 位作者， *Nucleic Acids Res.* **28** (2000) 316 - 319.

网址：

`http://transfac.gbf.de/TRANSFAC/`

`http://www.biobase.de/` (商业性用户请访问此网页)

北京大学的镜象在：

`http://www.cbi.pku.edu.cn/gbf/`

R-220 **IMD** 是从 TRANSFAC 等数据库计算出来的为寻找结合位点用的权重矩阵的集合。MATRIX SEARCH 等程序需访问此库。关于 IMD 和 MATRIX SEARCH 均请参见：

Q. K. Chen, G. Z. Hertz, and G. D. Stormo. *CABIOS* (*Bioinformatics*) **11** (1995) 563 - 566.

网址：

`ftp://beagle.colorado.edu`

从子目录 /pub 中取得的 `imd_1.1_tar.gz` 是 1997 年的 UNIX 版本，此后未见更新。

R-221 **TRRD**，真核生物基因组转录调控区数据库 (Transcription Regulatory Regions Database)。这个由俄国科学院新西伯利亚分院细胞和遗传研究所建立和维护的数据库，包含转录调控区结构与功能组织、所涉及的蛋白质因子，以及转录信号等数据。1999 年底的第 4.2.5 版包括 760 个基因、3 403 个表达图谱和 4 600 多个调控元件的描述，后者包含 3 604 个转录因子结合位点、600 个启动子和 152 个增强子。请参见：

N. A. Kolchanov 等 16 位作者， *Nucleic Acids Res.* **28** (2000) 298 - 301.

网址：

`http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/`

R-222 **OOTFD**，转录因子和基因表达数据库，库本身采用面向对象 (即

[21] E. Wingender, *Nucleic Acids Res.* **16** (1988) 1879 - 1902; *Crit. Rev. Eukaryot. Gene Expr.* **1** (1990) 11 - 48.

OOP) [R-51] 的程序设计技术。这是原来 TFD 关系数据库的 OO 版本。请参见:

D. Ghosh, *Nucleic Acids Res.* **28** (2000) 308 - 310.

网址:

`http://www.ifti.org/`

脊椎动物基因组中各种重复序列占有很高的比例。早在发明 DNA 测序方法之前,就靠密度梯度法发现了重复单元达数千碱基对,但在整个基因组中位点不多的高冗余度的重复序列。它们的组分不很均匀,集中在着丝粒附近和性染色体的异染色质 (heterochromatin) 区域。在早期空间技术发展的历史背景下,这类重复序列被称为"卫星" (satellites) 序列。它们很难被克隆,至今仍是妨碍真核生物基因组完全测序的重要因素。后来又发现了重复单元为 9 ~ 100 碱基对,重复可达数百次的"小卫星" (minisatellites) 序列,以及重复单元为 1 ~ 6 个碱基对,重复约 100 次的"微卫星" (microsatellites) 序列。前者较为集中在亚端粒 (telomere) 区域,后者散布在染色体各处。现在知道,微卫星序列的扩增是若干遗传病的原因。重复序列妨碍序列的联配和测序片段的组装。因此,已经发展了一些重复序列数据库。

R-223 **RepBase**,真核生物 DNA 中重复序列数据库,由非营利性的遗传信息研究所 (Genetic Information Research Institute,简称 GIRI) 维护。1999 年 12 月为 4.04 版。RepeatMasker 程序 [R-748] 即根据 RepBase 库工作。此库区别对待学术性和商业性用户。从事学术研究的个人可在登记注册并承诺履行协议后,免费下载。详情请见网址:

`http://www.girinst.org/~server/repbase.html`

R-224 **MicroSatellite**。微卫星重复序列数据库,在 Smithsonian 分子系统学实验室 (Laboratory of Molecular Systematics)。网址:

`http://nmnhgoph.si.edu/gopher-menus/MicroSatelliteDatabase.html`

R-225 **ALU** 数据库是人及其他灵长类代表性的 Alu 重复片段库。此库是从 RepBase 中提取的 Alu 序列。网址:

`ftp://ncbi.nlm.nih.gov (/pub/jmc/alu/)`

R-226 **Simple Repeats**,简单重复序列库。描述参见:

J. Jurka, and C. Pethiyagoda, *J. Mol. Evol.* **40** (1995) 120 - 126.

网址:

`ftp://ncbi.nlm.nih.gov`

访问文件 (`/repository/repbase/REF1/simple.ref`).

R-227 **COMPEL**, 复合元件 (composite elements) 数据库. 这是指相邻或部分重叠的蛋白质结合位点, 它们结合来自不同的因子家族或不同的信号途径的蛋白质, 提供转录的组合调控. 详见:

O.V. Kel-Margoulis, A. G. Romashchenko, N. A. Kolchanov, E. Wingender, and A. E. Kel, *Nucleic Acids Res.* **28** (2000) 311 - 315.

这个由俄国新西伯利亚细胞和遗传研究所建立的数据库, 最好从德国的网址访问:

`ftp://ftp.gbf-braunschweig.de (/pub/compel/)`

R-228 **MPDB**, 分子探针数据库, 包含大约 4 000 种人工合成的寡核苷酸, 每个序列可长达 100 个核苷酸. 请参见:

M. Giuseppina Campi 等 10 位作者, *Nucleic Acids Res.* **26** (1998) 147 - 149.

网址:

`http://www.biotech.ist.unige.it/interlab/mpdb.html`

R-229 **HvrBase**, 灵长类 mtDNA 调控区序列库, 主要是人的 HVI 和 HVII 两个高变异区的序列. 请参见:

F. Burckhardt, A. von Haeseler, and S. Meyer, *Nucleic Acids Res.* **27** (1999) 138 - 142.

网址:

`http://monolith.eva.mpg.de/hvrbase/`

R-230 **PlantCARE**, 植物顺式作用 (*cis*-acting) 调控因子数据库. 请参看:

S. Rombauts, P. Dehais, M. Van Montagu, and P. Rouze, *Nucleic Acids Res.* **27** (1999) 295 - 296.

网址:

`http://sphinx.rug.ac.be:8080/PlantCare/`

R-231 **PLACE** 是从文献中搜集的植物顺式作用 调控元件 DNA 模体 的数据库. 只涉及维管植物. 请参看:

K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga, *Nucleic Acids Res.* **27** (1999) 297 - 300.

网址:

http://www.dna.affrc.go.jp/htdocs/PLACE/

ftp://ftp.dna.affrc.go.jp (/pub/dna_place/place.seq 只有原始数据)

R-232 **Mendel** 数据库，搜集植物 STS 和 EST 序列，并加有相应基因家族的信息。网址:

http://jiio6.jic.bbsrc.ac.uk/

R-233 **HOX Pro** 同源异形盒基因数据库. 同源异形盒 (homeobox) 是 DNA 中高度保守的一段，长约 180 碱基对，编码 60 个氨基酸的同源异形结构域 (homeodomain) [R-435]。1984 年首先在果蝇中被发现，现在知道它普遍存在于真核生物基因组中. 请参看 [R-435] 和以下描述文章:

A. V. Spirov, T. Bowler, and J. Reinitz, *Nucleic Acids Res.* **28** (2000) 337 - 340.

网址:

http://spirov.iephb.nw.ru/hox_pro/hox-pro00.html

R-234 **OPD**，寡核苷酸探针数据库 (Oligonucleotide Probe DataBase)。它包含寡核苷酸探针用于膜杂交、原位 (*in situ*) 杂交及作为 PCR 引物的设计和使用信息，反映已发表和未发表的实验数据和文献. 网址:

http://www.cme.msu.edu/OPD/

R-235 **dbSTS**，序列标记位点 (Sequence Tagged Sites) 数据库. 网址:

http://www.ncbi.nlm.nih.gov/dbSTS/

ftp://ncbi.nlm.nih.gov (/repository/dbSTS)

R-236 **dbEST**，这是 GenBank [R-212] 的重要组成部分，它包含若干物种的已表达的序列标记 (Expressed Sequence Tag) 信息。此库开始于 1993 年. 当时的描述参见:

M. S. Boguski, T. M. J. Lowe, and C. M. Tolstoshev, *Nature Genetics* **4** (1993) 332 - 333.

网址:

http://www.ncbi.nlm.nih.gov/dbEST/

ftp://ncbi.nlm.nih.gov (/repository/dbEST)

R-237 **AMmtDB**，后生动物线粒体 DNA 多序列联配数据库，它搜集了脊椎动物线粒体中编码蛋白质和 tRNA 的多 DNA 序列对比数据，以及哺乳动物 mtDNA 主调控区 (D-loop) 序列联配数据。请参看：

C. Lanave, S. Liuni, F. Licciulli, and M. Attimonelli, *Nucleic Acids Res.* **28** (2000) 153 - 154.

网址：

http://bio-www.ba.cnr.it:8000/BioWWW/#AMMTDB

R-238 **HOVERGEN**，脊椎动物同源基因数据库 (HOmologous VERtebrate GENes)，它的特点是搜集非编码区的高度保守的多序列联配。描述见：

L. Duret, D. Mouchiroud, and M. Gouy, *Nucleic Acids Res.* **22** (1994) 2360 - 2365.

网址：

http://acnuc.univ-lyon1.fr/

ftp://biom3.univ-lyon1.fr (/pub/hovergen)

ftp://ftp.infobiogen.fr (/pub/db/acnuc/hovergen)

ftp://ncbi.nlm.nih.gov (/repository/hovergen)

R-239 **DNA 结构参数库**，BEND [R-751] 等 DNA 结构预测程序使用此库中参数。描述见：

R. Lavery, and H. Sklenar, *J. Biomol. Struct. Dyn.* **6** (1988) 63 - 91.

网址：

ftp://transfac.gbf.de (/pub/structure_library)

R-240 **NUCLEOSOME 数据库**，收集实验测定的核小体数据，用于预测 DNA 中与组蛋白八聚体结合的位点。描述见：

I. Ioshikhes, and E. N. Trifonov, *Nueceic Acids Res.* **21** (1993) 4857 - 4859.

网址：

ftp://ftp.ebi.ac.uk /pub/databases/nucleosomal_dna/

R-241 **SELEX_DB**，随机化序列库。俄国新西伯利亚细胞和遗传学研究所的理论遗传学研究室建立此库，提供专为基因组注释参考用的随

机化的 DNA 和 RNA 序列. 请参看:

J. V. Ponomarenko 等 7 位作者, *Nucleic Acids Res.* **28** (2000) 205 - 208.

网址:

`http://wwwmgs.bionet.nsu.ru/mgs/systems/selex/`

除了后面要介绍的 YIDB [R-361] 库之外, 最近又建立了几个有关内含子、外显子和 mRNA 前体剪接的新数据库:

R-242 **ASDB**, 交替剪接基因的数据库. 描述见:

I. Dralyuk, M. Brudno, M. S. Gelfand, M. Zorn, I. Dubchak, *Nucleic Acids Res.* **28** (2000) 296 - 297.

网址:

`http://hattrick.lbl.gov:8888/`

R-243 **Intronerator**, 秀丽线虫 [R-94] 内含子和交替剪接数据库. 描述见:

W. J. Kent, and A. M. Zahler, *Nucleic Acids Res.* **28** (2000) 91 - 93.

网址:

`http://www.cse.ucsc.edu/~kent/intronerator/`

R-244 **IDB 和 IEDB**, 前者是内含子序列数据库, 后者是内含子演化数据库. 1999 年 8 月 IDB 包含 63 000 个基因和 154 000 个内含子; IEDB 总结了 2 800 个物种的信息. 这两个库目前每两年更新一版. 请参看:

N. J. Schisler, and J. D. Palmer, *Nucleic Acids Res.* **28** (2000) 181 - 184.

网址:

`http://nutmeg.bio.indiana.edu/intron/index.html`

R-245 **EID**, 外显子、内含子数据库. 它尽可能完全地搜集了具有内含子的、编码蛋白质的基因数据. 1999 年 8 月 EID 包含 51 289 个基因和 287 209 个与内含子为邻的外显子. 请参看:

S. Saxonov, I. Daizadeh, A. Fedorov, and W. Gilbert, *Nucleic Acids Res.* **28** (2000) 185 - 190.

网址:

`http://mcb.havard.edu/gilbert/EID/`

R-246 **ExInt**，外显子、内含子数据库。请参看：

M. Sakharkar, M. Long, T. W. Tan, and S. J. de Souza, *Nucleic Acids Res.* **28** (2000) 191 - 192.

网址：

`http://intron.bic.nus.edu.sg/rint/exint.html`

R-247 **NDB**，核酸晶体结构数据库。请参看：

H. M. Berman 等 9 位作者， *Biophys. J.* **63** (1992) 751 - 759.

网址：

`ftp://ndbserver.rutgers.edu/`

`http://ndbserver.rutgers.edu/NDB/ndb.html`

有一些与 DNA 测序或基因工程关系更密切的数据库，也放在 DNA 数据库这一节略加介绍。首先是几个载体数据库。它们不仅用于设计载体，而且往数据库提交新序列前，以及从库中取来的 DNA 序列，如 EST 序列，往往要借助 VectorDB 和 Vector-ig 等库排除其中误带的载体序列片段。

R-248 **VectorDB**，载体数据库，搜集载体序列和常用载体的有关信息，均为 GenBank 格式 [R-179]。网址：

`http://vectordb.atcg.com/`

R-249 **Vector 和 Vector-ig** 库，包含分子生物学常用的许多载体的注释和序列信息。网址：

`ftp://ncbi.nlm.nih.gov (/repository/vector-ig)`

`ftp://ncbi.nlm.nih.gov (/repository/vector)`

R-250 另一个有用的载体库在：

`http://biology.queensu.ca/miseners/vector.html`

R-251 **UniVec 数据库**。NCBI [R-134] 的 VecScreen 服务使用 UniVec 数据库过滤序列中来自载体的片段。网址：

`http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html`

§4.7 RNA 序列和核糖体数据库

各种 RNA 在细胞中起着多种作用。除了信使 RNA (mRNA)、转运

RNA (tRNA)、构成核糖体骨架的 rRNA、导引 RNA (gRNA)，还有各种不翻译成蛋白质的 RNA，起着调控或催化作用，有些功能尚未阐明。作为蛋白质制造厂的核糖体，三分之二由 RNA 组成，它们的数据库也放在这一节介绍。

R–252 1993 年成立的 RNA 学会，在出版 *RNA* 刊物同时，还维护着两个信息网页：

http://www.pitt.edu/~rna1/

http://www.cup.org/Journals/JNLSCAT/rRNA/rna.html

此外，还请参看有关 RNA 的生物信息学会议文集 [R–829] 和 Ambion 公司发行的电子通信 RNA Flashnotes [R–806]。

R–253 **snoRNA 数据库**。小核仁 RNA (snoRNA) 是真核生物细胞核仁中的一类稳定的 RNA，在酵母中约有 75 ~ 100 种，在哺乳动物中可能达 200 种。它们不直接参与蛋白质合成，但与 rRNA 的切割和折叠有关。这是由酿酒酵母基因组提取的 snoRNA 数据库，详细描述请参看：

D. A. Samarsky, and M. J. Fournier, "A comprehensive database for the small nucleolar RNAs from *S. cerevisiae*", *Nucleic Acids Res.* **27** (1999) 161 – 164.

网址：

http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html

R–254 **Small RNA 数据库**。所谓小 RNA 是指哪些不直接参与蛋白质合成的 RNA 分子。真核生物核仁、细胞质、线粒体，以及一些原核生物和病毒都含有小 RNA。库的描述参见：

Jian Gu, Yahua Chen, and Ram Reddy,*Nucleic Acid Res.* **26** (1998) 160 – 162.

网址：

http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html

R–255 **RNAse P 数据库**，包含 RNA 水解酶 P 的 RNA 亚基序列、联配、二级结构和三维模型。描述见：

J. W. Brown, *Nucleic Acids Res.* **26** (1998) 351 – 352.

网址：

http://jwbrown.mbio.ncsu.edu/RNAseP/home.html

tmRNA 旧称 10Sa RNA，其遗传学名字为 SsrA，迄今只在真细菌和一些细胞器中发现。它在 mRNA 翻译成蛋白质的最后阶段有重要作用。现在有一个 tmRNA 网点 [R-256] 和一个 tmRDB 数据库 [R-257]。

R-256 **tmRNA** 网点，包含 tmRNA 序列、公认蛋白质水解标记、序列联配、确定新 tmRNA 的导引，以及简要综述等。见：

K. P. Williams, *Nucleic Acids Res.* **28** (2000) 168.

网址：

http://www.indiana.edu/~tmrna/

R-257 **tmRDB**，已经联配好的、加有注释的、按亲缘关系排列的 tmRNA 序列数据，详见：

C. Zwieb, and J. Wower, *Nucleic Acids Res.* **28** (2000) 169 - 170.

网址：

http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html

R-258 **gRNA**，导引 RNA (guide RNA) 数据库，包含已发表的 gRNA 序列和文献。其 3.0 版介绍见：

S. Hinz, and H. U. Geringer, *Nucleic Acids Res.* **27** (1999) 168.

网址：

http://www.biochem.mpg.de/~goeringe/

R-259 **SRPDB**，信号识别粒子数据库。这是研究信号识别粒子 (Signal Recognition Particle，简称 SRP) 功能与结构的工具。它提供真核生物和古细菌的带注释的 SRP 的 RNA 序列，按亲缘关系排列，并同它们的细菌等价序列联配。请参看：

C. Zwieb, and T. Samuelsson, *Nucleic Acids Res.* **28** (2000) 171 - 172.

关于 SRP 的较新报道，还可参看下文及其所引同期文章：

P. Walter, R. Keenan, and U. Schmitz, *Science* **287** (2000) 1212.

网址：

http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html

R-260 **TransTerm**，信使 RNA 的组分和翻译控制信号数据库。它包括 GenBank [R-212] 中许多物种的起始和终止密码子、密码子使用频度表、5′UTR 和 3′UTR 序列、编码区的某些参数如 GC 偏离等。1999 年 10 月 TransTerm 包含来自 1 万个物种的数据，其中有 20 个原核

生物和 3 个真核生物的完全基因组。请参看：

G. H. Jacobs 等 5 位作者， *Nucleic Acids Res.* **28** (2000) 293 - 295.

网址：

`http://biochem.otago.ac.nz/Transterm/`

R-261 类病毒 (viroids) 和类病毒样 (viroid-like) RNA 数据库。这是会自我复制的最小的 RNA 物种。请参看：

M. Pelchat, P. Deschenes, and J. P. Perreault, *Nucleic Acids Res.* **28** (2000) 179 - 180.

网址：

`http://www.callisto.si.usherb.ca/~jpperra/`

R-262 **UTRdb 和 UTRsite**。许多基因表达的调控元件在 DNA 的非翻译区中。UTRdb 是真核生物 mRNA 的 5′ 端和 3′ 端非翻译区序列的非冗余数据库，UTRsite 搜集这些非翻译区序列中的功能片段。此网页还提供两种分析工具：UTRFasta 检查用户提交的序列是否包含 UTRdb 中的序列；UTRScan 检查用户序列中是否有 UTRsite 中的片段。此库的最近描述见：

G. Pesole 等 7 位作者， *Nucleic Acids Res.* **28** (2000) 193 - 196.

网址：

`http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome/`

R-263 **ncRNA**，似 mRNA 的非编码 RNA 数据库。描述见：

V. A. Erdmann 等 5 位作者， *Nucleic Acids Res.* **28** (2000) 197 - 200.

网址：

`http://www.man.poznan.pl/5SData/ncRNA/index.html`

R-264 **RNAmods**， RNA 修饰数据库、是转录后 RNA 核苷酸修饰的清单，最初以书面表格形式发表于：

P. A. Limbach, P. F. Crain, and J. A. McCloskey, *Nucleic Acids Res.* **22** (1994) 2183 - 2196.

此库是该文数据的不断更新补充。较近情况见：

J. Rozenski, P. F. Crain, and J. A. McCloskey, *Nucleic Acids Res.* **27** (1999) 196 - 197.

网址：

`http://www-medlib.med.utah.edu/RNAmods/RNAmods.html`

ftp://medlib.med.utah.edu (/library/RNAmods)

R-265 **AARSDB**，酰氨基 tRNA 合成酶 (aminoacyl-tRNA synthetase) 数据库。请参看：

M. Szymanski, and J. Barciszewski, *Nucleic Acids Res.* **28** (2000) 326 - 328.

网址：

http://rose.man.poznan.pl/aars/index.html

R-266 **tRNA** 序列和基因、结构与功能数据库。请参看：

M. Sprinzl 等 5 位作者， "Compilation of tRNA sequences and sequences of tRNA genes", *Nucleic Acids Res.* **26** (1998) 148 - 153.

网址：

http://www.uni-bayreuth.de/departments/biochemie/trna/

R-267 **PLMItRNA**，基于 FastA [R-641] 的绿色植物 (25 种高等植物和 7 种绿藻) 线粒体 tRNA 分子和 tRNA 基因的数据库，包括 tRNA 基因多序列联配。详见：

V. Volpetti 等 6 位作者， *Nucleic Acids Res.* **28** (2000) 159 - 162.

网址：

http://bio-www.ba.cnr.it:8000/srs6/

http://www.ebi.ac.uk/services/

http://bigarea.area.ba.cnr.it:8000/BioWWW/fasta.htm

R-268 **16SMDB**、**16S-likeMDB**、**16SMDBexp**、**23SMDP**、**23S-likeMDB** 和 **23SMDBexp** 数据库。这是富兰克林和马歇尔学院 (Franklin and Marshall College) 生物系的 K. L. Triman 所维护的一批 16S 和 23S 核糖体 RNA 突变数据库，其较近的描述见：

K. L. Triman, A. Peister, and R. A. Goel, *Nucleic Acids Res.* **26** (1998) 280 - 284.

网址：

http://www.fandm.edu/departments/
biology/databases/rna.html

ftp://acad.fandm.edu (/nar/)

R-269 **RNA www**，RNA 二级结构网页，也有 16S RNA 和 23S RNA 的数据。描述见：

R. R. Gutel 等, *Nucleic Acids Res.* **21** (1993) 3055 - 3074; **22** (1994) 3051 - 3054.

网址:

http://pundit.colorado.edu:8080/RNA/

R-270 **uRNADB**, 已经联配好的、加有注释的、按亲缘关系排列的 uRNA 序列数据, 描述见:

C. Zwieb, *Nucleic Acids Res.* **25** (1997) 102 - 103.

网址:

http://psyche.uthct.edu/dbs/uRNADB/uRNADB.html

R-271 **U-insertion/deletion** 编辑序列数据库, 包含 5 个无脊椎动质体目 (Kinetoplastida) 物种的线粒体基因和编辑后的 mRNA 序列. 描述见:

L. Simpson, *Nucleic Acids Res.* **26** (1998) 170 - 176.

网址:

http://www.lifesci.ucla.edu/RNA/trypanosome/database.html

R-272 **PseudoBase**, 假扭结数据库. 自从 1985 年发现 RNA 二级结构中的假扭结 (pseudoknot) 以来, 文献中已有不少关于假扭结的报道. 1998 年建立的 PseudoBase 目的即在于汇总这方面的知识, 并提供一个发表数据的场所. 这里还有一个名为 STAR 的程序, 可以预测某些假扭结. 请参看:

F. H. D. van Batenburg 等 5 位作者, *Nucleic Acids Res.* **28** (2000) 201 - 204.

网址:

http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html

R-273 **RDP**, 核糖体数据库计划 (Ribosomal Database Project), 包含小亚基 (SSU) 和大亚基 (LSU) 的两部分 rRNA, 由已联配和未联配的 RNA 序列以及亲缘树组成, 详见:

B. L. Maidak 等 12 位作者, *Nucleic Acids Res.* **28** (2000) 173 - 174.

网址:

http://www.cme.msu.edu/RDP/

http://rdpwww.life.uiuc.edu/

ftp://rdp.life.uiuc.edu (/pub/)

mailto: server@rdp.life.uiuc.edu

北京大学生物信息中心 [R-166] 有镜象。

R-274 **GenCANS-RDP**，这是把原来为蛋白质数据自动分类而设计的 GenCANS(Gene Classification Artificial Neural System) 系统，推广到 RDP [R-273] 计划中 rRNA 序列而得到的分类。描述见：

C. Wu, and S. Shivakumar, *Nucleic Acids Res.* **22** (1994) 4291 - 4299.

网址：

http://diana.uthct.edu/~nih/cans/gencans_rdp.html

R-275 **SSU rRNA**，欧洲核糖体小亚基 RNA 结构数据库。描述见：

Y. Van de Peer 等 5 位作者，*Nucleic Acids Res.* **28** (2000) 175 - 176.

网址：

http://rrna.uia.ac.be/ssu/

ftp://rrna.uia.ac.be (/pub/)

R-276 **LSU rRNA**，欧洲核糖体大亚基 RNA 结构数据库。描述见：

P. De Rijk 等 5 位作者，*Nucleic Acids Res.* **28** (2000) 177 - 178.

网址：

http://rrna.uia.ac.be/lsu/

ftp://rrna.uia.ac.be (/pub/)

R-277 **5S rRNA** 数据库。当前版本包含 1 985 个 5S rRNA 及其基因 5S rDNA 的一级结构，按来源物种分类。描述见：

M. Szymanski, M. Z. Barciszewska, J. Barciszewski, and V. A. Erdmann, *Nucleic Acids Res.* **28** (2000) 166 - 167.

网址：

http://rose.man.poznan.pl/5SData/index.html

R-278 **DRC**(Database of Ribosomal Crosslinks)，核糖体交链数据库。为了理解翻译过程，必须阐明核糖体的高级结构。本数据库集中了大肠杆菌 rRNA 之间、rRNA 与核糖体蛋白质、核糖体蛋白质之间、核糖体大小亚基之间等各种层次的交链数据。请参看：

P. V. Baranov 等 5 位作者，*Nucleic Acids Res.* **27** (1999) 184 - 185.

网址：

http://www.mpimg-berlin-dahlem.mpg.de/~ag_ribo/ag_brimacombe/drc/

由于是德俄合作项目，在莫斯科还有一个镜象点：

http://ribosome.genebee.msu.su/DRC/

R-279 **ACTIVITY**， DNA 和 RNA 中功能位点数据库. 这是一个分布式的智能库， 1999 年底的第 1.1.5 版共有 511 个条目。描述见：

J. V. Ponomarenko 等 9 位作者， Activity: a database for activities of functional DNA/RNA sites, in *Proceedings of BGRS'98*, Novosibirsk, 1998, 62 - 65;

以及 [R-720] 的引文。两文均可从网页下载. 网址：

http://wwwmgs.bionet.nsu.ru/systems/Activity/

R-280 **RNA** 非正则配对数据库. 它搜集已知 RNA 结构中少见的非正则碱基配对。描述见：

U. Nagaswamy, N. Voss, Z. D. Zhang, and G. E. Fox, *Nucleic Acids Res.* **28** (2000) 375 - 376.

网址：

http://prion.bchs.uh.edu/bp_type/

§4.8 基因图谱数据库

R-281 **RHdb**，辐射杂交数据库。 1999 年 7 月的第 16.0 版包含人、大鼠和家鼠三个物种的 105 216 个辐射杂交条目。目前每夜发布一次进展报告。请参看：

P. Rodriguez-Tome, and P. Lijnzaad, *Nucleic Acids Res.* **28** (2000) 146.

网址：

http://www.ebi.ac.uk/RHdb

http://corbra.ebi.ac.uk/RHdb/species/HUMAN/gm99.html

ftp://ftp.ebi.ac.uk (/pub/databases/RHdb)

北京大学生物信息中心 [R-166] 有镜象。

R-282 **Mouse RH** 数据库。 Whitehead 生物医学研究所的这个库，包括小鼠的 19 个染色体和 X 染色体的辐射杂交数据。网址：

http://www-genome.wi.mit.edu/mouse_rh/

R-283 **GDB**，人类基因组数据库，是使用较多的一个重要数据库。这是

原来 John Hopkins 大学 (JHU) 医学院维护的人类基因组数据库。1998 年因经费危机险些寿终，该年底 GDB 主节点移至加拿大多伦多儿童医院生物信息超级计算中心，数据库的审读仍在 JHU 进行。GDB 是人类基因图谱和疾病的数据库，目的在于支持构建人类基因图谱和测序。请参看：

S. I. Letovsky, R. W. Cottingham, C. J. Porter, and P. W. D. Li, *Nucleic Acids Res.* **26** (1998) 94 - 99.

多伦多儿童医院的网址：

`http://www.bioinfo.sickkids.on.ca/`

GDB 的原网址：

`http://www.gdb.org/` 或

`http://wwwgdb.gdb.org/`

`ftp://ftp.gdb.org`

仍可继续使用。全世界有十多处 GDB 镜象点。中国镜象在北京大学生物信息中心 [R-166]，它有一个专用网址：

`http://gdb.pku.edu.cn/gdb/`

此外，请参看 VIRGIL[R-321] 数据库。

R-284 **GeneMap'99**，人类基因图谱 1999 年版，由国际辐射杂交图谱协作组提供，是以下论文的更新的电子附录：

P. Deloukas, *Science* **282** (1998) 744 - 746.

网址：

`http://www.ncbi.nih.gov/genemap/`

目前它包含 3 万多个基因位点。纯文本格式的人类基因图谱 99 可由以下网址获取：

`ftp://ftp.ebi.ac.uk` (`/pub/databases/RHdb/gm99.map`)

R-285 **HuGeMap**，人类基因遗传图谱和物理图谱的分布式集成数据库，提供图谱的相互联系和视象化表示。请参看：

E. Barillot 等 6 位作者， *Nucleic Acids Res.* **27** (1999) 119 - 122.

网址：

`http://www.infobiogen.fr/services/Hugemap/`

§4.9 人类基因组有关数据库

"人类基因组组织" (HUGO) 促进着人类基因组计划的国际合作。实际上，计划的主要完成者是美国国家卫生署和能源部支持的一批实验室，以及英国 Wellcome Trust 支持的若干研究中心。中国从 1999 年 9 月 1 日起正式承担了国际人类基因组测序任务的 1%，也就是第 3 号染色体上 3 000 万碱基对的测定，并在 2000 年 6 月 26 日与其他参加国家共同宣布完成了人类基因组的 "工作草图"。

4.9.1 人类基因组测序中心

关于人类基因组计划的由来和近况，可参看中文文集：

R-286 贺林主编，《解码生命。人类基因组计划和后基因组计划》，科学出版社，2000。

从国际上几个主要人类基因组计划资助机构的网页，可以获知该计划的一般情况：

R-287 **HUGO** 是人类基因组组织 (HUman Genome Organization) 的缩写。网址：

`http://hugo.gdb.org/`

R-288 HUGO 的太平洋部分有一个网页设在日本并发行 HUGO Pacific GENOME Newsletter。其反映中国情况的短文在：

`http://hugo-pacific.genome.ac.jp/3_2contents/china.html`

可惜内容已经几年没有更新了。

R-289 美国能源部支持的人类基因组计划见：

`http://www.er.doe.gov/production/ober/hug_top.html`

R-290 美国国家卫生署对人类基因组计划的支持，通过 **NHGRI** 即国家人类基因组研究所 (National Human Genome Research Institute) 体现，其网址是：

`http://www.nhgri.nih.gov/`

R-291 英国 **Wellcome Trust** 是人类基因组计划的另一个主要资助者。网址：

`http://www.wellcome.ac.uk/`

任何政府资助均源于纳税人的贡献，必须造福整个社会，而不容许同商业集团的利益混淆。在国际人类基因组计划中，这清楚表述在 1996 年 2 月在百慕大举行的第一届人类基因组测序战略会议通过的百慕大原则中：测序的中间和最终结果都必须迅速公开。

R-292 百慕大原则参见：

http://www.gene.ucl.ac.uk/hugo/bermuda.html

欧洲 EMBL 数据库 [R-211] 专门为此建立了 HTG 即 High-Throughput Genome 部分，存放各种基因组大规模测序的中间结果，以利于各个测序中心的交流与合作。一旦完成注释，即移入 EMBL 的相应门类，并从 HTG 中取消。美国 NCBI [R-134] 的 htgs 序列库也起着同样的作用。

事实上，人类基因的各种片段构成多种数据库的主要部分。本节注重与人类基因组计划有关的库，在基因图谱、基因表达、突变、病理、免疫等各节中都还有大量人类基因数据。

R-293 世界上主要人类基因组测序中心的名单见：

http://www-hgc.lbl.gov/inf/HGcenters.html

http://www.ornl.gov/hgmis/centers.html

每个测序中心维护着其特有的数据库，可直接去访问。表 4.6 按照染色体编号，列举了某些数据所在网址。随着人类基因组工作草图的完成，各个基因组中心正在调整他们的数据库。因此，这里提供的网址会有不少变动。请特别注意几个大的测序中心的网页，例如：

R-294 NCBI [R-134] 的 GenBank 数据库 [R-212] 从 1999 年 10 月起，建立了智人 (*Homo sapiens*) 基因组子目录，其下按染色体编号设子目录。网址：

http://ncbi.nlm.nih.gov/genbank/genomes/H_sapiens/

R-295 英国的 Sanger 中心的人类基因组计划网页，不仅有它们负责测序的染色体数据，还有到其他染色体数据的链接。网址：

http://www.sanger.ac.uk/HGP/

R-296 日本的 DDBJ [R-213] 和信息生物学中心 (Center for Information Biology，简称 CIB) 联合建立了一个 HUman Genomics Studio，可以按染色体编号检索和查找基因序列。网址：

http://studio.nig.ac.jp/

此外，第 22 号和第 21 号染色体的基本部分，即常染色质 (euchromatin) 的序列，已经在 1999 年 12 月和 2000 年 5 月发表，参见：

R-297 I. Dunham 等 217 位作者，"The DNA sequence of human chromosome 22", *Nature* **402** (1999) 489 - 495.

R-298 第 22 号染色体图谱和测序协助组，以及 M. Hattori 等 63 位作者，"The DNA sequence of human chromosome 21", *Nature* **405** (2000) 311 - 319.

R-299 **Sanger** 中心是世界上最大的 DAN 测序中心之一。它承担着人类基因组计划三分之一、即 10 亿碱基对的测序任务，以及一些其他物种的测序。人类基因组测序集中在以下各染色体：1、6、9、10、13、20、22 和 X。目前其测序进展统计每 20 分钟自动更新一次，参见：

http://www.sanger.ac.uk/HGP/stats.shtml

R-300 **LBNL**，Lawrence Berkeley 国家实验室，其人类基因测序部，现在是联合基因组研究所 JGI [R-303] 的一部分，网址：

http://www-hgc.lbl.gov/GenomeHome.html

R-301 **LLNL**，Lawrence Livermore 国家实验室，其生物学与生物技术研究计划 (Biology and Biotechnology Research Program，简称 BBRP) 完成了第 19 号染色体的高分辨率、可用以测序的图谱；其与华盛顿大学、Merck 公司等合作单位组成的 I.M.A.G.E. 协作组 [R 314]，拥有目前最大的、已测序的 cDNA 克隆。网址：

http://www-bio.llnl.gov/bbrp/genome/genome.html

R-302 **LANL**，美国洛斯阿拉莫斯国家实验室，其人类基因组研究中心 (Center for Human Genome Studies，简称 CHGS)，主要从事第 16 号染色体的图谱和测序。网址：

http://www-ls.lanl.gov/index.html

R-303 **JGI**，由美国能源部支持的，依托 LBNL [R-300]、LLNL [R-301] 和 LANL [R-302] 三个国家实验室的人类基因组研究部门组建的联合基因组研究所 (Joint Genome Institute)。它于 1999 年 1 月正式启用强大的高产出测序设备，目标是产生高质量的序列，足以区分单核

表 4.6 人类染色体数据网址

染色体	网址
1	http://linkage.rockefeller/chr1/
	http://www.sanger.ac.uk/HGP/Chr1/
2	http://www.sanger.ac.uk/HGP/Chr2/
3	http://mars.uthscsa.edu/
	http://www.genomics.org.cn/
4	http://www.sanger.ac.uk/HGP/Chr4/
5	http://www.jgi.doe.gov/
6	http://www.sanger.ac.uk/HGP/Chr6/
7	http://www.genet.sickkids.on.ca/chr7db/
8	http://gc.bcm.tmc.edu:8080/chr8/home.html
9	http://www.gene.ucl.ac.uk/chr9/
	http://www.sanger.ac.uk/HGP/Chr9/
10	http://www.cric.com/htdocs/chr10-mapping/
	http://www.sanger.ac.uk/HGP/Chr10/
11	http://chr11.bc.ic.ac.uk/
	http://mcdermott.swmed.edu/datapage/
	http://shows.med.buffalo.edu/database.html
12	http://paella.med.yale.edu/chr12/home.html
13	http://genome1.ccc.columbia.edu/~genome/
	http://www.sanger.ac.uk/HGP/Chr13/
14	http://www.sanger.ac.uk/HGP/Chr14/
15	http://www.sanger.ac.uk/HGP/Chr15/
16	http://www.jgi.doe.gov/
	http://www.tigr.org/tdb/humgen/c16.html
17	http://bioinformatics.weizmann.ac.il/
18	http://www.sanger.ac.uk/HGP/Chr18/
19	http://www.jgi.doe.gov/
	http://www-bio.llnl.gov/bbrp/genome/genome.html
20	http://www.expasy.ch/cgi-bin/lists?humchr20.txt
	http://www.sanger.ac.uk/HGP/Chr20/

表 4.6 (续表)

染色体	网址
21	http://www.expasy.ch/cgi-bin/lists?humchr21.txt
	http://www-ari.uchsc.edu/chr21/welcome.html
	http://www.cephb.fr/chromosome21.html
22	http://www.cbil.upenn.edu/HGC22.html
	http://www.expasy.ch/cgi-bin/lists?humchr22.txt
	http://www.sanger.ac.uk/hum22/ /HGP/Chr22/
	http://www.genome.ou.edu/gifs/
X	http://gc.bcm.tmc.edu:8080/chrX/home.html
	http://www.expasy.vh/cgi-bin/lists?humchrx.txt
	http://www.sanger.ac.uk/HGP/ChrX/
Y	http://www.expasy.ch/cgi-bin/lists?humchry.txt
线粒体	http://infinity.gen.emory.edu/mitomap.html

苷酸多态性 和测序错误，区分功能基因和假基因 等。人类基因组测序集中在第 5 、 16 和 19 号染色体。网址：

http://jgi.doe.gov/

R-304 **UWGC**，华盛顿大学基因中心，是国际上最活跃的测序中心之一。正在进行的工作包括人类第 7 号染色体，人白细胞抗原 HLA 第一类基因座、家鼠 T 细胞受体 α 区，以及绿脓假单胞菌 (*Pseudomonas aeruginosa*) 的图谱和测序。这里有不少与测序有关的软件，如 Phrap [R-691] ， RepeatMasker [R-748] 等。网址：

http://www.genome.washington.edu/

ftp://ftp.genome.washington.edu/

R-305 **SHGC** ，斯坦福大学人类基因中心，主要做高分辨率辐射杂交图谱，以及人类第 4 号染色体 BAC 克隆的测序。网址：

http://www-shgc.stanford.edu/

R-306 美国哥伦比亚大学基因中心，主要研究和人类疾病有关的基因和第 13 号染色体图谱。网址：

http://genome1.ccc.columbia.edu/~genome/

http://genome3.cpmc.columbia.edu/~legion/

R-307 **GÉNÉTHON** ，法国人类基因组研究中心。网址：

http://www.genethon.fr/genethon_en.html

4.9.2 人类基因组有关数据库

GenBank[R-212]、EMBL[R-211]、GSDB[R-214]、GDB[R-283]等综合数据库的主要内容都来自人. 下面再列举一批与人类基因组有关的数据库。

R-308 **UniGene**，人类基因序列集合，搜集了 GenBank [R-212] 中不同基因产物的序列. 描述见:

M. S. Boguski, and G. D. Schuler, *Nature Genetics* **10** (1995) 369 - 371.

可通过 NCBI [R-134] 的网页访问:

`http://www.ncbi.nlm.nih.gov/UniGene/`

R-309 **HIB**(Human Info Base) 数据库，是德国人类基因组计划中基因分析项目所建立的自动注释的基因集团数据库。网址:

`http://www.mips.biochem.mpg.de/proj/human/`

它的原始数据来自 UniGene [R-308]，而软件工具是 CAP3 [R-692] 和 PEDANT [R-755]。详情请参阅网址:

`http://www.mips.biochem.mpg.de/desc/human/`

通常把导致不同表现型或疾病的碱基改变称为突变 (mutation)，而不引起表现型或病变的称为多态性 (polymorphism)[22]。近来发现单核苷酸多态性 (Single Nucleotide Polymorphism，简称 SNP) 对于人类遗传学研究有重要意义，于是出现相应的数据库。目前至少有 4 个 SNP 数据库:

R-310 **dbSNP**，设在美国国家生物技术信息中心 NCBI [R-134] 的单核苷酸多态性数据库，收录单核苷酸置换，以及短的删除和插入所导致的多态性. 请参看:

E. M. Smigielski, K. Sirotkin, M. H. Ward, and S. T. Sherry, *Nucleic Acids Res.* **28** (2000) 352 - 355.

网址:

`http://www.ncbi.nlm.nih.gov/SNP/`

R-311 Whitehead 研究所 WI [R-157] 的人类单核苷酸多态性 (SNP) 数据库。这是与 Affymetrix 公司 [R-801] 等合作进行的项目. 见:

`http://www-genome.wi.mit.edu/SNP/human`

[22] Polymorphism 译为多态性较确切，并可有别于生物多样性的 diversity。

R-312 **HGBASE** 是人类的双等位基因序列 (Human Genic Bi-Allelic SEquences) 的缩写. 这是人类基因从启动子到转录终点, 即基因及其前后所发现的所有单核苷酸多态性 和其他变化的数据库. 这不是一个基因突变库, 而是"正常"人基因序列变异的目录. 它不限于双等位基因 SNP, 诸如启动子和非沉默密码子 (non-silent codon) 变异、内含子 变异等也包括在内. 2000 年 2 月 7 日的第 6 版包含 6 688 条基因内多态性记录. 请参看:

A. J. Brookes 等 8 位作者, *Nucleic Acids Res.* **28** (2000) 356 – 360.

网址:

`http://hgbase.interactiva.de/`

`http://hgbase.cgr.ki.se/`

R-313 位于 St. Louis 的华盛顿大学的 SNP 数据库, 网址:

`http://www.ibc.wustl.edu/SNP/`

cDNA 克隆和 BAC 图谱等, 在大规模基因组测序计划中起着重要作用. 这里列举一些有关网址.

R-314 **I.M.A.G.E** 协作组, 其名称缩写来自 Integrated Molecular Analysis of Genomes and their Expression. 他们共享高质量的 cDNA 克隆库, 并把有关序列、图谱和表达数据公开. 请参考长篇介绍:

G. Lenon, C. Auffray, M. Ploymeropoulos, and M. B. Soares, *Genomics* **33** (1996) 1 – 152.

I.M.A.G.E 的网址:

`http://www-bio.llnl.gov/bbrp/image/image.html`

I.M.A.G.E 的克隆识别号 (ID) 出现在许多数据库条目中, 或在 DE 项下, 或在性状表的 /clone= 之后. 这些 ID 可从 NCBI [R-134] 的 dbEST [R-236] 数据库获取, 也可用一个名为 LENS 的浏览器查找:

`http://agave.humgen.upenn.edu/lens/`

如果需要把 I.M.A.G.E 克隆 ID 换成克隆名字, 可以借助:

`http://www.hgmp.mrc.ac.uk/BIO/translate/`

R-315 **ATCC**, 美国菌种保藏中心 (American Type Culture Collection), 它提供包括 cDNA 克隆库在内的各种生物学和分子生物学试剂和材料. 这虽然不是一个生物信息学资源, 但在文献中时有提及. 因此, 我们给出网址:

http://www.atcc.org/

R-316 **GenMapDB**，V. Cheung [R-783] 实验室维护的一个人类 BAC 图谱数据库。它以 1Mbp 的间距覆盖了人类第 2, 14, 15, 16, 17, 18, 19, 20, 21, 22, X 和 Y 等染色体的 BAC 克隆，在 1999 年初总长度达到 1 156Mbp。网址：

http://w95vc1.neuro.chop.edu/vcheung/

R-317 **BAC Ends**，人类 BAC 末端数据库。BAC 末端序列可提供高度特异的标记，对基因组测序有重要作用 (见 3.6.5 小节末尾)。有关 BAC Ends 数据库请参看：

S. Y. Zhao, *Nucleic Acids Res.* **28** (2000) 129 – 132.

网址：

http://www.tigr.org/tdb/humgen/bac_end_search/

ftp://ftp.tigr.org (/pub/data/h_sapiens/bac_ends_sequences)

R-318 **HUGE**，人类未经实验证实的编码 (Human Unidentified Gene-Encoded) 基因的数据库。这是由日本 Kazusa DNA 研究所 cDNA 测序计划所确定的、尚未经实验证实的编码人类大蛋白质的基因数据的集合。请参看：

R. Kikuno 等 6 位作者，*Nucleic Acids Res.* **28** (2000) 331 – 332.

网址：

http://www.kazusa.or.jp/huge/

R-319 **IXDB**，集成的人类 X 染色体物理图谱数据库。数据来自其他数据库、文献和直接投稿。请参看：

U. Leser, H. Roest Crollius, H. Lehrach, and R. Sudbrak, *Nucleic Acids Res.* **27** (1999) 123 – 127.

网址：

http://ixdb.mpimg-berlin-dahlem.mpg.de/

R-320 **Genotype**，法国人类多态性研究中心 (Centre d'Etude du Polymorphisme Humain，简称 CEPH) 的基因型数据库。它搜集人类染色体联锁图谱 (linkage mapping) 中已定型的遗传标记的基因型。请参看：

J. C. Murray 等 27 位作者，"A comprehensive human linkage map with centimorgan density", *Science* **265** (1994) 2049 – 2054.

1998 年 12 月的 Genotype 数据库第 8.2 版包含 11 995 个遗传标记，包括 9 000 多个微卫星标记，其中 57% 是高度多态的. CEPH 数据库总共包含 250 万条以上基因型. 库的网址:

`http://www.cephb.fr/cephdb/`

R-321 **VIRGIL**，专门为 GDB [R-283] 中的人类基因和 GenBank [R-212] 中的 DNA 序列提供对应链接关系的数据库，与 GenBank 同步更新. 请参看:

F. Achard, G. Vaysseix, P. Dessen, and E. Barillot, *Nucleic Acids Res.* **27** (1999) 113 – 114.

原始网址在法国 INFOBIOGEN [R-148]:

`http://www.infobiogen.fr/services/virgil/HPvirgil.html`

`ftp://ftp.infobiogen.fr (/pub/db/virgil/virgil.ffl)`

北京大学生物信息中心 [R-166] 有镜象.

R-322 **KinMutBase**，人类致病蛋白质激酶突变数据库. 请参看:

K. A. E. Stenberg, P. T. Riikonen, and M. Vihinen, *Nucleic Acids Res.* **28** (2000) 369 – 371.

网址:

`http://www.uta.fi/imt/bioinfo/KinMutBase/`

R-323 **CpGIsle**，人类基因中 CpG 岛数据库. CpG 岛是指在同一条 DNA 链中相邻的 CG，写成 CpG 以有别于双链间的 CG 配对. CpG 中的 C 容易被甲基化修饰而产生 C→T 突变. 因此，哺乳动物基因组中 CpG 明显少于 GpC. 这可以作为在 DNA 序列中寻找基因的一种参考. 详见:

A. Bird, "CpG islands as gene markers in the vertebrate nucleus", *Trends in Genetics* **3** (1987) 342 – 347.

CpGIsle 数据库基于对 EMBL [R-211] 数据库中所有人类基因和假基因的分析，这包括含有全部外显子的完整基因序列，也包括部分测序但外显子全部确定而且至少有一个片段长于 2 000 个核苷酸的序列. 短于 2 000 的完整基因也做了分析. 第一个外显子未知或 5′ 端少于 200 的序列均排除在外. 请参看:

F. Larsen, G. Gundersen, R. Lopez, and H. Prydz, *Genomics* **13** (1992) 1095 – 1107.

网址：

ftp://bioslave.uio.no (/cpgisle/)

ftp://ftp.infobiogen.fr (/pub/db/cpgisle/)

ftp://ftp.ebi.ac.uk (/pub/databases/cpgisle)

北京大学生物信息中心 [R-166] 有镜象.

人类肿瘤抑制基因 p53 因其产物分子量为 53kD 而得名，它的位点在 17p13，约有半数癌症与 p53 突变有关。下面是几种 p53 突变数据库。

R-324 **p53 数据库**，建于 1991 年。这是研究人类肿瘤及肿瘤细胞系 p53 基因突变的数据库和软件，实际上由 4 个数据库组成： p53 库、体细胞突变库、种系突变库和细胞系突变库。详情请参看：

C. Béroud, and T. Soussi, *Nucleic Acids Res.* **26** (1998) 200-204.

网址：

http://perso.curie.fr/tsoussi/

相应软件在网页上运行。需要数据库和软件在本地计算机上运行的学者，请与作者联系：

mailto: thieryy.soussi@curie.fr

或

mailto: beroud@ceylan.necker.fr

R-325 **IARC p53 数据库**。法国国际癌症研究会 (International Agency for Research on Cancer，简称 IARC) 的肿瘤和细胞系 p53 基因突变数据库，包括可视化工具。此库的描述见：

P. Hainaut 等 8 位作者， *Nucleic Acids Res.* **26** (1998) 205-213.

网址：

http://www.iarc.fr/p53/homepage.html

http://www.ebi.ac.uk/ (经 services 进入 db 进入 IARC p53)

ftp://ftp.ebi.ac.uk (/pub/databases/p53/)

R-326 **p53 数据库**。具有癌症倾向家族的 p53 种系突变数据库。请参看：

Z. Sedlacek, R. Kodet, A. Poustka, and P. Goetz, *Nucleic Acids Res.* **26** (1998) 214 – 215.

网址：

http://www.lf2.cuni.cz/projects/germline_mu_p53.htm

ftp://ftp.lf2.cuni.cz (/pub/doc/medical/)

北卡罗林纳大学的 Neal F. Cariello 等人维护着包括 p53 在内的四个突变数据库：

R-327 人类 **p53** 基因突变库及软件，可从以下网址下载：

http://metalab.unc.edu/dnam/des_p53.htm

R-328 人类 **hprt** 即次黄嘌呤鸟嘌呤磷酸核糖基转移酶 (hypoxanthine guanine phosphoribosyl transferase) 基因突变数据库和在 PC 视窗下运行的分析软件。其新版包含 2 500 多突变，须向作者订阅：

mailto: cariello@sunsite.unc.edu

但在以下网址的较旧版本可自由下载：

http://metalab.unc.edu/dnam/des_hprt.htm

R-329 转基因啮齿动物 **LacI** 数据库，可从以下网址下载：

http://metalab.unc.edu/dnam/des_laci.htm

R-330 转基因啮齿动物 **LacZ** 突变库，可从以下网址下载：

http://metalab.unc.edu/dnam/des_lacz.htm

以上四个数据库及相应软件的较近描述见：

R-331 N. F. Cariello 等 6 位作者， *Nucleic Acids Res.* **26** (1998) 198.

R-332 **WT1**，基因突变数据库及分析软件。在人类染色体 11p13 区域的 WT1 基因，编码一种含锌指结构的转录因子，后者与胚性癌肉瘤 (维尔姆斯瘤， Wilms' tumor) 有关。数据库和软件的描述见：

C. Jeanpierre, C. Baroud, P. Niaudet, and C. Junien, *Nucleic Acids Res.* **26** (1998) 271 – 274.

需要此库者应与引文第一作者联系：

mailto: jeanpierre@necker.fr

R-333 **WRN** 基因突变与遗传病 Werner 综合征有关，它导致少年早衰。WRN 基因突变、多态性和文献均收录在此网址：

http://www.pathology.washington.edu/werner/ws-wrn.html

R-334 **LDL**，人类 LDL 受体基因突变数据库和分析软件。描述见：

M. Varret 等 6 位作者， *Nucleic Acids Res.* **25** (1997) 172 – 181.

R-335 **OMIM**，在线人类孟德尔遗传 (Online Mendelian Inheritance in Man) 数据库，是从 1963 年开始的使用计算机管理的库发展起来的网络数据库。它搜集人类正常基因和基因失常的信息，除电子数据库

外，每隔几年还印刷成书：

V. A. McKusick, *Mendelian Inheritance in Man*, John Hopkins University Press, 1966, 1968, 1971, 1975, 1978, 1983, 1988, 1990, 1992, 1994, 1998.

网址：

http://www3.ncbi.nlm.nih.gov/omim/

北京大学生物信息中心 [R-166] 有镜象。

R-336 **STACK**，南非国家生物信息中心 SANBI [R-154] 维护的一个序列标记联配和代表序列 知识库 (Sequence Tag Alignment and Consensus Knowledgebase)。其目的是通过广泛处理已 知的 EST 片段，尽可能地提取人类基因组中已表达基因的序列，对每个基因提供一组仔细拼接起来的代表序列。 2000 年初 STACK 库中有 94 000 条 3′ 端序列。这里还有一个软件工具 stackPACK 和供学术界做标定用的人类 EST 序列文件 benchmark10000.seq 。所有学术性单位都可以自由下载这些软件和文件．网址：

http://www.sanbi.ac.za/Dbases.html

R-337 **SANIGENE** 是与 STACK [R-336] 密切相关的一个数据库，它包含所有经过计算机处理联配过的人类基因 EST 的集团，每个集团中的序列至少带有两个重叠的 EST 以便形成代表序列．代表序列的质量要求是至少应有 99% 的残基匹配。 SANIGENE 库中没有单个的 EST ．请参阅 STACK [R-336] 数据库的网址．

§4.10 其他物种基因组数据库

本节重点是各类物种的完全基因组或完整染色体序列的数据库．各个物种基因组的大小，可以用实验方法粗估．相应数据可以在下面的 DOGS 数据库中查到：

R-338 **DOGS**，基因组尺寸数据库 (Database Of Genome Sizes)．网址：

http://www.cbs.dtu.dk

应当特别指出，美国 GenBank [R-212] 的 /genomes/ 子目录从 1999 年 10 月起，做了大幅度的扩充．目前已开辟了人、家鼠、果蝇、线虫、

酵母、细菌、病毒、细胞器等多个子目录，专门搜集完全基因组、完整染色体以及其他长 DNA 序列. 北京大学生物信息中心 [R-166] 已经备有此目录的副本，见 [R-339] .

R-339 GenBank [R-212] 的 /genomes/ 子目录:
ftp://ftp.cbi.pku.edu.cn (/pub/databases/genbank/genomes/)

关于真核生物基因的综合知识，请参看印第安那大学的 euGenes 数据库:

R-340 euGenes，真核生物基因综合知识库，目前包括果蝇、人、小鼠、拟南芥、线虫、酵母和斑马鱼的数据. 网址:
http://iubio.bio.indiana.edu/eugenes/)

4.10.1 原核生物基因组

原核生物的基因组测序，集中在病原和模式生物. 截至 2000 年 8 月初, 已经有 32 个完全基因组数据保存在 GenBank 的 /genomes/bacteria/ 子目录下. 这些完全基因组的大小和由计算机预测的蛋白质或开放读框数目列举在表 4.7 中. 此外，还有十几个基因组已经完成测序，正在进行注释，正式发表之日相应数据库就会对公众开放，70 个基因组正在进行测序. 细菌基因组计划的进展情况，可随时从以下网址查询:

R-341 http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html

对于许多尚未完成测序和公开发表的细菌基因组，通常也可以从有关测序中心获取不带注释的原始序列. 为此可参看 EBI [R-131] 的每周更新的基因组测序进展表 MOT[R-342] 和日本 DDBJ 的 GIB [R-343] :

R-342 **MOT**，欧洲生物信息研究所 EBI [R-131] 的基因组测序进展表 (Genome Monitoring Table)，每周更新. 网址:
http://www.ebi.ac.uk/~sterk/genome-MOT/

R-343 **GIB**，日本 DDBJ [R-213] 设立的 Genome Information Broker for microbial genomes 的缩写. 这是按物种组织的微生物基因组信息网页. 网址:
http://mol.genes.nig.ac.jp/gib/

R-344 MAGPIE 测序计划清单也可以参考. 它比较详细，但更新不够及时. 网址:

http://www-fp.mcs.anl.gov/~gaasterland/genomes.html

法国的 EMGLib 数据库也反映一些微生物基因组的综合数据。

R-345 **EMGLib**，增补微生物基因组库 (Enhanced Microbial Genomes Library)，它反映细菌和酵母的完全基因组。详见：

G. Perriere, P. Bessieres, and B. Labedan, *Nucleic Acids Res.* **28** (2000) 68 - 71.

网址：

http://pbil.univ-lyon1.fr/emglib/emglib.html

美国能源部支持的微生物基因组计划，在完成了最初确定的生殖道支原体、甲烷球菌、热自养甲烷菌和闪烁古生球菌之后，又增加了与全球二氧化碳循环有关的四个细菌。

下面从模式细菌开始，介绍一批各个物种的基因组或基因图谱数据库。

大肠杆菌 (*Escherichia coli* [R-92]) 是研究得最多的模式生物。下面列举一些与它有关的数据库。

R-346 大肠杆菌 K12 菌株的完全基因组序列，可由 GenBank 的子目录 /genomes/ [R-339] 获取，或从华盛顿大学大肠杆菌基因组中心，即 Blattner 实验室的网页读取：

http://www.genetics.wisc.edu/pub/sequence/

K-12 菌株完全基因组的报告见：

F. R. Blattner 等 17 位作者， *Science* **277** (1997) 1453 - 1462.

日本 DDBJ [R-213] 中有大肠杆菌另一个菌株的完全基因组序列：

ftp://monet.genes.nig.ac.jp (/data/ecoli/4.64M.seq.Z)

R-347 **ECDC**，大肠杆菌菌株 K12 的基因序列库、包括基因、读框、调控区、启动子、终止子、 tRNA 和 rRNA 等。描述见：

R. Wahl, and M. Kroeger, *Microbiol. Res.* **150** (1995) 7 -61.

另一个类似的库 **ECD** 已被 ECDC 取代， ECDC 的网址：

http://susi.bio.uni-giessen.de/ecdc/ecdc.html

ftp://ftp.ebi.ac.uk (/pub/databases/ecdc)

R-348 **EcoGene** 和 **EcoWeb**，大肠杆菌 K12 菌株基因组数据库，包括基因、蛋白质、基因间区域，以及蛋白质组信息。事实上，它已经发展

表 4.7 公开数据库中的细菌完全基因组(带 * 号者为古细菌)

名称	拉丁名	碱基数	ORF 数
	*Aeropyrum pernix**	1 669 695	2 694
产液菌	*Aquifex aeolicus*	1 551 335	1 522
闪烁古生球菌	*Archaeoglobus fulgidus**	2 178 400	2 407
枯草芽孢杆菌	*Bacillus subtilis*	4 214 814	4 100
布氏疏螺旋体	*Borrelia burgdorferi*	910 724	850
空肠弯曲杆菌	*Campylobacter jejuni*	1 641 481	1 654
肺炎衣原体	*Chlamydia pneumoniae* CWL029	1 230 230	1 052
肺炎衣原体	*Chlamydia pneumoniae* AR39	1 229 853	997
肺炎衣原体	*Chlamydia pneumoniae* J138	1 228 267	1 017
衣原体	*Chlamydia muridarum*	1 069 412	818
砂眼衣原体	*Chlamydia trachomatis*	1 042 519	894
耐放射微球菌	*Deinococcus radiodurans*	2 648 638	2 580
大肠杆菌	*Escherichia coli*	4 639 221	4 289
流感嗜血菌	*Haemophilus influenzae*	1 830 138	1 709
幽门螺杆菌	*Helicobacter pylori* 26695	1 667 867	1 566
幽门螺杆菌	*Helicobacter pylori* J99	1 643 831	1 491
热自养甲烷杆菌	*M. thermoautotrophicum**	1 751 377	1 869
詹氏甲烷球菌	*Methanococcus jannaschii**	1 664 970	1 715
结核分枝杆菌	*Mycobacterium tuberculosis*	4 411 529	3 918
生殖道支原体	*Mycoplasma genitalium*	580 073	467
肺炎支原体	*Mycoplasma pneumoniae*	816 394	677
脑膜炎奈瑟氏球菌	*Neisseria meningitidis* MC58	2 272 325	2 025
脑膜炎奈瑟氏球菌	*Neisseria meningitidis* Z2491	2 184 406	2 121
热球菌	*Pyrococcus abyssi**	1 765 118	1 765
热球菌	*Pyrococcus horikoshii**	1 738 505	1 979
普氏立克次氏体	*Rickettsia prowazekii*	1 111 529	834
集胞蓝细菌	*Synechocystis* PCC6803	3 573 470	3 169
海栖热袍菌	*Thermotoga maritima1*	860 725	1 846
梅毒密螺旋体	*Treponema pallidum*	1 138 011	1 031
解脲尿支原体	*Ureaplasma urealyticum*	751 719	611
霍乱弧菌	*Vibrio cholerae* El Tor N16961	4 033 460	3 885
苛养木杆菌	*Xylella fastidiosa*	2 679 305	2 904

成一个名为 EcoWeb 的专门网页，把信息、文献和链接集成为一体。请参看：

K. E. Rudd, *Nucleic Acids Res.* **28** (2000) 60 - 64.

网址：

`http://bmb.med.miami.edu/EcoGene/EcoWeb/`

此外，还有 GenProtEc 数据库，包含大肠杆菌的基因组和蛋白质组，并有详细的与序列有关的蛋白质家族清单。网址：

`http://genprotec.mbl.edu/`

R 349 **RegulonDB**，大肠杆菌转录调控和操作子数据库。其 3.0 版描述见：

H. Salgado 等 6 位作者， *Nucleic Acids Res.* **28** (2000) 65 - 67.

网址：

`http://www.cifn.unam.mx/Computational_Biology/regulondb/`

下面列举一些其他细菌的基因组数据库。

R 350 **NRSub**，非冗余枯草芽孢杆菌 DNA 数据库，包括完全基因组、密码子使用表、基因图谱和基因家族，有对 SWISS-PROT [R-401]、ENZYME [R-415]、HOBACGEN [R-421] 等数据库的交叉引用。所谓非冗余即剔除了重复序列。详见：

G. Pirriese 等， *Nucleic Acids Res.* **26** (1998) 60 - 62.

此库的原始网址在法国里昂大学的 PBIL [R-150]：

`http://acnuc.univ-lyon1.fr/nrsub/nrsub.html`

`ftp://biom3.univ-lyon1.fr (/pub/nrsub)`

许多大的生物信息中心设有镜象点。例如日本镜象点在：

`http://ddbjs4h.genes.nig.ac.jp/`

`ftp://ftp.nig.ac.jp (/pub/db/nrsub)`

R 351 **HIDB**，流感嗜血菌完全基因组的原始数据库。描述见：

R. D. Fleischmann 等， *Science* **269** (1995) 496 - 512.

网址：

`ftp://ftp.tigr.org/pub/data/h_influenzae`

`http://www.tigr.org:80/tdb/mdb/hidb/hidb.html`

R 352 **HIDC**，流感嗜血菌基因序列库，其组织方式与大肠杆菌的 ECDC [R 347] 库相似。网址：

http://susi.bio.uni-giessen.de/ecdc/hidc.html

R-353 **CyanoBase**，蓝细菌数据库，实际上是集胞蓝细菌 (*Synechocystis* sp. PCC6803) 的基因组数据库。蓝细菌具有氧化和光合作用所需的全套基因，这一菌株的完全基因组已在 1996 年测定，请参看：

Y. Nakamura, T. Kaneko, and S. Tabata, *Nucleic Acids Res.* **28** (2000) 72.

网址：

http://www.kazusa.or.jp/cyano/cyano.html

R-354 **MJDB**，詹氏甲烷球菌基因组数据库。其完全基因组描述见：

C. J. Bult 等， *Science* **273** (1996) 1058 - 1073.

网址：

ftp://ftp.tigr.org (/pub/data/m_jannaschii)

http://www.tigr.org/tdb/mdb/mjdb/mjdb.html

R-355 **MycDB**，分枝杆菌数据库。这是由世界卫生组织 WHO 等支持的一个交互式数据库，其中最重要的部分涉及麻风分枝杆菌 (*Mycobacterium leprae*) 和结核分枝杆菌 (*Mycobacterium tuberculosis*)。库的描述见：

S. Gergh, and S. T. Cole, *Mol. Microbiol.* **12** (1994) 517 - 534.

网址：

http://www.biochem.kth.se/MycDB.html

R-356 **RsGDB**，类球红细菌 (*Rhodobacter sphaeroides*) 基因组数据库。类球红细菌有两个环形染色体，大者 CI 约有 3Mbp，小者 CII 约有 0.9Mbp。这是 CII 的数据库，请参看：

M. Choudhary, C. Mackenzie, N. J. Mouncey, and S. Kaplan, *Nucleic Acids Res.* **27** (1999) 61 - 62.

文中所给网址很难进入，有必要时请与上文第一作者联系：

mailto:/ madhu@utmmg.med.uth.tmc.edu

R-357 **PGI**，疫霉属基因预研究计划 (Phytophthora Genome Initiative) 的数据库。这个由 NCGR [R-135] 支持的项目，研究破坏性很大的植物病原，即卵菌纲 (Oomycetes) 的疫霉属的基因与演化，目的在于了解其感染和抗性机理。目前在做 *Phytophthora infestans* 和 *Phytophthora sojae* 的 EST 和后者的 BAC 库的测序。详见：

M. Waugh 等 8 位作者， *Nucleic Acids Res.* **28** (2000) 87 - 90.

网址：

`http://www.ncgr.org/pgi/`

4.10.2 真菌基因组

真菌 (fungi) 界的基因组，首先是与模式生物酿酒酵母 (*Saccharomyces cerevisiae*) 有关的一批数据库，例如：

R-358 **SGD**，酿酒酵母基因组数据库，它把功能基因组学信息集成到数据库中，参看：

C. A. Ball 等 17 位作者， *Nucleic Acids Res.* **28** (2000) 77 - 80.

网址：

`http://genome-www.stanford.edu/Saccharomyces/`

`ftp://genome-ftp.stanford.edu (/pub/yeast)`

R-359 **LISTA, LISTA-HOP 和 LISTA-HON** 是酿酒酵母基因组中蛋白质编码序列及其同源性的数据库，详见：

R. Dolz 等 5 位作者， *Nucleic Acids Res.* **24** (1996) 50 - 52.

网址：

`http://www.ch.embnet.org/`

`ftp://bioftp.unibas.ch`

请注意，这是酵母完全基因组测定之前形成的数据库。

R-360 **MYGD**，酵母基因组、蛋白质和同源关系的数据库，描述见：

H. W. Mewes 等 12 位作者， *Nucleic Acids Res.* **28** (2000) 37 - 40.

网址：

`http://www.mips.biochem.mpg.de/proj/yeast/`

R-361 **YIDB**，酵母内含子数据库，详见：

P. J. Lopez, and B. Seraphin, *Nucleic Acids Res.* **28** (2000) 85 - 86.

网址：

`http://www.EMBL-Heidelberg.DE/`
`ExternalInfo/seraphin/yidb.html`

此外，还请参看酵母蛋白质组数据库 YPD [R-499]、酵母基因功能数据库 TRIPLES[R-508] 等。关于其他真菌，可参看：

R-362 **MNCDB**，由德国 MIPS 所维护的粗糙链孢霉 (*Neurospora crassa*) 基因组数据库。其基因组总长约 4 300 万碱基对，MIPS 负责其 7 个染色体中的第 II 和第 V 号。网址：

`http://www.mips.biochem.mpg.de/desc/neurospora/`

R-363 真菌基因组资源的网址：

`http://fungus.genetics.uga.edu:5080/main.html`

这里有指向念珠菌 (*Candida*)、链孢霉 (*Neurospora*) 和肺囊虫 (*Pneumocystis*) 基因组计划的链接，不再一一列出。

R-364 **FGSC**，真菌遗传学信息中心 (Fungal Genetics Stock Center)。其网址是：

`http://www.fgsc.net/`

4.10.3 原生生物和线虫基因组

关于原生生物基因组测序的进展，可以参看：

R-365 欧洲生物信息研究所 EBI [R-131] 的原生生物网页：

`http://www.ebi.ac.uk/Projects/Protozoa/`

R-366 人恶性疟原虫 (*Plasmodium falciparum*) 的染色体，现已测完第 II 号和第 III 号，其描述分别见：

M. J. Gardner 等， *Science* **282** (1998) 1126 – 1132. GenBank 编号 AE001362。

S. Bowman 等， *Nature* **400** (1999) 572.

顺便提一下，蚊子的基因图谱数据库可参看 MsqDB [R-375]。

与秀丽线虫 [R-94] 有关的数据库，除了前面已经提到过的 Intronerator [R-243] 和后面还要介绍的 WormPD [R-500] 等，请特别注意：

R-367 **ACeDB**，线虫综合数据库。它的原始库在 Sanger 中心 [R-299]，但可从许多其他网点读取：

`ftp://sanger.ac.uk (/pub/acedb)`

`ftp://ncbi.nlm.nih.gov (/repository/acedb)`

`ftp://lirmm.lirmm.fr (/pub/acedb)`

应当特别指出，ACeDB 数据库本身基于面向对象的程序设计 (OOP) [R-51] 思想，可以从网络上自由下载。目前许多研究单位用它建立自

已的数据库，详见第 5 章的介绍 [R-851]。

R-368 关于线虫发育特别是化学感觉神经的研究，可以参阅 C. Bargmann 实验室的网页：

http://devbio-mac1.ucsf.edu/

4.10.4 昆虫基因组

果蝇 (*Drosophila melanogaster*) 的研究，近一个世纪以来对遗传学的发展始终起着重要作用。全世界果蝇研究者大约有 6 000 人。果蝇的全基因组有 1.8 亿碱基对，其富含基因的常染色质 (euchromatin) 部分，计 1.2 亿碱基对，已经由 Celera Genomics 公司 [R-798] 为主的协作组基本上测定，并于 2000 年 3 月 24 日发表在美国《科学》周刊的果蝇专号上。果蝇有 13 600 个基因，比线虫略少。

R-369 M. D. Adams 等 34 个单位的 195 位作者，"The genome sequence of *Drosophila melanogaster*", *Science* **287** (2000) 2185 - 2195.

相应数据已送交 GenBank[R-212]，索取号为 AE002566 - AE003403。

下面再列举一些与果蝇有关的数据库或研究中心：

R-370 斯坦福大学的果蝇基因组中心，已经独立出来。它的网址是：

http://www.fruitfly.org/

R-371 **FlyBase**，果蝇基因和分子数据库，由国际果蝇协作组织维护，描述见：

The Flybase Consortium, *Nucleic Acids Res.* **27** (1999) 85 - 88.

实际上现在已经发展出一个名为 The Interactive Fly 的网页，涵盖果蝇基因，胚胎、组织和器官发育，生化和发育途径等各方面的信息。

网址：

http://flybase.bio.indiana.edu/

ftp://flybase.bio.indiana.edu/

在各主要国际生物信息中心均有镜象。

R-372 **FlyNets**，果蝇分子和遗传相互作用数据库。请参看：

C. Sanchez 等 8 位作者，*Nucleic Acids Res.* **27** (1999) 89 - 94.

网址：

http://gifts.univ-mrs.fr/FlyNets/

R-373 **GIF-DB**，果蝇胚胎发育过程中基因相互作用的 WWW 数据库，其格式与 EMBL [R-211] 库类似。请参看：

E. Mohr 等 8 位作者， *Nucleic Acids Res.* **26** (1998) 89 – 93.

相应 GIFTS 服务器的网址：

`http://www-biol.univ-mrs.fr/~lgpd/GIFTS_home_page.html`

R-374 哈佛大学的果蝇网页：

`http://morgan.harvard.edu/`

果蝇以外其他昆虫的基因图谱数据库，我们只指出：

R-375 **MsqDB**，蚊子基因数据库，包括多种蚊子的遗传和物理图谱。网址：

`http://klab.agsci.colostate.edu/acedb/MsqDB-acedb.html`

`ftp://klab.agsci.colostate.edu`

4.10.5 鱼类数据库

模式生物斑马鱼 的研究信息和数据库，可访问以下网址：

R-376 美国国家卫生署 (NIH) 1997 年建立的斑马鱼网页：

`http://www.nih.gov/science/models/zebrafish/`

这里有一批与斑马鱼信息资源有关的链接。

R-377 **ZFIN**，斑马鱼基因组、发育突变和野生种系数据库。网址：

`http://zfish.uoregon.edu/ZFIN/`

注意：此网页只能用 Netscape 3.0 [R-68] 以上的浏览器访问，Internet Explorer [R-69] 不能正确工作。

R-378 **Fugu** 是河豚 (*Fugu rubripes*) 的简称，英文又叫 Puffer fish 。它的基因组大小只有人的七分之一，但基因数目与人相近，因此也被列为与人类基因组计划有关的模式生物。其数据库网址：

`http://fugu.hgmp.mrc.ac.uk/`

4.10.6 啮齿动物基因组

家鼠 (*Mus musculus*) 的 DNA 序列长度和基因总数都与人类相近，可以通过基因剔除等实验增进对人类基因的认识。因此，家鼠基因组计划与人类基因组计划密切相关。原来预计家鼠的完全基因组将在 2008 年测

出，看来可能大为提前。下面列举一些与家鼠有关的数据库。

R-379 **M. Musculus 基因组库**。GenBank 已经在 1999 年 10 月底在基因组目录下建立了家鼠子目录，网址：

ftp://ncbi.nlm.nih.gov (/genbank/genomes/M_muslulus)

这个子目录中的文件按染色体编号。

R-380 **MGD**，家鼠基因组库，现在又称 MGI 即家鼠基因组信息库 (Mouse Genome Informatics)，并且正在成为家鼠集成数据库 MGEIR [R-509] 的组成部分，它包含实验室中培育的家鼠的遗传和基因、图谱和文献信息，还有到其他哺乳类数据库的链接。详见：

J. A. Blake, J. T. Eppig, J. E. Richardson, M. T. Davisson, 以及家鼠基因组数据库小组， *Nucleic Acids Res.* **28** (2000) 108 - 111.

网址：

http://www.informatics.jax.org/mgd.html/

ftp://ftp.informatics.jax.org/

在英国、法国和日本设有镜象点。

R-381 **Cre 转基因家鼠系的数据库**。Cre 重组酶由大肠杆菌噬菌体 P1 的 Cre 基因编码，是基因靶位操作的一种工具。可参看：

汪亚平、朱作言，“基因靶位操作的原理与策略”，《遗传》 **21** (1999) 第 3 期；

http://www.chinainfo.gov.cn/periodical/yc/yc9903/990314.htm

加拿大 Nagy 实验室的 Cre 转基因数据库的网址：

http://www.mshri.on.ca/nagy/cre.htm

R-382 **RatMap**，大鼠基因图谱数据库，包含大鼠染色体基因和 DNA 标记、与小鼠和人的同源关系等。网址：

http://ratmap.gen.gu.se/

4.10.7 细胞器数据库

细胞器数据库目前主要搜集关于线粒体和叶绿体基因的数据。

R-383 **MitoNuc 和 MitoAln** 是关于编码线粒体蛋白的细胞核基因的两个相互关联的数据库。请参看：

G. Pesole 等 8 位作者， *Nucleic Acids Res.* **28** (2000) 163 - 165.

网址：

http://bio-www.ba.cnr.it:8000/srs6/

R-384 **GOBASE**，细胞器基因组数据库，目前数据集中在线粒体基因组，下一步将扩展到叶绿体以及被认为与线粒体和叶绿体的共同祖先有关的细菌．请参看：

M. Korab-Laskowska 等 7 位作者， *Nucleic Acids Res.* **26** (1998) 138 - 144.

网址：

http://megasun.bch.umontreal.ca/gobase/

R-385 **MitBASE**，线粒体 DNA 数据库，集成所有已知线粒体基因信息，包括人、动物、植物和微生物，也提供一些检索工具．此库的较近介绍见：

M. Attimonelli 等 22 位作者， *Nucleic Acids Res.* **28** (2000) 148 - 152.

网址：

http://www3.ebi.ac.uk/Research/Mitbase/mitbase.pl/

R-386 人类线粒体数据库：

http://bio-www.ba.cnr.it:8000/Tutorials/MitBASE/

R-387 **MitBASE Pilot**，酵母线粒体中核基因数据库．网址：

http://www3.ebi.ac.uk/Research/Mitbase/mitbase.pl/

R-388 植物和藻类线粒体数据库：

http://www.biologie.uni-ulm.de/bio2/
knoop/mitbase/plant_mt_gene.gif

http://tonic.ebi.ac.uk:8889/mitbase/
plsql/pla_qry.pla_show_qry_opts/

R-389 原生生物线粒体数据库：

http://bio-www.ba.cnr.it:8000/Tutorials/
MitBASE/protist_table.html

R-390 脊椎动物线粒体数据库：

http://bio-www.ba.cnr.it:8000/Tutorials/
MitBASE/vertebrate.html

4.10.8 拟南芥基因组

目前研究得最多的模式植物是拟南芥 (*Arabidopsis thaliana*)，它的基因组总长度约 1.2 亿碱基对，约编码 25 000 个基因，将在 2000 年基本上测序完毕。下面是几个与拟南芥基因组有关的数据库：

R-391 **MATDB**，国际拟南芥基因组计划 (Arabidopsis Genome Initiative，简称 AGI) 的数据汇总。关于此计划请参看：
M. Bevan 等 5 位作者，*Bioessays* **21** (1999) 110 - 120.
数据库网址：
`http://www.mips.biochem.mrg.de/desc/thal/`

R-392 **AtDB**，拟南芥基因组数据库。详见：
S. Y. Rhee 等 7 位作者，*Nucleic Acids Res.* **27** (1999) 79 - 84.
网址：
`http://genome-www.stanford.edu/Arabidopsis/`
`ftp://genome-ftp.stanford.edu` (`/pub/arabidopsis`)
GenBank 在 1999 年底开辟了拟南芥基因组的子目录：
`ftp://ftp.ncbi.nlm.nih.gov` (`/genbank/genomes/A_thaliana/`)
此子目录中现有第 II 号和第 IV 号两个染色体的子目录。关于第 IV 号染色体的描述见下一条目。

R-393 欧洲共同体拟南芥基因组计划组织，以及 M. Reven 等 68 位作者，*Nature* **391** (1998) 485 - 488.

R-394 **DAtA**，拟南芥基因组注释库。详见：
C. J. Palm, N. A. Federspiel, and R. W. Davis, *Nucleic Acids Res.* **28** (2000) 102 - 103.
网址：
`http://luggagefast.Stanford.edu/group/arabprotein/`

R-395 **TAIR**，拟南芥信息资源 (The Arabidopsis Information Resources)，是 NCGR [R-135] 和卡内基研究会 (Carneigie Institution) 在 1999 年 10 月共同建立的拟南芥基因组和文献数据库。网址：
`http://www.arabidopsis.org/`

R-396 **AGR**，拟南芥基因组资源 (Arabidopsis Genome Resource)，是英国 CropNet [R-567] 网上植物生物信息的一部分。网址：

http://synteny.nott.ac.uk/agr/agr.html

ftp://thale.nott.ac.uk (/pub/uk-crop/db/AGR/)

这是 UK-CropNet 的一个镜象，每天当地时间凌晨 3 点更新一次。

R-397 **TIGR-AT**， TIGR [R-156] 研究所的拟南芥 EST 和基因序列数据库。描述见：

S. D. Rounsley 等 7 位作者， *Plant Physiol.* **112** (1996) 1177 - 1183.

网址：

http://www.tigr.org/tdb/at/at.html

ftp://ftp.tigr.org (/pub/data/a_thaliana)

4.10.9 病毒数据库

最后，提几个与病毒有关的数据库：

R-398 **ICTVdB**，病毒数据库。这是国际病毒分类委员会 (International Committee on Taxonomy of Viruses，简称 ICTV) 指导下建立的病毒命名、显微镜照片和基因序列的数据库。原始库在澳大利亚国立大学：

http://life.anu.edu.au/viruses/ICTVdB/ictvdb.html

中国科学院微生物研究所 [R-170] 设有镜象：

http://www1.im.ac.cn/ictvdb/

R-399 **VIDEdB**，病毒鉴定交换数据库 (Virus Identification Data Exchange)。原始库在澳大利亚国立大学：

http://biology.anu.edu.au/research-groups/MES/vide/

中国科学院微生物研究所 [R-170] 设有镜象，网址见 [R-398]。

R-400 **RDV**，水稻矮缩病毒 (Rice Dwarf Virus) 基因组数据库，由北京大学生物信息中心王建民和顾孝诚建立。它不仅包含序列和图谱信息，还有文献目录。网址：

http://www.cbi.pku.edu.cn/rdv/

§4.11 蛋白质序列数据库

最重要的蛋白质氨基酸序列数据库是瑞士的 SWISS-PROT [R-401] 和美、德、日三国合建的国际 PIR 库 [R-404]。

R-401 **SWISS-PROT** 是对数据人工审读很严格的库。可以说，只有实际存在的蛋白质才被收入。每一条数据都有详细注释，包括功能、结构域、翻译后的修饰等，以及齐全的引文和到许多其他数据库的链接。此库的冗余度也较低。一般说，任何蛋白质序列数据的搜寻和比较都应当从 SWISS-PROT 开始。最近的描述见：

A. Bairoch, and R. Apweiler, *Nucleic Acids Res.* **28** (2000) 45 - 48.

网址：

`http://www.expasy.ch/sprot/`

`ftp://ftp.expasy.ch (/databases/swiss-prot/)`

北京大学生物信息中心 [R-166] 有 SWISS-PROT 镜象，可通过检索工具 SRS [R-203] 查询。

R-402 **TrEMBL** 是从 EMBL 库中的核酸序列翻译出来的氨基酸序列，已经完成了自动注释。它又分成两部分： SP-TrEMBL 的条目已由专家人工分类并且赋予了 SWISS-PROT 库的索取号，但是还没有通过人工审读被最终收入 SWISS-PROT ； REM-TrEMBL (REMaining TrEMBL) 包含由于某种原因而还没有被收入 SWISS-PROT 的条目。参看 [R-401] 引文。 1999 年 4 月这个库里有 77 977 条序列。

网址：

`ftp://ftp.ebi.ac.uk (/pub/databases/trembl/)`

`http://www.ebi.ac.uk:5000`

如果想取得 SWISS-PROT 和 TrEMBL 中全部条目的清单，可访问：

`http://www.expasy.ch/sprot/sprot-retrieve-list.html`

取 SWISS-PROT+TrEMBL 非冗余库用：

`ftp://ftp.expasy.ch (/databases/sp_tr_nrdb/)`

北京大学生物信息中心 [R-166] 有镜象，可通过检索工具 SRS [R-203] 查询。

R-403 **TrEMBL-NEW** 是从 EMBL 库中的核酸序列翻译出来的氨基酸序列，但是还没有赋给 SWISS-PROT 索取号，因此只能借助蛋白质标识符检索。

R-404 **PIR** 是蛋白质信息资源 (Protein Information Resource) 的缩写。这是一个国际蛋白质序列数据库，它包含所有序列已知的自然界中野生型蛋白质的信息。此库主要目的是提供按同源性和分类学组织的

综合的、非冗余的数据库。它由美国华盛顿的全国生物医学研究基金会 (National Biomedical Research Foundation，简称 NBRF) 所支持的 PIR、德国马普学会的慕尼黑蛋白质序列信息中心 MIPS [R-139] 和日本的 JIPID [R-138] 共同维护。自 1984 年以来，PIR 库每周更新，每季度发行新版。PIR 内容分为 4 级，其 2000 年 1 月底的 63.03 版的收藏情况见表 4.8，2000 年 1 月 21 日的 63.02 版收入 171 197

表 4.8 PIR 的收藏情况

分级	说明	条目数
PIR1	完全分类清楚	20 049
PIR2	已检查和分类	150 497
PIR3	未检查	781
PIR4	未解码翻译	369
总计		171 696

条蛋白质，共计 59 721 663 个氨基酸残基。PIR 库实际上是 PSD [R-408]、PATCHX [R-409]、ARCHIVE [R-410]、NRL-3D [R-451]、FAMBASE [R-452]、PIR-ALN [R-456]、RESID [R-460]、ProClass [R-411]、ProtFam [R-453] 和 PIR-ASDB [R-412] 等多个数据库的集成和链接。最近的综述见：

W. C. Barker 等 14 位作者，*Nucleic Acids Res.* **28** (2000) 41 – 44.

网址：

`http://www-nbrf.georgetown.edu/pir/`

`http://www.mips.biochem.mpg.de/proj/protseqdb/`

`ftp://nbrf.georgetown.edu (/pir/)`

北京大学生物信息中心 [R-166] 有镜象。

R-405 **GenPept** 是由 GenBank [R-212] 中的 DNA 序列翻译得到的蛋白质序列，与 TrEMBL [R-402] 相似，但没有像后者那样经专家审读。

网址：

`http://www.infobiogen.fr/srs/`

`ftp://ftp.ncifcrf.gov (/pub/genpept)`

`ftp://ftp.infobiogen.fr (/pub/db/genpept)`

ftp://bioinformatics.weizmann.ac.il

访问子目录 /pub/databases/genpept .

R-406 **PROSITE**，由专家根据生物知识审编的 SWISS-PROT [R-401] 蛋白质序列中有生物意义的位点 (sites)、模式 (patterns) 和轮廓 (profiles) 的数据库，包括酶活性位点、辅助因子结合位点、二硫键 S-S 位点等。此库可以帮助确定新的蛋白质序列是否属于已知的家族。用户可用 PrositeScan [R-407] 服务器搜索此库，请参看：

K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch, *Nucleic Acids Res.* **27** (1999) 215 - 219.

网址：

http://www.expasy.ch/prosite/

ftp://ftp.expasy.ch (/databases/prosite)

ftp://ncbi.nlm.nih.gov (/repository/PROSITE)

北京大学生物信息中心 [R-166] 有镜象。

R-407 **PrositeScan** 服务器，根据用户填表提交的蛋白质序列搜索 PROSITE 模式。它接受所有 ReadSeq [R-699] 程序所能转换的序列格式，也可按 SWISS-PROT 的 ID 或 AC 号，GenPept [R-405] 的 GI 号指定序列。网址：

http://www.isrec.isb-sib.ch/software/PSTSCAN_form.html

R-408 **PSD**，蛋白质序列数据库 (Protein Sequence Database)，是 PIR 的主体。描述请参看 PIR [R-404] 的引文。网址：

http://pir.georgetown.edu/pirwww/dbinfo/textpsd.html

R-409 **PATCHX**，PIR 的子库之一，收入尚未纳入 PIR 库的蛋白质序列。请参看 PIR [R-404] 的引文。网址：

http://pir.georgetown.edu/pirwww/dbinfo/patchx.html

R-410 **ARCHIVE**，PIR 的子库之一，保存 PIR [R-404] 库中条目的原始文献或最初提交的序列。请参看 PIR 的引文。网址：

http://pir.georgetown.edu/pirwww/dbinfo/archive.html

R-411 **ProClass**，蛋白质类数据库，是根据 PROSITE 库 [R-406] 和 PIR 库 [R-404] 中超家族的关系组织起来的非冗余蛋白质库，详见：

H. Huang, C. L. Xiao, and C. H. Wu, *Nucleic Acids Res.* **28** (2000) 273 - 276.

网址：

http://pir.georgetown.edu/gsfserver/prolclass.html

ftp://nbrfa.georgetown.edu (/pir/databases/proclass/)

http://diana.uthct.edu/proclass.html

ftp://diana.uthct.edu (tt /pub/ProClass/)

R-412 **PIR-ASDB**，PIR 的注释和相似性数据库。它集中了 PSD [R-408] 中所有相似条目的注释。网址：

http://www-nbrf.georgetown.edu/pir/

R-413 **KIND**，瑞典斯德哥尔摩生物信息中心维护的非冗余蛋白质序列库，可由 KI [R-145] 的 ftp 服务器下载：

ftp://ftp.mbb.ki.se (/pub/KIND)

在介绍 ENZYME 等酶数据库之前，必须说明一下酶的命名系统。国际生物化学和分子生物学联合会下属的命名委员会 (Nomenclature Committee，简称 NC-IUBMB) 赋予每种新刻画清楚的酶一个由 4 个数字组成的号码，称为 EC 号。例如，腺三磷酶 (ATPase) 的 EC 号是 3.6.1.37。酶的发现者应当向 NC-IUBMB 提出申请。联系地址和申请表格可在 ENZYME [R-415] 数据库的使用手册中找到。ENZYME、BRENDA [R-416]、EMP [R-549]、PUMA [R-551]、WIT [R-548]、LIGAND [R-557] 等多种与酶和代谢途径有关的数据库均使用 EC 号。NC-IUBMB 不定期地发表酶的命名，例如：

R-414 NC-IUBMB, *Enzyme Nomenclature*, Academic Press, 1992.

R-415 **ENZYME**，基于命名系统的酶数据库。可按照酶的 EC 号、分类、学名和俗名、化合物、辅助因子等查询。每一个条目下列出所催化的反应和酶的来源、功能等，并有指向其他多种数据库、以及文献库 MEDLINE [R-599] 和代谢途径图 [R-555] 的链接。截至 2000 年 1 月 15 日，ENZYME 第 25 版包含 3705 个条目，详见：

A. Bairoch, *Nucleic Acids Res.* **28** (2000) 304 - 305.

网址：

http://www.expasy.ch/enzyme/

ftp://ftp.expasy.ch (/databases/enzyme)

北京大学生物信息中心 [R-166] 有镜象。

R-416 **BRENDA**，这是一个内容广泛的酶的信息库。网址：

http://www.brenda.uni-koeln.de/

R-417 **OWL**，蛋白质序列库，是由 SWISS-PROT [R-401]，PIR [R-404]，GenBank [R-212] 翻译序列和 PDB [R-441] 等数据库产生的非冗余的蛋白质序列库。网址：

http://bmbsgi11.leeds.ac.uk/bmb5dp/owl.html

ftp://ftp.hgmp.mrc.ac.uk (/pub/database/owl)

ftp://bmbsgi11.leeds.ac.uk (/pub/owl/)

ftp://ncbi.nlm.nih.gov (/repository/OWL)

北京大学生物信息中心 [R-166] 有镜象。

R-418 **GeneCards**，由以色列魏茨曼科学研究所 [R-164] 维护的关于基因及其产物，以及它们的生物医学应用的文献库。描述见：

M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, *Bioinformatics* **14** (1998) 656 - 664.

网址：

http://bioinfo.weizmann.ac.il/cards

中国镜象在北京医科大学 [R-169]。

R-419 **SWISS-2DPAGE**，由二维聚丙烯酰胺凝胶电泳 (PolyacrylAmide Gel Electrophoresis，简称 PAGE) 所确定的蛋白质的参考图谱数据库，包括文本和图象信息，通向其他 2D-PAGE 数据库的链接等。最近描述见：

C. Hoogland 等 7 位作者，*Nucleic Acids Res.* **28** (2000) 286 - 288.

网址：

http://www.expasy.ch/ch2d/

ftp://www.expasy.ch (/databases/swiss-2dpage/)

商业性用户使用此库，须付费取得许可。北京大学生物信息中心有其镜象：

http://expasy.pku.edu.cn/

ftp://expaxy.pku.edu.cn

R-420 **HDB**，组蛋白数据库，包括联配好的组蛋白序列以及已确认包含有组蛋白折叠模体的非组蛋白序列，以及所有已知组蛋白和组蛋白折叠的结构，同时指出不同数据库中类似序列的差异。请参看：

S. A. Sullivan 等 5 位作者，*Nucleic Acids Res.* **28** (2000) 320 - 322.

网址：

http://genome.nhgri.nih.gov/histones/

ftp://ncbi.nlm.nih.gov (/pub/baxevanis/histones)

R-421 **HOBACGEN** 数据库，包含按家族组织的所有细菌的蛋白质序列，有助于从各种细菌选取同源家族，作多序列联配和构建亲缘树。

网址：

http://pbil.univ-lyon1.fr/databases/hobacgen.html

R-422 **MITOP**，线粒体蛋白质组数据库，包括与线粒体有关的基因、蛋白质和疾病信息。请参看：

C. Scharfe 等 12 位作者，*Nucleic Acids Res.* **28** (2000) 155 – 158.

网址：

http://www.mips.biochem.mpg.de/proj/medgen/mitop/

R-423 **MITOMAP**，人类线粒体基因组数据库，请参看：

A. M. Kogelnik 等 5 位作者，*Nucleic Acids Res.* **26** (1998) 112 – 115.

网址：

http://www.gen.emory.edu/mitomap.html

R-424 **REBASE**，限制性内切酶和甲基化酶数据库，包括它们的识别位点、剪切位点、甲基化特异性、由哪些微生物分离得到，以及供应商和文献。请参看：

R. J. Roberts, and D. Macelis, *Nucleic Acids Res.* **28** (2000) 306 – 307.

网址：

http://www.neb.com/rebase

ftp://www.neb.com (/pub/rebase)

北京大学生物信息中心 [R-166] 有镜象。

R-425 **ProtoMap**，蛋白质分类数据库。这是对 SWISS-PROT [R-401] 数据库中的全部蛋白质由计算机自动进行层次分类、把相关者聚集分组所得到的数据库。所列出的许多分组与自然的蛋白质家族和超家族有关联。它有助于对已知蛋白质家族做更细致的划分，并阐明家族之间的关系。这个网点提供交互式、视象化的分析工具，以便显示有关的大量数据和在蛋白质图谱中“导航”。请参看：

G. Yona, N. Linial, and M. Linial, *Nucleic Acids Res.* **28** (2000) 49 – 55.

网址：

http://www.protomap.cs.huji.ac.il/

R-426 **ISSD**，蛋白质序列数据库，其每个条目包含一个基因的编码序列，同相应的氨基酸序列对比，并给出相应多肽链的结构数据．核苷酸序列取自 GenBank [R-212]，结构参数取自 PDB [R-441]，包括多肽骨架原子坐标、二面角，还有 DSSP 程序所预测的二级结构． ISSD 的 2.0 版描述见：

I. A. Adzhubei, and A. A. Adzhubei, *Nucleic Acids Res.* **27** (1999) 268 - 271.

网址：

http://www.protein.bio.msu.su/issd/

R-427 **PRF**，日本蛋白质研究基金会 (Protein Research Foundation) 维护着三个蛋白质和多肽数据库：PRF/LITDB 文献库、PRF/SEQDB 序列库及 PRF/SYNDB 合成产物库．它们的特点是包括了一些不到 50 个氨基酸残基的多肽链和一些人工合成的、非天然的产物。网址：

http://prfsun2.prf.or.jp/

R-428 **MEROPS**，肽酶数据库．它提供所有肽酶，即蛋白质水解酶的目录和基于结构的分类．这是很大的一群蛋白质，占基因总产物约 2%，在医学和生物技术中有重要作用．通过肽酶名称索引，可以访问名为 PepCards 的一组文件，每个文件提供一种肽酶的分类和命名信息、蛋白质和核酸序列、三级结构，以及通向其他人类遗传数据库中有关条目的界面． PepCards 的另一个索引可按物种名称查找其全部已知肽酶．库中肽酶根据其对活性起主要作用的“肽酶单元”部分序列的统计置信的相似性分成家族 (families)．有共同演化来源并且预期有类似的三级折叠的家族，再归并到一起称为宗族 (clans)． MEROPS 中另有名为 FamCards 和 ClanCards 的两组文件．每个 FamCard 文件含有通向其他数据库的链接，以便查找序列模体 和二、三级结构，并显示该家族在主要生物界中的分布情况。请参看：

N. D. Rawlings, and A. J. Barrett, *Nucleic Acids Res.* **28** (2000) 323 - 325.

网址：

http://www.bi.bbsrc.ac.uk/Merops/Merops.htm

R-429 **PKR**，蛋白激酶信息库 (Protein Kinase Resource)。描述见：

M. Gribskov, P. Bourne, and C. M. Smith, in [R-19] (1999) 241 - 246.

网址：

`http://www.sdsc.edu/Kinases/pkr/pk_catalytic/pk_cat_list.html`

`http://www.sdsc.edu/Kinases/pkr/pk_structure.html#Analysis`

`http://www.sdsc.edu/pb/Software.html`

R-430 **Wnt** 基因网页。Wnt 蛋白质是高度保守的分泌性的信号分子家族，对调控胚胎发育中的细胞相互作用有重要意义。其名称来自家鼠的 int-1 基因和果蝇的 wingless 基因。本网页总结了 Wnt 在从非洲爪蟾到脊椎动物和人类的基因型和表现型数据。网址：

`http://vonbaer.ana.ed.ac.uk/rnusse/wntwindow.html`

`http://www.stanford.edu/~rnusse/wntwindow.html`

另外，请参看信号转导知识环境 STKE [R-852] 中关于 Wnt 途径的信息。

R-431 **PhosphoBase**，磷酸化位点数据库。除了直接检索，还可用来预测给定序列包含何种激酶磷酸化位点。其 2.0 版描述见：

A. Kreegipuu, N. Blom, and S. Brunak, *Nucleic Acids Res.* **27** (1999) 237 - 239.

网址：

`http://www.cbs.dtu.dk/databases/PhosphoBase/`

R-432 **SYSTERS**，蛋白质集团数据库，它使用系统重复搜寻方法 (SYSTEmatic Re-Searching) 构建。此方法描述见：

A. Krause, and M. Vingron, *Bioinformatics* **14** (1998) 430 - 438.

数据库近况请参看：

A. Krause, J. Stoye, and M. Vingron, *Nucleic Acids Res.* **28** (2000) 270 - 272.

网址：

`http://www.dkfz-heidelberg.de/tbi/services/cluster/`

R-433 **DIP**，蛋白质相互作用数据库。描述见：

I. Xenarios 等 6 位作者，*Nucleic Acids Res.* **28** (2000) 289 - 291.

网址：

http://URLdip.doe-mbi.ucla.edu/

R-434 **DExH/D** 数据库。DExH/D 蛋白质对 RNA 代谢和加工有多方面的重要作用。请参看：

E. Jankowsky, and A. Jankowsky, *Nucleic Acids Res.* **28** (2000) 333 - 334.

网址：

http://www.columbia.edu/~ej67/dbhome.htm

R-435 **Homeodomain**，同源异形结构域数据库。由同源异形盒 (参看 [R-233]) 编码的蛋白质结构域，构成一个大的蛋白质家族。此库搜集其序列、结构和基因组信息。请参看：

S. Banerjee-Basu, J. F. Ryan, and A. D. Baxevanis, *Nucleic Acids Res.* **28** (2000) 329 - 330.

网址：

http://genome.nhgri.gov/homeodomain/

R-436 **InBase**，新英格兰生物实验公司 (New England BioLab，简称 NEB) 的蛋白质剪接数据库。请参看：

F. B. Perler, *Nucleic Acids Res.* **28** (2000) 344 - 345.

网址：

http://www.neb.com/neb/inteins.html

R-437 **LGICdb**，配体门控离子通道数据库 (Ligand Gated Ion Channel database)。它包含从细胞外激活的通道受体的基因、RNA 和蛋白质序列。来自其他大数据库的信息已经再处理过，以减少冗余。此库也包含多序列联配、亲缘关系、原子坐标 (PDB [R-441] 格式) 的数据。请参看：

N. Le Novée, and J. C. Changeux, *Nucleic Acids Res.* **27** (1999) 340 - 342.

网址：

http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html

R-438 **SENTRA**，信号传递蛋白质数据库。请参看：

M. D'Souza, M. F. Romine, and N. Maltsev, *Nucleic Acids Res.* **28** (2000) 335 - 336.

网址：

http://wit.mcs.anl.gov/WIT2/Sentra/

R-439 **ICN**，离子通道网络 (Ion Channel Network)，是由美国神经科学数据库中心等单位联合建立的一个内容丰富的网页。网址：

http://pain.med.umn.edu/csn/

R-440 **AAindex**，氨基酸索引数据库。它包含 20 种氨基酸的各种物理化学和生物学参数的数值，以及序列联配用的各种置换矩阵，例如 PAM [R-619] 和 BLOSUM [R-620] 矩阵。请参看：

S. Kawashima, and M. Kanehisa, *Nucleic Acids Res.* **28** (2000) 374.

网址：

http://www.genome.ad.jp/aaindex/

ftp://ftp.genome.ad.jp (/db/genomenet/aaindex/)

北京大学生物信息中心 [R-166] 有镜象。

§4.12 蛋白质结构和分类数据库

我们在 3.5.6 小节中已经提到蛋白质结构的几个层次，讨论了一些尚未统一的名词术语的译法。简而言之，一级结构是氨基酸的排列顺序，即 §4.11 节中的蛋白质序列。二级结构主要是由氢键维持的 α 螺旋和 β 片。三级结构是完全折叠好的蛋白质的空间结构。四级结构是多个蛋白质亚基组成蛋白质复合体的结构。在最细的层次，由 X 射线衍射和核磁共振 (NMR) 等实验方法确定的蛋白质中原子的三维坐标，构成 PDB [R-441] 这样的蛋白质结构数据库的主要内容。二级结构和三级结构之间的模体 (motif)、结构域 (domain) 和“折叠” 或“折叠单元” (fold)，对于蛋白质结构的分类和预测有重要作用。

R-441 **PDB**，蛋白质结构数据库 (Protein Data Bank)。1971 年建立于美国布鲁克海文国家实验室 [R-163]，当时只有 7 个结构。它搜集由 X 射线衍射和核磁共振实验测定的生物大分子三维结构数据。从 1998 年 10 月 1 日起 PDB 的管理交给 RCSB[R-442]。2000 年 6 月 7 日 PDB 库中有 12 474 个条目。关于 PDB 库的较近介绍见：

H. M. Berman 等 8 位作者，*Nucleic Acids Res.* **28** (2000) 235 – 242.

网址：

http://www.rcsb.org/pdb/

在世界许多地方设有 PDB 镜象点。北京大学生物信息中心 [R-166] 和北京大学生物信息服务器 [R-167] 都有镜象。

R-442 **RCSB**，结构生物信息学合作研究组织 (Research Collaboration for Structural Bioinformatics)，现在是 PDB [R-441] 数据库的管理者。

网址：

http://www.rcsb.org/

R-443 **MSD**，大分子结构数据库 (Macromolecular Structure Database)，乃是交由 RCSB 管理后的 PDB 库的正式名称，不过 PDB 仍然是当前通用的名字。请看 PDB [R-441]。

R-444 **PDBNEW**，下一版 PDB 库正式发布前收到的全新或更新条目。

网址：

http://www.pdb.bnl.gov/

北京大学生物信息中心 [R-166] 设有镜象。

R-445 **PDBFinder**，在 PDB [R-441]、DSSP [R-465]、HSSP [R-466] 基础上建立的二级库，它包含 PDB 序列、作者、R 因子、分辨率、二级结构等。这些信息不易从 PDB 中直接读取。随着 PDB 库每次发布新版，PDBFinder 在 EBI [R-131] 自动生成，可能有几天延迟。请参看：

R. W. W. Hooft, C. Sander, M. Scharf, and G. Vriend, *CABIOS* **12** (1996) 525 – 529.

网址：

http://www.sander.embl-heidelberg.de/pdbfinder/

ftp://swift.embl-heidelberg.de (/pdbfinder)

R-446 **PDB at a Glance 清单**。PDB [R-441] 数据库中的每个条目由 4 位数字和字母编号，无法简单地从编号看出是什么样的蛋白质。NIH 的分子模拟网页上名为“PDB at a Glance”的这个超文本清单，帮助用户按蛋白质的功能分类迅速查找其 PDB 编号。网址：

http://cmm.info.nih.gov/modeling/pdb_at_a_glance.html

R-447 **PDBselect 数据库**。PDB 库中有大量同源蛋白的数据。研究工作中往往需要从中挑选出每个同源家族的代表，形成不含高度同源蛋白的结构数据子集合。PDBselect 库就是这样一个子集合。其最初描

述见：

U. Hobohm, and C. Sander, *Protein Science* **3** (1994) 522.

网址：

`http://swift.embl-heidelberg.de/pdbsel/`

`ftp://ftp.embl-heidelberg.de` (/pub/databases /protein_extras/pdb_select)

R-448 **PDBsum** 是 PDB [R-441] 库中数据的更便于阅读的总结和分析，以及一些衍生数据．例如，原来的坐标数据变成了图形，增加了从 CATH [R-455] 、 PROSITE [R-406] 等库得到的简明信息等．这是 University College London 维护的一个项目，描述见：

R. A. Laskowski 等 6 位作者， *Trends Biochem. Sci.* **22** (1997) 488 - 490.

网址：

`http://www.biochem.ucl.ac.uk/bsm/pdbsum/index.html`

R-449 **BioMagResBank**，简称 **BMRB**，是关于多肽、蛋白质和核酸的核磁共振数据库．它的结构数据与 PDB [R-441] 有些重复，但也收入了化学位移、 J 耦合、弛豫速率等 PDB 中没有的数据．网址：

`http://www.bmrb.wisc.edu/`

R-450 **CSD**，剑桥结构数据库 (The Cambridge Structural Database)．这实际上是最老的一个结构数据库．它不限于生物大分子，目前包含 20 万种以上有机和金属有机化合物的由 X 射线或中子衍射测定的结构数据．每一条目按“维数”组织：一维是文献数据，二维化学式，三维分子结构和三维晶体结构．此库虽不常用于蛋白质折叠的模拟，但对于配位结合位点的模拟以及蛋白质设计颇为有益．请参看：

D. G. Watson, *J. Res. Natl. Inst. Stand. Technol.* **101** (1996) 227 - 229.

网址：

`http://www.ccdc.cam.ac.uk/prods/csd.html`

R-451 **NRL-3D**，三维结构已经确定的蛋白质序列库．可以把新的蛋白质序列与此库中序列比较，以判断是否与结构已知的蛋白质相似．2000 年 1 月底的第 26.01 版收入 14 791 个蛋白质．网址：

`http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html`

http://www.gdb.org/Dan/proteins/nrl3d.html

R-452 **FAMBASE** 是每个蛋白质家族的代表序列的集合，它有助于加速同源性搜索。请参看 PIR [R-404] 的引文。网址：

http://pir.georgetown.edu/pirwww/dbinfo/fambase.html

R-453 **ProtFam**，蛋白质超家族的序列联配数据库。它是 PIR [R-404] 库的有机组成部分。网址：

http://www.mips.biochem.mpg.de/proj/protfam/protfam/

R-454 **SCOP**，蛋白质结构分类数据库 (Structural Classification Of Proteins)。这是对已知的蛋白质三维结构进行手工分类得到的数据库。请参看：

L. Lo Conte 等 6 位作者， *Nucleic Acids Res.* **28** (2000) 257 - 259.

网址：

http://scop.mrc-lmb.cam.ac.uk/scop/

它在世界许多地方设有镜象点。中国镜象在北京大学物理化学研究所：

http://www.ipc.pku.edu.cn/scop/

R-455 **CATH**，蛋白质结构与功能关系分类数据库。这是把蛋白质结构域按四个层次进行分类的数据库。这四个层次是“类别” (Class 即 C)，“构架” (Architecture 即 A)，拓扑 (Topology 即 T)，以及同源超家族 (Homologous superfamily 即 H)。库名即来自这四个字母。它有通向 PDB 总结文件和 OWL 库的超链接。详细描述见：

F. M. G. Pearl 等 8 位作者， *Nucleic Acids Res.* **28** (2000) 277 - 282.

网址：

http://www.biochem.ucl.ac.uk/bsm/cath/

R-456 **PIR-ALN**，蛋白质序列联配数据库，包括同一家族内 (彼此差异在 55% 以内) 序列的联配，一个超家族内不同家族代表序列的联配，以及不同蛋白质的同源结构域序列片段的联配。2000 年 1 月底的 22.03 版收入 4 076 个条目。库的描述见：

G. Y. Srinivasarao 等 6 位作者， *Nucleic Acids Res.* **27** (1999) 284 - 285.

G. Y. Srinivasarao 等 5 位作者， *Bioinformatics* **15** (1999) 382 - 390.

网址：

http://pir.georgetown.edu/pirwww/dbinfo/piraln.html
http://www-nbrf.georgetown.edu/pir/alndb.html

R-457 **3Dee**，蛋白质结构域定义的数据库，包括了 PDB [R-441] 库中含 20 个以上残基的蛋白质序列的结构域定义，但不包括理论模型。所有结构域按序列相似性和结构相似性分成聚类。所得家族按层次组织存储。3Dee 具有与 SCOP [R-454] 类似的、到本地计算机上 RasMol 程序 [R-777] 的接口，可用后者显示三维图象。网址：
http://circinus.ebi.ac.uk:8080/3Dee/

R-458 **ProTherm**，蛋白质及其变异体热力学数据库，包括几种热力学参数的数值，如吉布斯自由能、焓、热容、转变温度等。这些参数有利于理解蛋白质变异的结构和稳定性。它还包括关于二级结构、野生型残基、实验条件 (pH 值、温度等)、每种数据的测量方法等信息。ProTherm 2.0 版的描述见：
M. M. Gromiha 等 7 位作者，*Nucleic Acids Res.* **2** (2000) 283 - 285。
网址：
http://www.rtc.riken.go.jp/protherm.html

R-459 **ASTRAL** 是基于 SCOP [R-454] 数据库的一组分析蛋白质结构和蛋白质序列用的数据库和工具，包括 SCOP 结构域对应的序列库、按所需相似度组织的低冗余子集、由 SCOP 1.38 产生的结构对比库，以及工具和索引。请参看：
S. E. Brenner, P. Koehl, and M. Levitt, *Nucleic Acids Res.* **28** (2000) 254 - 256.
网址：
http://astral.stanford.edu/

R-460 **RESID**，蛋白质翻译后修饰情况的数据库，包括描述性的关于化学、结构和文献的信息。2000 年 1 月底的第 20.02 版共收入 275 个条目。详见：
J. S. Garavelli, *Nucleic Acids Res.* **28** (2000) 209 - 211.
网址：
http://pir.georgetown.edu/pirwww/search/textresid.html
http://www-nbrf.georgetown.edu/resid/get.html

R-461 **SMART**，是简单模块构架搜索工具 (Simple Modular Architecture Research Tool) 的缩写．它的最初目的是研究涉及真核生物信号转导 (signal transduction) 的蛋白质结构域，描述见：

J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting, *Proc. Natl. Acad. Sci.* USA **95** (1998) 5857 - 5864.

此库后来扩充到细胞外蛋白质的活动结构域、细菌双组元调控系统，以及与 DNA 、 RNA 、染色质和细胞骨架功能有关的结构域．这个基于网页的数据库的最近描述见：

J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork, *Nucleic Acids Res.* **28** (2000) 231 - 234.

网址：

`http://SMART.embl-heidelberg.de/`

R-462 **PROMISE** 数据库．其名称来自 The PROsthetic groups and MEtal Ions in protein SitEs 短语中的一些字母，即蛋白质活性位点的辅基中心 (prosthetic center) 和金属离子这些有生物学意义的无机组分 的数据库．详见：

K. N. Degtyarenko, A. C. T. North, and J. B. C. Findlay, *Nucleic Acids Res.* **27** (1999) 233 - 236.

网址：

`http://bmbsgi11.leeds.ac.uk/bmbknd/promise/MAIN.html`

R-463 **MMDB**，蛋白质分子模型数据库 (Molecular Modeling Database)，由 NCBI 的 MMDB 组维护．这是 Entrez 检索工具所使用的三维结构数据库，它以 ASN.1 格式 [R-180] 反映 PDB 库中的结构和序列数据，引文链接到 MEDLINE [R-599]． MMDB 有一个配套的三维结构显示程序 Cn3D，请参看 [R-779]．详见：

Y. L. Wang 等 7 位作者， *Nucleic Acids Res.* 28 (2000) 243 - 245.

网址：

`http://www.ncbi.nlm.nih.gov/Structure/`

`ftp://ncbi.nlm.nih.gov (/mmdb)`

R-464 **VAST**，矢量联配搜索工具 (Vector Alignment Search Tool)．此库包含 PDB 中所有结构域的结构和序列的联配数据，是寻找邻近三维结构时的原始数据．但它使用 ASN.1 格式 [R-180]，一般用户不易

直接阅读，描述见：

J. F. Gibrat, T. Madej, and S. Bryant, *Curr. Opin. Struct. Biol.* **6** (1996) 377 - 385.

网址：

`http://www.ncbi.nlm.nih.gov/Structure/vast.html`

`ftp://ncbi.nlm.nih.gov` (`/mmdb/vastdata/`)

R-465 **DSSP**，PDB 库中所有蛋白质条目的二级结构归属数据库 (Database of Secondary Structure assignments for all Protein entries)，网址：

`http://swift.embl-heidelberg.de/dssp/`

`ftp://ftp.embl-heidelberg.de` (`/pub/databases/dssp/`)

此库最早的描述见：

W. Kabsch, and C. Sander, *Biopolymers* **22** (1983) 2577 - 2637.

北京大学生物信息中心 [R-166] 有镜象。

R-466 **HSSP**，按同源性导出的蛋白质二级结构数据库。每一条 PDB [R-441] 项目都有一个对应的 HSSP 文件。因此，应先按蛋白质的 PDB 编号，例如 1dba 在 HSSP 的 INDEX 中查找 1dba.hssp，然后再读取压缩文件 1dba.hssp.Z。当然、通过 WWW 服务器查找更为方便。关于 HSSP 请参看：

C. Dodge, R. Schneider, and C. Sander, *Nucleic Acids Res.* **26** (1998) 313 - 315.

网址：

`http://www.sander.embl-heidelberg.de/hssp/`

`ftp://ftp.embl-heidelberg.de` (`/pub/databases/hssp`)

`ftp://ftp.embl-ebi.ac.uk` (`/pub/databases/hssp`)

北京大学生物信息中心 [R-166] 有 HSSP 的镜象。

R-467 **Dali/FSSP**，基于 PDB 数据库中现有蛋白质三维结构，用自动结构对比程序 Dali 逐一比较而形成的折叠单元和家族分类库。详见：

L. Holm, and C. Sander, *Nucleic Acids Res.* **27** (1999) 244 - 247.

此库在 PDB 库每次新版后自动更新，其网址：

`http://www.embl-ebi.ac.uk/dali/`

`http://croma.ebi.ac.uk/dali/fssp/`

ftp://ftp.ebi.ac.uk (/pub/databases/fssp)

北京大学生物信息中心 [R-166] 有镜象。

R-468 **3d_ali** 数据库，搜集彼此相关的蛋白质序列和结构数据．描述见：

S. Pascarella, F. Milpetz, and P. Argos, *Prot. Eng.* **9** (1996) 249 - 251.

网址：

http://www.embl-heidelberg.de/argos/ali/ali.html

ftp://ftp.embl-heidelberg.de (/pub/databases/3d_ali/)

ftp://ftp.ebi.ac.uk (/pub/databases/3d_ali)

R-469 **DEF**，蛋白质折叠类的预测数据库 (Database of Expected Fold classes)．它的构建基于 3d-ali [R-468] 数据．请参看：

M. Reczko, D. Karras, and H. Bohr, *Nucleic Acids Res.* **25** (1997) 235.

网址：

http://zeus.cs.uoi.gr/neural/biocomputing/def.html

R-470 **INFOGENE**，Sanger 中心计算基因组学小组维护的、各基因组测序计划所提供的序列中已知的蛋白质和预测出的基因与蛋白质的数据库．它有一个图形界面．描述见：

V. V. Solovyev, and A. A. Salamov, *Nucleic Acids Res.* **27** (1999) 248 - 250.

网址：

http://genomic.sanger.ac.uk/inf/infodb.html

R-471 **TMBase**，跨膜蛋白数据库．主要基于 SWISS-PROT [R-401] 的跨膜蛋白质片段．描述见：

K. Hoffmann, and W. Stoffel, *Biol. Chem. Hoppe-Seyler.* **374** (1993) 166.

网址：

ftp://ulrec3.unil.ch (/pub/tmbase)

ftp://ncbi.nlm.nih.gov (/repository/TMbase)

R-472 **PRESAGE** 是关于结构基因组学的一个数据库，它为库中每个蛋白质搜集了反映当前实验状况、结构、模型和研究建议的注释．详见：

S. E. Brenner, D. Barken, and M. Levitt, *Nucleic Acids Res.* **27** (1999) 251 - 253.

网址:

http://presage.stanford.edu/

R-473 **SBASE**，带有注释的蛋白质序列片段、即蛋白质结构域 的数据库，由 ICGEB [R-152] 建立和维护．关于其 7.0 版的介绍见:

J. Murvai, K. Vlahovicek, E. Barta, B. Cataletto, and S. Pongor, *Nucleic Acids Res.* **28** (2000) 260 - 262.

网址:

http://www.icgeb.trieste.it/sbase/

ftp://icgeb.trieste.it (/pub/SBASE)

北京大学生物信息中心 [R-166] 有镜象。

由于从测序得到的 DNA 翻译出来的氨基酸序列迅速增加，对这些可能的新蛋白质的功能和结构的预测越来越多地依靠同已知的蛋白质序列比较．我们在 3.5.6 小节中提到过，蛋白质结构域的比较对于确定同源性极为重要．现在已经有一批把各种蛋白质数据库中的模体、轮廓、结构域等局域模式信息集成起来的数据库，如 InterPro[R-474]、BLOCKS+[R-477] 等．

R-474 **InterPro**、集成的蛋白质结构域和功能位点数据库，目前仍在试运行．它把 SWISS-PROT [R-401]、TrEMBL [R-402]、PROSITE [R-406]、PRINTS [R-479]、PFAM [R-478]、ProDom [R-480] 等数据库提供的蛋白质序列中的各种局域模式 (pattern)，如 结构域、模体 等信息统一起来．此库在果蝇基因组 [R-369] 的注释和酵母、线虫与果蝇的比较基因组学研究中已经发挥作用．网址:

http://www.ebi.ac.uk/interpro/

R-475 **HITS**，瑞士 ISREC [R-143] 新近建立的一个蛋白质结构域数据库，它的方便之处在于给定蛋白质序列立即回答其中含有哪些模体，给出模体立即返回 SWISS-PROT 等数据库中含有该模体的蛋白质清单，并且带有相关链接．网址:

http://www.isrec.isb-sib.ch/cgi-bin/hits/hits_index

R-476 **BLOCKS**，蛋白质分类与同源性数据库，包含蛋白质家族中保守区域的组块 (blocks) 多序列联配的数据．这个数据库是根据 PROSITE [R-406] 中的条目，用 BLOSUM [R-620] 打分矩阵作序列联配生成，并随 PROSITE 库的每个新版更新，详见:

J. G. Henikoff, E. A. Greene, S. Pietrokovski, and S. Henikoff, *Nucleic Acids Res.* **28** (2000) 228 – 230.

原始数据库在美国西雅图的 FHCRC，即 Fred Hutchinson 癌症研究中心，网址：

`http://www.blocks.fhcrc.org/`

`ftp://ncbi.nlm.nih.gov (/repository/blocks/UNIXDOS)`

关于 BLOCKS 库的查询，还可用电子邮件 (在主文中写 HELP)：

`mailto: blocks@howard.fhcrc.org`

北京大学生物信息中心 [R-166] 有镜象。

R-477 **BLOCKS+ 数据库**。BLOCK 数据库基于专家审读过的 PROSITE 库，质量较好，但库中条目有限。因此，同一批作者又发展了一个 BLOCK+ 数据库。它由三个经过专家审读的数据库 PROSITE [R-406]、PRINTS [R-479] 和 PFAM-A [R-478]，以及两个自动产生的库 ProDom [R-480] 和 DOMO [R-482] 出发，使用 PROTOMAT 程序逐步添加新的组块。目前，

`http://www.blocks.fhcrc.org/`

网页的首选库就是 BLOCK+。请参看：

S. Henikoff, J. G. Henikoff, and S. Pietrokovski, *Bioinformatics* **15** (1999) 471 – 479.

R-478 **PFAM 或 PFAM-A，高质量的蛋白质结构域家族数据库**。它搜集蛋白质多序列联配和隐马可夫模型数据，已经达到同 SWISS-PROT [R-401] 和 TrEMBL [R-402] 中半数以上蛋白质匹配。2000 年 1 月发行的 5.0 版，有 2 008 个蛋白质结构域家族，与 SWISS-PROT [R-401] (第 38 版) 中 64% 的序列有匹配。PFAM 的重要用途是迅速自动地把 DNA 序列中预测出的蛋白质分成结构域家族，从而有助于对翻译出的蛋白质做注释。这时或者使用 HMMer [R-739] 软件，或者用 Wise2 程序包，后者的网址：

`http://www.sanger.ac.uk/Software/Wise2/`

PFAM 库第 4.3 版的描述见：

A. Bateman 等 6 位作者，*Nucleic Acids Res.* **28** (2000) 263 – 266.

网址：

`http://www.sanger.ac.uk/Software/Pfam/` (英国网点)

http://www.cgr.ki.se/Pfam/ (瑞典网点)

http://pfam.wustl.edu/ (美国网点)

ftp://ftp.sanger.ac.uk (/pub/databases/Pfam)

北京大学生物信息中心 [R-166] 也有镜象.

R-479 **PRINTS** 数据库最近改名为 **PRINTS-S**，这是一个蛋白质家族的指纹 (fingerprint) 和模体 数据库. 详见:

T. K. Attwood 等 8 位作者, *Nucleic Acids Res.* **28** (2000) 225 - 227.

网址:

http://www.bioinfo.man.ac.uk/dbbrowser/PRINTS/

ftp://ftp.ebi.ac.uk (/pub/databases/prints/)

ftp://ncbi.nlm.nih.gov (/repository/PRINTS/)

北京大学生物信息中心 [R-166] 有镜象.

R-480 **ProDom**，自动产生的蛋白质结构域 家族数据库，详见:

F. Corpet, F. Servant, J. Gouzy, and D. Kahn, *Nucleic Acids Res.* **28** (2000) 267 - 269.

网址:

http://www.toulouse.inra.fr/prodom.html

http://protein.toulouse.inra.fr/prodom.html

ftp://ftp.toulouse.inra.fr (/pub/prodom)

北京大学生物信息中心 [R-166] 有镜象.

R-481 **ProDomCG** 数据库与 ProDom [R-480] 类似, 是从完全基因组自动产生的蛋白质结构域 家族数据库. 请参看 ProDom 的引文和网址.

R-482 **DOMO**，蛋白质结构域 数据库. 法国国家生物信息中心 INFOBIOGEN [R-148] 维护的 DOMO 数据库, 自动分析蛋白质一级序列库 SWISS-PROT [R-401] 和 PIR [R-404]，找出其中的结构域并且把它们分组. 1999 年 7 月 DOMO 2.0 版中共有来自 83 054 个蛋白质序列的 99 058 个结构域, 后者又分为 8 877 组. 请参看:

J. Gracy, and P. Argos. *Bioinformatics* **14** (1998) 164 - 173.

网址:

http://www.infobiogen.fr/services/domo/

ftp://ftp.infobiogen.fr (/pub/domo/)

R-483 **GRBase**，这是参与基因调控的蛋白质的数据库 (Gene Regulation dataBase)．描述见：

B. Collier, and M. Danielsen, *Nucleic Acids Res.* **24** (1996) 219 220.

网址：

`http://www.access.digex.net/~regulate/`

`ftp://ftp.trevigen.com (/pub/Tfactors/)`

R-484 **PMD**，蛋白质突变体数据库 (Protein Mutant Database)，是一个集成了蛋白质序列和三维结构的显示和提取系统．描述见：

T. Kawabata, M. Ota, and K. Nishikawa, *Nucleic Acids Res.* **27** (1999) 355 – 357.

网址：

`http://pmd.ddbj.nig.ac.jp/`

R-485 **O-GLYCBASE**，蛋白质糖基化位点数据库．它搜集了至少有一个实验证实的糖基化位点的序列．它的一个子集 O-Unique 是不含相同糖基化位点的库． 1999 年初的 4.03 版有 180 个 G 蛋白条目．请参看：

R. Gupta 等 5 位作者， *Nucleic Acids Res.* **27** (1999) 370 – 372.

网址：

`http://www.cbs.dtu.dk/databases/OGLYCBASE/`

`ftp://ftp.cbs.dtu.dk`

在子目录 `/pub/Oglyc` 中取 `Oglyc.base` 和 `O-Unique.seq` 两个文件。

R-486 **ORDB**，嗅觉受体蛋白质序列数据库．嗅觉受体 (olfactory receptor) 是最大的真核生物基因家族． ORDB 库提供分析这些与 G 蛋白结合的受体功能的工具．详见：

E. Skoufos 等 5 位作者， *Nucleic Acids Res.* **28** (2000) 341 – 343.

网址：

`http://ycmi.med.yale.edu/senselab/ordb/`

`http://paella.med.yale.edu/`
`cgi-bin/receptor_top/DB_CGI.p/`

`http://paella.med.yale.edu/cgi-bin/receptor_top`
`/Public/cgiwrap/healy/`

此库有一部分是不公开的。

R-487 **CarbBank** 亦称 **CCSD**，复杂碳水化合物结构数据库，通常与蛋白质结构数据库归在一起。网址：

http://www.ccrc.uga.edu

http://mond1.ccrc.uga.edu

ftp://ncbi.nlm.nih.gov (/repository/carbbank)

中国科学院微生物研究所 [R-170] 设有镜象。

R-488 **SWISS-3DIMAGE**，蛋白质三维图象和 PDB [R-441] 浏览器。请参看：

M. C. Peitsch, T. N. C. Wells, D. R. Stampf, and J. L. Sussman, *Trends Biochem. Sci.* **20** (1995) 82 – 83.

网址：

http://www.expasy.ch/sw3d/

http://pdb.pdb.bnl.gov/expasy/sw3dimg/sw3d-top.html

ftp://ftp.expasy.ch (/databases/swiss-3dimage/)

北京大学生物信息中心有镜象：

http://exapsy.pku.edu.cn/sw3d/

ftp://ftp.expasy.pku.edu.cn (/databases/swiss-3dimage/)

R-489 **IMB**，大分子三维图象库。德国耶那的生物大分子三维图象库强调视象化和分析工具，它提供所有 PDB [R-441] 和 NDB [R-247] 库中条目的形象信息。请参看：

J. Reichert, A. Jabs, P. Slickers, and J. Suhnel, *Nucleic Acids Res.* **28** (2000) 246 – 249.

网址：

http://www.imb-jena.de/IMAGE.html

R-490 **BioImage**，多维生物学图象数据库。请参看：

J. M. Carazo 等 16 位作者，*Nucleic Acids Res.* **27** (1999) 280 – 283.

网址：

http://www-embl.bioimage.org/

http://www.bioimage.org/

R-491 **MolMovDB**，耶鲁大学的生物信息学研究室维护的分子运动数据库。网址：

http://bioinfo.mbb.yale.edu/MolMovDB/

R-492 **ModBase**，蛋白质结构模型比较数据库。请参看：

R. Sanchez 等 6 位作者， *Nucleic Acids Res.* **28** (2000) 250 – 253.

网址：

`http://pipe.ruckefeller.edu/modbase/`

§4.13 比较基因组学和蛋白质组学数据库

蛋白质同源家族的划分，对于确立物种亲缘关系和预测新蛋白质序列的功能有重要意义。同源蛋白质 (homolog) 进一步区分为直系同源 (ortholog) 和旁系同源 (paralog)。直系同源是指在不同物种中具有相同功能和共同起源的基因，例如哺乳动物的胰岛素基因。旁系同源是指在同一物种内具有不同功能、但有共同起源的基因，例如，同是起源于珠蛋白的 α 珠蛋白、β 珠蛋白和肌红蛋白。关于旁系和直系同源的定义请参看：

R-493 W. M. Fitch, *Syst. Zool.* **19** (1970) 99.

迅速增长的蛋白质数据库，为蛋白质分类和同源家族的划分提供了基础。1997 年在同一期美国《科学》周刊上曾有两篇文章讨论这个问题。文章作者多是某些数据库的作者：

R-494 S. Henikoff 等 6 位作者， "Gene families: the taxonomy of protein paralogs and chimeras", *Science* **278** (1997) 609 – 614.

R-495 R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families", *Science* **278** (1997) 631 – 637.

文献 [R-495] 的作者们引入了直系同源聚类 (Cluster of Orthologous Groups，简称 COG) 的概念，并且以分属于 17 个亲缘系的 21 个完全基因组中的蛋白质为基础，建立了 COG 数据库 [R-496]。

R-496 **COG**，直系同源聚类数据库。目前收入 2 091 个 COG。关于这个数据库及其检索工具的描述见：

R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, *Nucleic Acids Res.* **28** (2000) 33 – 36.

网址：

`http://www.ncbi.nlm.nih.gov/COG/`

还请参考 WIT [R-548] 网页下的 Ortholog Clusters 选项，那里的定

义更松散一些。

R-497 **GeneCensus**，耶鲁大学生物信息学研究室维护的各物种基因组的比较数据库，着重于折叠单元的结构对比。网址：

`http://bioinfo.mbb.yale.edu/genome/`

R-498 **XREFdb**，哺乳动物和模式生物的基因和遗传学交叉引用数据库。参看：

R. Ploger 等 7 位作者， *Nucleic Acids Res.* **28** (2000) 120 - 122.

网址：

`http://ncbi.nlm.nih.gov/XREFdb/`

R-499 **YPD**，酿酒酵母蛋白质组数据库。原来 YPD 中的 P 是指蛋白质，现在的意义是蛋白质组 (proteome)，更强调其生物性质及功能。详见：

M. C. Costanzo 等 14 位作者， *Nucleic Acids Res.* **28** (2000) 73 - 76.

网址：

`http://www.proteome.com/YPDhome.html`

`ftp://isis.cshl.org` (`/pub/yeast/YPD`)

北京大学生物信息中心 [R-166] 有镜象。

R-500 **WormPD**，线虫蛋白质组学数据库。详见 YPD [R-499] 的引文和网址。

§4.14 基因表达数据库

虽然每个细胞里都有包含全套基因信息的 DNA，但在特定的组织里，在一定的发育阶段和环境中，只有一部分基因被“表达”，即最终翻译为蛋白质。不同的生理或病理条件下，同一基因的表达速率也不断变化着。DNA 芯片和微阵列技术的发展，使人们得以研究大量基因同时表达的情况 (参看 §5.10 节关于大规模基因表达算法的讨论)。与单个基因表达有关的转录因子数据库 TRANSFAC [R-219] 等，已在前面列举，此节不再复述。

R-501 **Flyview**，果蝇基因表达数据库，描述见：

W. Janning, *Sem. Cell. Dev. Biol.* **8** (1997) 469 - 475.

网址：

http://flyview.uni-muenster.de/

R-502 **Flybrain**，果蝇神经系统图谱和数据库，描述见：

M. Heisenberg, and K. Kaiser, *Trends Neurosci.* **8** (1995) 481.

网址：

http://flybrain.uni-freiburg.de/

R-503 **NEXTDB**，线虫基因表达模式数据库 (Nematode Expression Pattern Database)。可以通过浏览器访问：

http://watson.genes.nig.ac.jp:8080/db/

R-504 **MAGEST** 数据库，其名字来自 MAboya Gene Expression patterns and Sequence Tags 短语的缩写。日文名字叫 Maboya 的海鞘 (*Halocynthia roretzi*) 是一种低等脊索动物。此库包含受精卵发育过程中的基因表达图谱和序列标记。描述见：

T. Kawashima 等 5 位作者， *Nucleic Acids Res.* **28** (2000) 133 - 135.

此库基于 Sybase 关系数据库，可通过 WWW 访问。网址：

http://star.scl.kyoto-u.ac.jp/magest/

R-505 **BodyMap**，人类和家鼠基因表达数据库，可按 DNA 序列查询。它最初基于大规模 cDNA 测序和基因表达的定性定量分析，目的在于通过系统地分析 cDNA 和构建数据库来发现新基因。详见：

T. Hishiki, S. Kawamoto, S. Morishita, and K. Okubo, *Nucleic Acids Res.* **28** (2000) 136 - 138.

网址：

http://bodymap.ims.u-tokyo.ac.jp/

R-506 **Axeldb**，非洲爪蟾基因表达数据库。它本身是用 ACeDB [R-851] 实现的。请参看：

N. Pollet 等 5 位作者， *Nucleic Acids Res.* **28** (2000) 139 - 140.

网址：

http://www.dkfz-heidelberg.de/abt0135/axeldb.htm

http://www.dkfz-heidelberg.de/tbi/axeldb_images/

docs/help.html (文件和显示示例)

R-507 **XMMR**，非洲爪蟾分子标记资源 (Xenopus Molecular Marker Resource)。它提供有关非洲爪蟾发育过程各个方面的链接以及相应研究单位的信息。网址：

http://vize222.zo.utexas.edu/

R-508 **TRIPLES**，酵母基因功能数据库，设在耶鲁大学医学院的基因组分析中心 (Yale Genome Analysis Center，简称 YGAC)。TRIPLES 是 TRansposon-Isertion Phenotypes, Localization, and Expression in Saccharomyces 的缩写。库的描述见：

A. Kumar 等 6 位作者，*Nucleic Acids Res.* **28** (2000) 81 – 84.

网址：

http://ygac.med.yale.edu/triples/

R-509 **MGEIR**，集成的家鼠基因表达信息资源 (Mouse Gene Expression Information Resource)。网址：

http://genex.hgu.mrc.ac.uk/

R-510 **GXD**，家鼠基因表达数据库，详见：

M. Ringwald, J. T. Eppig, J. A. Kadin, J. E. Richardson, 以及基因表达数据库小组，*Nucleic Acids Res.* **28** (2000) 115 – 119.

网址：

http://www.informatics.jax.org/
searches/gxdindex_form.shtml

R-511 **EpoDB**，脊椎动物红细胞生成 (erythropoiesis) 基因表达分析数据库。请参看：

C. J. Stoeckert Jr., F. Salas, B. Brunk, and G. C. Overton. *Nucleic Acids Res.* **27** (1999) 200 – 203.

网址：

http://cbil.humgen.upenn.edu/epodb/

R-512 **KidneyDB**，肾脏发育数据库，有通向引文的链接。网址：

http://www.ana.ed.ac.uk/anatomy/kidbase/kidhome.html

R-513 **ToothExp**，牙齿基因表达数据库。网址：

http://honeybee.helsinki.fi/toothexp/toothexp.html

§4.15 基因突变、病理和免疫数据库

我们把有关基因突变的数据库同病理和免疫放在一起列举。

R-514 关于人类基因突变的命名规则，请参看：

Human Mutation **8** (1996) 197 - 202; **11** (1998) 1 - 3.

R-515 **HGMD**，人类基因突变数据库，可用于预测基因疾病。描述见：

D. N. Cooper, E. V. Ball, and M. Krawczak, *Nucleic Acids Res.* **26** (1998) 285 - 287.

网址：

http://uwcm.web.cf.ac.uk/uwcm/mg/hgmd0.html

R-516 **Marfan**，人类 FBN1 基因突变数据库及分析软件。其第 3 版描述见：

G. Collod-Baroud 等 19 位作者， *Nucleic Acids Res.* **26** (1998) 229 - 233.

网址：

http://uwcm.web.cf.ac.uk/uwcm/mg/hgmd0.html

R-517 **Collagen**，人类胶原数据库。它搜集所有已知的人类第 1 类胶原 α1 链和 α2 链基因突变，以及第 III 类胶原 α1 链突变 (COL1A1，COL1A2 和 COL3A1) 的数据。请参看：

R. Dalgleish, *Nucleic Acids Res.* **26** (1998) 253 - 255.

网址：

http://www.le.ac.uk/genetics/collagen/

R-518 人类 **PAX2** 等位基因变异数据库。请参看 [R-519] 的引文。网址：

http://www.hgu.mrc.ac.uk/Softdata/PAX2/

R-519 人类 **PAX6** 等位基因突变数据库。请参看：

A. Brown, M. McKie, V. van Heyningen, and J. Prosser, *Nucleic Acids Res.* **26** (1998) 259 - 264.

网址：

http://www.hgu.mrc.ac.uk/Softdata/PAX6/

R-520 **Androgen**，雄激素受体突变数据库，包含与男性性器官发育不良、前列腺癌等有关图谱，密度、频度以及基因型和表现型关联数据。描述见：

B. Gottlieb, M. Trifiro, R. Lumbroso, and L. Pinsky, *Nucleic Acids Res.* **25** (1997) 158 - 162.

网址：

http://www.mcgill.ca/androgendb/

ftp://ftp.ebi.ac.uk (/pub/databases/androgen/)

R-521 **ALFRED** 为 ALlele FREquency Database 的缩写。这是由耶鲁大学 K. K. Kidd 实验室维护的一个针对人口多样性和 DNA 多态性的等位基因数据库。描述见:

K. H. Cheung 等 6 位作者, *Nucleic Acids Res.* **28** (2000) 361 – 363.

网址:

http://alfred.med.yale.edu/alfred/

R-522 **CD40LBASE**, CD40L 基因突变数据库。CD40L 突变导致与 X 染色体相联系的血免疫球蛋白过多综合征 (X-linked hyper IgM syndrome, 简称 X-HIM)。此库的一部分是文献目录。

网址:

http://www.expasy.ch/cd40lbase/

ftp://ftp.expasy.ch (/databases/cd40lbase)

北京大学生物信息中心 [R-166] 有镜象。

R-523 **KMDB**, 由日本庆应义塾 (Keio) 大学医学院建立的一组与人类疾病有关的基因突变数据库。最早只有眼病数据库 KMeyeDB, 现在已发展出与心脏、耳、脑和癌症有关的 KMheartDB、KMearDB、KMbrainDB 和 KMcancerDB。它们都是借助一个名叫 MutationView 的数据库软件建立的。关于这些数据库的总描述, 请参看:

S. Minoshima 等 5 位作者, *Nucleic Acids Res.* **28** (2000) 364 – 368.

从 KMDB 的网页, 可以进入任何一个库, 但在访问时须先注册。网址:

http://mutview.dmb.med.keio.ac.jp/

R-524 **KMeyeDB**, 人类疾病和眼病基因突变数据库。设在日本庆应义塾大学医学院。访问时须先注册。网址:

http://mutview.dmb.med.keio.ac.jp/mutview3/kmeyedb/

R-525 **KMheartDB**, 人类心脏病基因突变数据库。设在日本庆应义塾大学医学院。请参看 KMDB[R-523] 的网址。

R-526 **KMearDB**, 人类耳病基因突变数据库。设在日本庆应义塾大学医学院。请参看 KMDB[R-523] 的网址。

R-527 **KMbrainDB**，人类脑病基因突变数据库。设在日本庆应义塾大学医学院。请参看 KMDB[R-523] 的网址。

R-528 **KMcancerDB**，人类癌症基因突变数据库。设在日本庆应义塾大学医学院。请参看 KMDB[R-523] 的网址。

R-529 **OMIA** 是一大批动物的孟德尔遗传、疾病、基因型和表现型的数据库，其组织与 OMIM[R-335] 库有相似之处。请访问澳大利亚的 OMIA 在线服务器：

http://www.angis.su.oz.au/BIRX/omia/omia_form.html

R-530 **Atlas**，法国建立的针对肿瘤学和血液学的遗传与细胞遗传交互数据库 (Atlas of Genetics and Cytogenetics in Oncology and Heamatology)，正在完善之中。其描述可见：

J. L. Huret 等 5 位作者，*Nucleic Acids Res.* **28** (2000) 349 - 351.

网址：

http://www.infobiogen.fr/services/chromcancer/

R-531 **F7MD**，凝血因子 VII 突变位点数据库，详见 HAMSTeRS [R-532] 的引文和网址。

R-532 **HAMSTeRS**，凝血因子 VIII 结构和突变位点数据库。HAMSTeRS 是 Haemophilia A Mutation Search Test and Resource Site 的缩写。这是所有从 A 型血友病患者身上发现的点突变、插入和删除的总汇。网页上还有凝血因子 VIII 蛋白质结构和基因分析的信息，以及 A 型血友病分子遗传学的综述。HAMSTeRS 第 4 版描述见：

G. Kemball-Cook, E. G. D. Tuddenham, and A. I. Wacey, *Nucleic Acids Res.* **26** (1998) 216 - 219.

网址：

http://europium.mrc.rpms.ac.uk/

ftp://ftp.ebi.ac.uk (/pub/databases/hamsters)

R-533 **HaemB**，B 型血友病凝血因子 IX 点突变和短插入或删除序列的数据库，其第 8 版描述见：

F. Giannelli 等 11 位作者，*Nucleic Acids Res.* **26** (1998) 265 - 268.

网址：

http://www.umds.ac.uk/molgen/haemBdatabase.htm

ftp://ftp.ebi.ac.uk (/pub/databases/haemb)

R-534 **TTMD**，转基因动物和靶突变数据库 (Transgenic/Targeted Mutation Database)。网址：

http://tbase.jax.org/

下面主要列举与人类有关的病理和免疫数据库。

R-535 **FIMM**，功能分子免疫学数据库。它搜集以细胞免疫为重点的、与功能分子免疫学有关的数据，包括蛋白质抗原、主要组织相容性复合体 MHC 分子、与 MHC 有关的多肽、以及相关疾病等。请参看：

C. Schonbach, J. L. Y. Koh, X. Sheng, L. Wong, and V. Brusic, *Nucleic Acids Res.* **28** (2000) 222 – 224.

网址：

http://sdmc.krdl.org.sg:8080/fimm/

R-536 **MTB**，家鼠肿瘤生物学 (Mouse Tumor Biology) 数据库。以家鼠作为遗传性癌症的模型生物，描述其肿瘤和肿瘤细胞系、肿瘤病理报告和图象、与肿瘤发展有关的遗传因子、发病率、以及通向其他网上资源的链接。详见：

C. J. Bult, D. M. Krupke, J. P. Sundberg, and J. T. Eppig, *Nucleic Acids Res.* **28** (2000) 112 – 114.

网址：

http://tumor.informatics.jax.org/

R-537 **BCGD**，人类乳腺癌基因数据库。网址：

http://condor.bcm.tmc.edu/ermb/bcgd/bcgd.html

R-538 **PDD**，人类体液中蛋白质与疾病关系的数据库 (Protein Disease Database)。请参看：

C. R. Merril, *Appl. Theor. Electrophoresis* **5** (1995) 49 – 54.

网址：

http://www-lmmb.ncifcrf.gov/PDD/

http://www-pdd.ncifcrf.gov/

R-539 **PAH** 是导致人类苯丙酮尿症 (phenylketonuria) 的苯丙氨酸羟化酶特异位点 (PhenylAlanine Hydroxylase locus) 数据库。这是一个经过人工审读的关系数据库。描述见：

P. Nowacki, S. Byck, L. Prevost, and C. R. Scriver, *Nucleic Acids Res.* **26** (1998) 220 – 225. 网址：

http://www.mcgill.ca/pahdb/

R-540 **CFTR**，囊性纤维变跨膜调控子 (Cyctic Fibrosis Transmembrane conditional Regulator) 突变数据库。网址：

http://www.genet.sickkids.on.ca/cftr/

R-541 **NRR**，核受体资源 (Nuclear Receptor Resource) 计划，包括糖类皮酯激素 (glucocorticoid，见 GRR)、矿质肾上腺皮质激素 (mineralocorticoid)、甲状腺激素、维生素 D 受体、类固醇受体等信息的数据库。请参阅：

E. Martinez 等 9 位作者， *Nucleic Acids Res.* **25** (1997) 163 - 165.

网址：

http://nrr.georgetown.edu/nrr/nrr.html

http://nrr.georgetown.edu/GRR/GRR.html

R-542 **IMGT**， 1989 年建立的国际免疫遗传学数据库 (International ImMunoGeneTics database)。它包括各种脊椎动物免疫球蛋白 (Ig)、T 细胞受体 (TcR) 和主要组织相容性复合体 (MHC) 分子。它由两个库组成： IMGT/LIGN-DB 为人及脊椎动物 Ig 和 TcR 数据库，包括带详细注释序列的翻译；以及 IMGT/HLA-DB，即人类白细胞抗体数据库。由 IMGT 服务器可以访问各种免疫遗传学数据。详见：

M. Ruiz 等 12 位作者， *Nucleic Acids Res.* **28** (2000) 219 - 221.

网址：

http://imgt.cines.fr:8104/

ftp://imgt.cines.fr (/pub/IMGT)

http://www.ebi.ac.uk/imgt/

北京大学生物信息中心 [R-166] 有镜象。

R-543 **HIG**，Anthony Nolan 骨髓和白血病基金会的人类白细胞抗体 HLA 信息组 (HLA Informatics Group)。它的 HLA 序列数据库包含第 I 类和第 II 类 HLA 的核酸与蛋白质序列的联配结果。这里还有 HLA 等位基因命名规则等信息。网址：

http://www.anthonynolan.com/HIG/

R-544 **Kabat**， 30 年前由 E. A. Kabat 建立的具有免疫学意义的蛋白质序列数据库。 1991 年书面出版的第 5 版为三卷巨著：

E. A. Kabat, T. T. Wu, H. Perry, K. Gottesman, and C. Foeller,

Sequences of Proteins of Immunological Interest, NIH Publications, No. 91-3242, 5th ed. 1991.

1999 年 9 月底， Kabat 库容量为 1991 年的五倍，所包含的抗体轻链和重链分别有 1 599 375 和 2 517 756 个核苷酸。详情请参看：

G. Johnson, and T. T. Wu, *Nucleic Acids Res.* **28** (2000) 214 - 218.

网址：

`http://immuno.bme.nwu.edu/`

`ftp://ttwu.bme.nwu.edu` (`/pub/database/`)

许多国际生物信息中心有镜象。北京大学生物信息中心 [R-166] 也有镜象。

R-545 **PEDB**，前列腺表达数据库，由 Leroy Hood 领导的华盛顿大学 (西雅图) 分子肿瘤与发育实验室维护。最近描述见：

P. S. Nelson 等 7 位作者， *Nucleic Acids Res.* **28** (2000) 212 - 213.

网址：

`http://www.mbt.washington.edu/PEDB/`

`http://chroma.mbt.washington.edu/mod_www/`

R-546 **HIV**，艾滋病分子免疫学数据库。网址：

`http://hiv-web.lanl.gov/immunology/immuno-main.html`

R-547 斯坦福大学的 **HIV RT** 数据库，包含几乎全部已发表的 HIV RT(反转录酶) 和蛋白酶序列，是研究抗 HIV 药物靶分子演化和与药物有关变化的原始资料。这个库的重要性在于，最近发现鸡尾酒疗法等混合药物的疗效与病人过去的治疗史有关。许多艾滋病药物都可能诱发相互之间的抗药性。描述见：

R. W. Shafer 等 5 位作者， *Nucleic Acids Res.* **28** (2000) 346 - 348.

网址：

`http://hivdb.stanford.edu/hiv/`

§4.16 代谢途径和细胞调控数据库

基因组学和蛋白质组学的迅猛进展，展现了从整体上研究细胞内代谢途径和调控网络的前景。有关数据库和网页处在不断更新和重组中。

R-548 **WIT** 是 What Is There 的缩写。这是美国阿贡 (Argonne) 国家实验室的一个集成的重构代谢途径和模型的系统。它允许在网页上交互式地进行大量基因组序列的分析和建立模型，对代谢途径、酶、模型、操作子等提出查询。1995 年 WIT 第 1 版的网址：

http://www.cme.msu.edu/WIT/

WIT2 是其新版。现在第一次访问这个网页时须先注册，以后才能自由使用。其描述请参看：

R. Overbeek 等 9 位作者， *Nucleic Acids Res.* **28** (2000) 123 - 125.

网址：

http://wit.mcs.anl.gov/WIT2/

R-549 **EMP** 是酶与代谢途径 (Enzymes and Metabolic Pathways) 的缩写。

网址：

http://biobase.com/emphome.html/

http://www.biobase.com/EMP/

R-550 **MPW**，代谢途径 (Metabolic PathWays) 数据库，是 EMP [R 549] 库的一个子集。请参看：

E. Selkov, Jr., Y. Grechkin, N. Mikhailova, and E. Selkov, *Nucleic Acids Res.* **26**(1998) 43 - 45.

网址：

http://www.cme.msu.edu/MPW/

http://beauty.isdn.msc.anl.gov/MPW/

在上面第二个网点，可用类似电路图的方式绘制代谢途径。

R-551 **PUMA**，原是单细胞生物代谢途径亲缘联配数据库 (Phylogeny of the Unicellular organisms Metabolism pathways Alignment)。它的功能已经完全被 WIT [R-548] 数据库覆盖。以下网页也不复存在：

http://www.msc.anl.gov/home/compbio/PUMA/

R-552 **EcoCyc** 数据库和 **MetaCyc** 数据库。前者试图描述大肠杆菌的全部生化网络，包括基因、代谢途径、信号转导 途径和运输蛋白等。后者是以微生物为主的多个物种的酶和代谢途径数据库。两者在大肠杆菌数据上有重复。请参看：

P. D. Karp 等 6 位作者， *Nucleic Acids Res.* **28** (2000) 56 - 59.

两个库的网址都在：

http://ecocyc.panbio.com/ecocyc/

http://ecocyc.PangeaSystems.com/ecocyc/

http://www.ai.sri.com/ecocyc/ecocyc.html

R-553 **PathDB**，代谢途径数据库。由 NCGR [R-135] 发展和维护的这个生物化学和代谢途径数据库，搜集了丰富的有关酶、生化反应、代谢途径、输运步骤和化合物的信息。所有数据按物种分类组织。可以通过网页寻访，也可以下载一个 Java 工具来访问。网址：

http://www.ncgr.org/Software/pathdb/

R-554 **KEGG**，京都基因与基因组百科全书 (Kyoto Encyclopedia of Genes and Genomes)，它包含核酸分子、蛋白质序列、基因表达、基因组图谱、代谢途径图等。此库的建立参考了 Boehringer Mannheim 公司的代谢途径挂图 [R-555] 和日本生物化学学会的收藏。详见：

M. Kanehisa, and S. Goto, *Nucleic Acids Res.* **28** (2000) 27 - 30.

网址：

http://www.genome.ad.jp/kegg/

ftp://kegg.genome.ad.jp/

此库每天更新，在日本基因网络中心有镜象：

http://www.tokyo-center.genome.ad.jp/kegg/

http://www.tokyo-center.genome.ad.jp/kegg2.html

http://www.genome.ad.jp/kegg/kegg.html

R-555 由 Boehringer Mannheim 公司提供的代谢途径图，悬挂在许多生化实验室的墙壁上。与大型地图类似，它分别以字母和数字标识图中小块。各种酶和反应物的盘根错节关系，现在可以分块显示在屏幕上，并且上下左右跟踪。例如从 ENZYME 数据库查得，腺三磷酶 ATPase 在挂图 S3 区，可由此开始追踪。有关此挂图及其索引，请参看：

http://www.expasy.ch/cgi-bin/search-biochem-index/

R-556 **SMILES** 是一个辅助性数据库，它搜集与代谢途径有关的化合物名称。网址：

http://www.daylight.com/dayhtml/smiles/

R-557 **LIGAND**，酶反应化学数据库，由日本京都大学化学研究所维护。它从酶反应角度提供化学与生物学的联系。请参看：

S. Goto, T. Nishioka, and M. Kanehisa, *Nucleic Acids Res.* **28** (2000)

380 – 382.

网址：

http://www.genome.ad.jp/htbin/show_man?ligand

ftp://ftp.genome.ad.jp (/db/genomenet/ligand)

R-558 **CSNDB**，细胞中信号网络的数据库 (Cell Signaling Networks DataBase)。日本国立健康科学研究所建立的这个数据库，是人类细胞中信号途径的数据和知识库。它汇编了有关信号传输的生物分子、序列、结构、功能和生物化学反应，并可自动绘图表示信号途径。库的构建基于 ACeDB [R-367]，并有通向 TRANSFAC [R-219] 的链接。

网址：

http://geo.nihs.go.jp/csndb/

R-559 **Biocatalysis/Biodegradation**，生物催化与生物降解数据库。关于这个由 Minnesota 大学建立的数据库，可参看：

L. B. M. Ellis, C. D. Hershberger, and L. P. Wackett, *Nucleic Acids Res.* **28** (2000) 377 – 379.

网址：

http://dragon.labmed.umn.edu/~lynda/index.html

§4.17 与农林牧有关数据库

与农作物、树木和家禽、家畜有关的基因图谱数据库很多，在列举之前，先介绍几个主要的机构及其网址：

R-560 美国农业部 (USDA) 国家农业图书馆 (NAL) 基因组信息系统 (Agricultural Genome Information System，简称 AGIS)，它本身的服务器基于 ACeDB [R-851]。其旧网址：

http://probe.nalusda.gov:8000/

已不适用。植物基因组和其他物种基因组数据库已经转到下面 ARS 的网址。

R-561 **ARS**，农业研究服务处新设立在康奈尔大学的 USDA-ARS 生物信息学和比较基因组学中心，ARS 是 Agricultural Research Service 的缩写。网址：

http://ars-genome.cornell.edu/

这里也是日本小麦网 KOMUGI [R-575] 的美国镜象点。

R-562 **AgDB**，农业数据库和信息资源总清单。设在美国农业网络信息中心 (Agricultural Network Information Center，简称 AgNIC) 的 AgDB，是与农业有关的数据库和信息资源的总清单，内容相当丰富。我们不一一列举，请参见网址：

`http://www.agnic.org/agdb/`

R-563 设在英国爱丁堡的 Roslin 研究所的生物信息组，发展了名为“方舟” (Ark) 的系统来搜集和比较各种动物基因图谱。详情请参看他们的网址：

`http://www.ri.rrsrc.ac.uk/bioinformatics/ark-overview.html`

这里是猪、鸡、马、猫、火鸡、鹿、绵羊、丽鲷 (tilapia) 和鲑鱼基因图谱数据库的原始网址，还有牛类基因图谱库的镜象。进入的办法是在下面的 URL 的 == 后面填写物种名字，例如：

`http://www.ri.rrsrc.ac.uk/cgi-bin/arkdb/broswers/broswer.sh?species=pig`

R-564 **INRA**，法国国家农业研究所 (Institut National de la Recherche Agronomique)。这里有牛 [R-590]、水牛 [R-593]、山羊 [R-588]、兔 [R-596]、鳟鱼 [R-597]、马 [R-589] 等动物的基因图谱数据库，以及美国谷物基因库 GrainGenes [R-572] 的镜象。网址：

`http://locus.jouy.inra.fr/`

R-565 美国得克萨斯 A& M 大学是牛类基因图谱数据库的原始网址和绵羊、马数据库的镜象点：

`http://bos.cvm.tamu.edu/`

R-566 美国衣阿华州立大学有猪和鸡基因图谱数据库的镜象点：

`http://www.genome.iastate.edu`

4.17.1 农作物

下面是一批农作物基因图谱数据库的网址。

R-567 **UK CropNet**，英国农作物植物生物信息网络。这里有许多谷物类植物基因组数据库和其他生物信息，如大麦数据库 barleydb、牧草 (forage grasses) 数据库 foggdb、狗尾草数据库 milletgenes、芸薹数据库 BrassicaDB 等。请参看：

J. Dicks 等 16 位作者， *Nucleic Acids Res.* **28** (2000) 104 - 107.

网址：

`http://synteny.nott.ac.uk/`

R-568 **INE**，水稻基因组数据库。INE 是集成水稻基因探索者 (INtegrated rice genome Explorer) 首尾两字头的缩写，又是日文“稻”字的拼音。以日本为主的国际水稻基因组计划 (RGP) 集中对 *O. sativa* ssp. *japonica* 栽培亚种，又称日本粳稻 (nipponbare) 或 GA3 进行测序。此计划的开始见：

N. Kurata 等， *Nature Genetics* **8** (1994) 365 - 372.

新建的 INE 数据库基于 OOP [R-51] 概念，包含各种基因图谱的彩色显示。描述见：

K. Sakata 等 7 位作者， *Nucleic Acids Res.* **28** (2000) 97 - 101.

网址：

`http://www.staff.or.jp/giot/INE.html`

`ftp://ftp.staff.or.jp`

要想了解国际水稻基因组计划的进展情况，可以访问美国布鲁克海文国家实验室的基因组 ftp 服务器：

`ftp://genome1.bio.bnl.gov/`

在它的 /pub/maize/ 子目录里，保存着历次国际水稻基因组工作会议的记要。2000 年 4 月初，孟山都公司宣布它已经完成水稻 12 个染色体测序任务的 80%，并将与国际科学界共享测序成果。请参看：

`http://www.monsanto.com/monsanto/mediacenter/2000/00apr4_rice.html`

R-569 我国水稻基因组计划针对水稻的籼稻亚种。关于已经完成的物理图谱，有关文章见：

G. F. Hong 等 15 位作者， “A 120 kilobase resolution contig map of the rice genome”, *DNA Seq.* **7** (1997) 319 - 335.

此图谱数据，可访问国家基因研究中心 [R-175] 的网页：

`http://www.ncgr.ac.cn/`

`ftp://ftp.ncgr.ac.cn`

测序工作过去集中于第 4 号染色体。最近中国科学院遗传研究所人类基因组中心 [R-174] 已启动以籼稻为亲本的超级杂交水稻的大规

模测序。

R-570 美国 TIGR 研究所 [R-156] 维护着几个与水稻基因组有关的数据库，包括水稻基因组注释库 (Rice Genome Annotation Database，简称 RGAD)、水稻重复序列库 (估计水稻基因组中重复序列约占 50%)，以及水稻基因索引 (Oryza sativa Gene Index，简称 OsGI)。这里还有一个指向水稻基因组计划其他参与单位的链接表。请参看网址：

http://www.tigr.org/tdb/rice/

R-571 **RiceGenes** 是美国康奈尔大学的水稻基因组数据库，它包括水稻遗传学、遗传图谱、探针、种质 (germplasma)、QTL[23] 和比较图谱等方面的信息。此库采用 ACeDB [R-851] 软件，过去要通过美国国家农业图书馆的网址访问，现在直接由 [R-561] 进入：

http://ars-genome.cornell.edu/rice/

R-572 **GrainGenes** 是由美国农业部和国家农业图书馆的植物基因组计划支持的麦、燕麦和甘蔗遗传数据库。它搜集遗传和细胞遗传图谱、基因探针、核酸序列、基因、等位基因和基因产物、相关的表现型、QTL、病理和病原、昆虫、麦属和燕麦属的分类，以及此领域研究人员地址名单、文献等。GrainGenes 的原始数据库在康奈尔大学，过去要通过美国农业部的 AGIS [R-560] 服务器访问，现在可由 [R-561] 直接进入：

http://ars-genome.cornell.edu/

GrainGenes 数据库和服务器在法国 INRA [R-564] 有一个镜象点，称为 WWW GRAIN。网址：

http://grain.jouy.inra.fr/

ftp://grain.jouy.inra.fr (/pub/database)

R-573 关于世界范围的水稻生产和市场等情况，可以访问网址：

http://www.riceweb.org/

http://www.riceworld.org/

R-574 **WHEAT**，小麦基因图谱数据库，可访问网址：

http://wheat.pw.usda.gov/

[23] QTL 即数量性状基因座 (Quantitative Trait Loci)，性状决定比单一的孟德尔因子复杂，由多个“微效”基因决定，请参看 [R-37] 一书。

http://wheat.pw.usda.gov/ggpages/newquery-request.html

R-575 **KOMUGI**，日本小麦网，由 16 所大学和研究所联合维护。KOMUGI 是日文麦字。这是有关小麦属 (*Triticum*) 以及相近的燕麦属 (*Avena*) 和山羊草属 (*Aegilops*) 农作物的基因图谱和他信息的数据库。网址：

http://www.shigen.nig.ac.jp/wheat/wheat.html

此网址有时不能自由进入，可试其美国镜象点 [R-561]。

R-576 **MaizeDB**，玉米基因组数据库。此网页上还有不少其他信息，包括每年一度的玉米遗传学会议消息。网址：

http://www.agron.misouri.edu/top.html

R-577 **ZmDB**，玉米基因组数据库。请参看：

X. W. Gai, S. Lal, L. Q. Xing, V. Brendel, and V. Walbot, *Nucleic Acids Res.* **28** (2000) 94 – 96.

网址：

http://zmdb.iasstate.edu/

R-578 **ILDIS**，国际豆科植物数据库和信息服务 (International Legume Database and Information Service)，可通过其 LehumeWeb 检索有关豆科 (Leguminosae) 植物的信息。网址：

http://www.ildis.org/LegumeWeb/

R-579 豆类 (beans) 基因图谱：

http://scaffold.biologie.uni-kl.de/Beanref/

R-580 **Soybase**，大豆 (*Clycine max*) 数据库。这是美国农业部植物基因组计划资助的衣阿华大学豆类数据库的一部分。它结构上与 ACeDB [R-851] 类似，具有方便的链接和图形界面。可通过 ARS [R-561] 的网址访问。

R-581 **MGI**，NCGR [R-135] 和 Samuel Roberts Noble 基金会联合开展的豆科苜蓿属植物 *Medicago truncatula* 的基因组研究，在 2000 年 4 月已经提交 15 000 多条 EST，网址：

http://www.ncgr.org/research/mgi/

R-582 **cottonDB**，美国南方平原农业研究中心 (Southern Plains Agricultural Research Center，简称 SPARC) 所维护的棉花数据库，包括棉花遗传学知识库和基因组学数据库，后者含有棉花的 BAC 文库。网

址：

http://algodon.tamu.edu/

R-583 **TreeGenes**，树木遗传图谱数据库，采用 ACeDB [R-851] 数据库格式．网址：

http://probe.nalusda.gov:8000/plant/abouttreegenes.html

ftp://probe.nalusda.gov (/pub/treegenes/)

4.17.2 家畜、家禽和鱼类

下面是家禽、家畜、鱼类和其他有关动物的基因图谱数据库。

R-584 **ChickGBASE**，鸡基因图谱计划，搜集全世界鸡基因图谱信息，包括标记、图谱、微卫星、命名规则、文献，以及与鸡基因图谱计划有关的单位名单等．请参看：

D. W. Burt 等 5 位作者，*Trends in Genet.* **11** (1995) 190 - 194.

网址：

http://www.ri.bbsrc.ac.uk/chickmap/
chickgbase/manager.html

R-585 **Swinemap**，猪基因图谱计划，包含染色体图谱和标记。网址：

http://sol.marc.usda.gov/genome/swine/swine.html

http://www.ri.bbsrc.ac.uk/pigmap/pigbase/pigbase.html

R-586 **PiGBASE**，猪基因图谱信息库，设在英国 Roslin 研究所、美国农业部的家畜基因组计划和美国衣阿华大学．网址分别是：

http://www.ri.bbsrc.ac.uk/pigmap/arkpig/

http://www.public.iastate.edu/~pigmap/pigmap.html

http://probe.nalusda.gov:8000/animal/aboutpigbase.html

R-587 **SheepBase**，已发表的绵羊基因位点数据库，由新西兰牧业研究所建立，格式与 PigBASE [R-586]、ChickGBASE [R-584]、和 BovGBASE [R-592] 一致．描述见：

J. A. Sise, A. L. Hillyard, and G. W. Montgomery, *Mammalian Genome* **7** (1996) 1.

网址：

http://dirk.invermay.cri.nz/

http://www.ri.bbsrc.ac.uk/sheepmap/ (Roslin 研究所)

http://tetra.gig.usda.gov:8400/sheepgbase/manager.html （美国农业部）

R-588 **Goatmap**，山羊 (*Capra hircus*) 基因图谱数据库。网址：
http://locus.jouy.inra.fr/

R-589 **HorseMap**，马 (*Equus caballus*) 基因图谱数据库。网址：
http://www.ri.bbsrc.ac.uk/horsemap/arkhorse/
http://locus.jouy.inra.fr/

R-590 **Bovmap**，法国的牛 (*Bos taurus*) 基因图谱数据库。网址：
http://locus.jouy.inra.fr/cgi-bin/bovmap/intro.pl

R-591 **BovBase**，英国的牛基因图谱数据库，设在 Roslin 研究所 [R-563]，网址：
http://www.ri.bbsrc.ac.uk/bovmap/arkbov/

R-592 **BovGBASE**，美国农业部的家畜基因组图谱计划中的牛基因数据库。网址：
http://probe.nalusda.gov:8000/animal/aboutbovgbase.html

R-593 **Buffmap**，水牛 (*Babalus bubalis*) 基因图谱数据库。网址：
http://locus.jouy.inra.fr/

R-594 **DogMap**，狗基因图谱数据库。网址：
http://ubeclu.unibe.ch/itz/dogmap.html
http://mendel.berkeley.edu/dog.html

R-595 **CatMap**，猫基因图谱数据库，设在 Roslin 研究所。网址：
http://www.ri.bbsrc.ac.uk/catmap/ark/

R-596 **RabbitMap**，兔 (*Oryctolagus cuniculus*) 基因图谱数据库。网址：
http://locus.jouy.inra.fr/

R-597 **RainMap**，彩虹鳟鱼 (Rainbow trout, *Oncorhynchus mykiss*) 基因图谱数据库。网址：
http://locus.jouy.inra.fr/

R-598 另一个鳟鱼科 (Salmonids) 基因数据库在美国华盛顿州立大学：
http://www.wsu.edu:8000/~thorglab/DATA.HTML

§4.18 生物医学文献数据库

本节主要列举一些文献摘要和检索、名词术语定义，以及引用情况查询的网点和数据库.

R-599 **MEDLINE** (MEDlars onLINE) 是美国国家医学图书馆的文献摘要库，反映美国及其他国家 3 800 多种医学和生物期刊的作者摘要和引用情况. 网址:

`http://www.nlm.nih.gov/databases/medline.html`

北京大学生物信息中心 [R-166] 有 MEDLINE 的镜象.

R-600 最为方便的查询 MEDLINE 的方式，是通过 NCBI 的 **PubMed** 服务:

`http://www.ncbi.nlm.nih.gov/PubMed/`

R-601 **SeqAnalRef**，这是由 A. Bairoch 个人维护的有关序列分析的文献目录，可以用多种方式检索. 请参看:

A. Bairoch, *CABIOS* (*Bioinformatics*) **7** (1991) 268.

网址:

`http://www.expasy.ch/seqanalref/`

`ftp://ftp.expasy.ch (/databases/seqanalref)`

此库的问题是自 1996 年 2 月的 67.0 版以来还没有更新，但在各主要生物信息中心有镜象. 北京大学生物信息中心 [R-166] 也有镜象.

R-602 **SCI** 是设在美国费城的科学信息研究所 (Institute of Scientific Information，简称 ISI) 所提供的文献引用情况的检索服务. 只有付费订阅单位可访问其 Web of Science 网页:

`http://webofscience.com/`

R-603 **CancerWeb**，癌症网页:

`http://www.graylab.ac.uk/cancerweb.html`

包含关于癌症患者、临床治疗、教育、文献等多方面内容.

R-604 **HUMAT**，人体解剖学数据库. 网址:

`http://ed.ac.uk/anatomy/database/humat/`

R-605 **KeyNet**，按生物序列功能组织的基因和蛋白质名称关键字库. 描述见:

D. Catalano, F. Licciulli, D. D'Elia, and M. Attimonelli, *Nucleic Acids*

Res. **28** (2000) 372 - 373.

网址:

http://www.ba.cnr.it/keynet.html/

R-606 **BioABACUS**，生物学与生物技术以及计算机科学缩写字表，包括原名、意义、常见用法和到更详细解释的链接. 请参看:

M. Rimer, and M. O'Connell, *Bioinformatics* **14** (1998) 888 - 889.

网址:

http://www.nmsu.edu/~molbio/bioABACUShome.htm

§4.19 其他数据库

R-607 **Taxonomy**，分类学数据库。这是 NCBI [R-134] 和 GenBank [R-212] 所维护的生物分类数据库。任何物种，只要 GenBank 中至少有一条核酸或蛋白质序列，就在此库中有所反映。1999 年底收录的物种超过 55 000 。它的上层界面是一个分类学浏览器。网址:

http://www3.ncbi.nlm.nih.gov/Taxonomy/tax.html

R-608 **ETI**，世界生物多样性数据库设在荷兰的分类鉴定专家中心 (Expert Centre for Taxonomic Identifications，简称 ETI)。网址:

http://www.eti.uva.nl/

R-609 位于美国麻省的 Woods Hole 海洋生物研究室 (Marine Biology Laboratory，简称 MBL) 有一个海洋动物数据库 (Marine Animal Models)，它搜集了 210 种海生脊椎动物和鱼类的有关信息。网址:

http://database.mbl.edu/SPECIMENS/

如不能直接进入，可从 MBL 的主页选“数据库”项. 网址:

http://www.mbl.edu/

R-610 **TAED**，适应性演化数据库 (The Adaptive Evolution Database)。它目前包括脊索动物和植物的 TAED . 网址:

http://www.sbc.su.se/~liberles/TAED.html

第 5 章　服务、软件和算法

生物信息学和生物计算往往相提并论．生物信息学当然涉及大量算法和软件，也是一种生物计算．然而，某些传统的生物计算，诸如生物大分子的结构和相互作用的分子动力学模拟、代谢机制和免疫网络的模拟等等，本身都基于专业知识，是设备齐全的专业实验室的研究课题，一般不归入生物信息学范围，本书也不叙述．

生物信息学的首要任务，是从数据中提取知识．一般地说，分子生物学、遗传学、分子演化和基因工程所涉及的基于数据库的计算，通常在三个层次进行．

第一，享用网上服务：目前许多生物信息中心或条件较好的实验室都在网络上提供现成服务．不久以前，主要靠电子邮件提交作业和获取结果．现在，越来越多的服务可直接在互联网浏览器上实现．这在目前仍是大部分生物学者的主要工作方式，我们将扼要介绍这类服务．它的局限性在于只能有什么用什么，而且参数选择不当时容易被“自动”返回的结果误导．

第二，利用现成软件：这包括购买商业性软件包，或者从互联网下载免费软件，在本地计算机系统上实现．商业性软件通常有相应安装和维护服务，也不在本书介绍之例．使用免费软件，对用户有更多要求，首先要知道从何处获取，其次要自己安装，再次往往要作剪裁修改．这一般只能靠配套提供的使用说明来摸索，不宜过多麻烦原作者，何况有些作者早已改换课题，不再关心往事．我们将主要介绍一些获取软件的线索，而不谈程序的安装实现问题．

第三，创造信息环境：真正开展研究工作时，会发现任何成套供应的软件都不可能恰到好处地满足需要．这时就要自己动手或者同数理、计算工作者合作编制程序．这当然离不开对算法的研究和发展．网上自由软件带有的源程序，往往很值得参考．这里最重要的是把网上服务、下载软件和自编程序集成为一体，创造一个生物信息学的工作环境．这首先是

国家或地区生物信息中心的使命。每个研究集体也应因地制宜、逐步创造自己的信息环境。

按照这三个层次，本章标题为“服务、软件和算法”。然而叙述起来，却不得不把算法的简略介绍放在前面第 §5.2 节。否则说不清楚后面许多服务和软件使用中的注意事项。

对于许多生物学者，工作实践中最常见的需求，是把自己的核酸或蛋白质序列送去同国际数据库中收藏的海量序列进行比较，寻找同源关系和对结构、功能的启示。在我国现实条件下，借助主要国际生物信息中心的 BLAST 和 FASTA 服务器，仍然是较为可行的办法。因此，§5.3 提前介绍这两种服务。它们恰好也是说明 §5.2 中不少概念的良好实例。有了这些基础，就可以分类成批地介绍其他软件和服务。这是 §5.4 节到 §5.13 节的内容。最后，在 §5.14 和 §5.15 节中，罗列一些非软件性质的网络资源，主要是电子期刊、新闻和讨论组、会议、讲义等。

§5.1 软件和服务目录

国际互联网上有许多生物信息、生物计算软件和服务的目录，还有一些免费软件的档案库。下面列举若干网址。许多目录和档案库的缺点，是更新不够及时。读者查到某个感兴趣的条目，最好追踪到原作者的网址，查验有无更新消息。此外，用户还必须自己从相应网址下载，在计算机上安装。与商业软件不同，安装时往往会出现一些小问题，计算机系统经验不足时，会耽误时间。它们的优点是可以从源程序学到不少知识和技巧。有些后来成为商品的软件，在网上仍可能查到曾经免费的老版本。

R-611 美国印地安那大学的 IUBio 生物学软件档案是重要的信息资源之一。它的软件有分类目录，许多数据库每日更新。网址：

`http://iubio.bio.indiana.edu/`

`http://iubio.bio.indiana.edu/soft/molbio/`

`ftp://iubio.bio.indiana.edu (/molbio/)`

IUBio 在世界各地有多处镜象点，请参看上面第一个网址。

R-612 **BioCatalog** 是由欧洲生物信息研究所 EBI [R-131] 维护的分子生物学和遗传学自由软件目录。它按照不同的平台 (UNIX 、PC 等) 列

举软件功能、作者、引文、硬件要求以及从网络上获取的方法. 它欢迎学者提供自己的软件信息. BioCatalog 的网址:

ftp://ftp.ebi.ac.uk (/pub/databases/bio_catal/)

http://www.ebi.ac.uk/biocat/

同时, 欧洲生物信息研究所 EBI [R-131] 的 ftp 服务器

ftp://ftp.ebi.ac.uk/

上还保存着另外一套生物软件目录. 用无记名 ftp 进入之后, 可在

/pub/software/dos 和 /pub/software/unix

子目录中分别找到适用于 PC 和 UNIX 平台的软件信息. 北京大学生物信息中心有镜象:

ftp://ftp.cbi.pku.edu.cn (/pub/software/)

R-613 美国国家生物技术信息中心 NCBI [R-134] 的网页和 ftp 服务器上, 有两大类免费软件. 一类是 NCBI 自己研制的高质量的生物软件, 它们大都与 NCBI 的各种数据库和 Entrez [R-199] 检索工具集成在一起, 也可以下载后独立运行. 本书提及的 Cn3d [R-779], Sequin [R-790] 等, 以及 Entrez [R-199] 本身都属于这一类. 另一类是学者们提供的自由软件.

R-614 **GenomeWeb** 是由 HGMP [R-140] 维护的一个详尽的与基因组有关的链接地址和单位的清单. 它按基因中心、核酸、蛋白质、亲缘树、基因组数据库、图谱等项目分类, 并可按字母检索. 问题是有些地址"埋藏"甚深, 要多次辗转才能找到. 网址:

http://www.hgmp.mrc.ac.uk/GenomeWeb/

北京大学生物信息中心有镜象:

http://www.cbi.pku.edu.cn/GenomeWeb/

R-615 **BioMedNet**, 生物医学研究人员互联网团体 (The Internet Community for Biological and Medical Researchers) 的网页, 是另一个新消息的来源. 第一次使用时需先进入网站注册:

http://www.bmn.com/

从这里也可以访问 MEDLINE [R-599]. BioMedNet 的电子期刊目录很有用. 网址:

http://journals.bmn.com/

它还不定期地开放某些重要的生物学刊物供免费阅读和下载.

R-616 **GeneInfo**，堪萨斯大学医学中心维护的遗传专业人员信息 (Information for Genetic Professionals) 网页。它有通向遗传学学术组织、临床遗传学数据库、以及遗传学计算机资源的链接。网址是：
`http://www.kumc.edu/gec/geneinfo.html`
请参看 [R-848]。

R-617 较新的软件信息来源，是第 1 章提到的期刊 *Bioinformatics* [R-5]，即过去的 *CABIOS*。它所发表的算法，均要求作者在两年内公开相应源程序。事实上，许多文章发表时，程序已可自由获取。

R-618 另一个期刊 *Computer and Chemistry* 也经常发表与生物信息学有关的算法和程序描述。

此外，第 5.14.6 小节提到的一些个人网页，也可参考。

§5.2 序列分析算法概要

生物信息学计算的核心是序列的比较，这包括同一个序列内不同片段的比较，以及两个或多个序列的对比。比较的内容，从序列的组分变化、寻找特殊的字段，到序列间字母的对应。比较的主要目的在于阐明序列之间的同源关系，以及从已知序列预测新序列的结构和功能。所用方法也从半经验的直观手段，到具备较深刻数学背景的复杂算法，跨度很大。本书 §1.3 节点到的书名，从 [R-24] 到 [R-38]，都是探讨序列分析方法。我们在这一节里，只能极其简要地介绍一些基本概念。

人们之所以在算法问题上大作文章，是因为涉及核酸和蛋白质序列的计算，很容易在存储容量和计算时间两个方面都超出现代计算机的处理能力。

任何问题都有一个特征尺度 N，例如生物序列的长度，N 可能从几百到几百万。如果计算时间比例于 N、N^2 等有限的幂次，或者说计算时间按 N 的某种多项式增长，现代计算机还可以处理到比较大的 N。如果计算时间随 N 的指数 e^N 增加，那就根本不可能处理稍大的问题。这类需要“非多项式时间”的问题 (NP 问题)，超出了当代计算机的能力，属于真正的计算难题。NP 问题中有一大类是互相等价的，它们可以用多项式时间彼此转换。因此，解决其中任何一个就解决了全类。这类问题

称为 NP 完备问题. 然而, 至今既没有证明存在着解决它们的多项式时间的算法, 也没有证明不存在多项式时间的解法. DNA 序列分析中的许多问题, 都属于 NP 完备类.

NP 问题还有一个特点, 即一旦知道了它的解, 只要用多项式时间就可以演示. 其实, 需要多项式时间的某些问题也是求解难, 演示易.

5.2.1 序列联配基本概念

两个或多个符号序列按字母比较, 尽可能确切地反映它们之间的相似和相异, 称为序列的 联配 (alignment)[24] .

我们先讨论序列联配算法所涉及的一些主要概念. 核酸和蛋白质序列联配的前提是, 假定两个序列来自同一个祖先 ("同源"), 它们在演化过程中由于变异的积累而成为不同的序列. 作为符号序列看待, 点变异包括字母的代换 (substitution) 、删除 (deletion) 和插入 (insertion) ; 插入和删除统称为 "插删" (indel)[25] . 两个序列联配时, 往往要插入空位 (gap) , 以达到总体上更好的排列效果. 每当第一次插入空位时, 要计一定的 "罚分" (penalty) ; 连续插入空位时通常按比例给以稍小的罚分. 因此, 计算一组连续空位罚分的公式是 $p = a + b \times n$, 其中 n 是连续空位总数. 两个常数 a 和 b 的值, 与所比较的是核酸还是蛋白质序列有关, 而且要同下面讲到的打分矩阵的选择和数值范围适应. 例如, 选用 BLOSUM62 矩阵 [R-620] 比较蛋白质序列时, 可以取 $a = -12$ 和 $b = -2$. 这是基于统计和经验的 "艺术", 而不是可以简单论证的定量结果. 最初使用联配程序时, 宜先接受程序为 a 和 b 设定的 "补缺" (default) 值.

两个核酸序列的联配较为简单. 序列中一个嘌呤被嘧啶代换或反之, 称为颠换 (transversion) ; 嘌呤或嘧啶互换称为置换 (transition) . 如果根本不 "鼓励" 字母替换, 可以用单位矩阵作打分矩阵 (score matrix) 或代换矩阵 (substitution matrix) , 即令:

---

[24] 目前 alignment 一词有 "比对"、"对比"、"对排"、"阵排列" 等多种译法. 从词意和实际使用方便看, 在其具有特定意义的上下文中, 不宜译为普通词汇. 因此, 我们以为 "联配" 似更为合适.

[25] "插删" 是我们为从 insertion 和 deletion 缩并出的 indel 一词建议的译名.

	a	c	g	t
a	1	0	0	0
c	0	1	0	0
g	0	0	1	0
t	0	0	0	1

这就是说，两个序列中相应的核苷酸相同，记 1 分；否则记 0 分。有人用如下的打分矩阵：

	a	c	g	t
a	0.9	−0.1	−0.1	−0.1
c	−0.1	0.9	−0.1	−0.1
g	−0.1	−0.1	0.9	−0.1
t	−0.1	−0.1	−0.1	0.9

BLASTN [R-631] 程序使用的打分矩阵：

	a	c	g	t
a	M	N	N	N
c	N	M	N	N
g	N	N	M	N
t	N	N	N	M

它要求 $M > 0$ 和 $N < 0$，补缺值是 $M = 5$ 和 $N = -2$，用户可以另外设置.

20 种氨基酸彼此之间的代换，远比核苷酸复杂. 残基代换所引起的后果，与它们的具体物理化学性质有关. 因此，对各种代换的效果，要有所估计，计算出各种打分矩阵. 常用的打分矩阵有两类，即 PAM 矩阵和 BLOSUM 矩阵，可参看 AAindex 数据库 [R-440].

R-619 **PAM 矩阵.** M. Dayhoff 等 [26] 在 20 世纪 70 年代后期引入了 PAM(Point Accepted Mutation) 概念：取一个蛋白质序列中的氨基酸变异 1% 作为演化距离的单位，称之为 1 个 PAM. 注意，100PAM 并不意味着序列变得完全不同，因为有些变异是互相抵消的.

[26] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins", in *Atlas of Protein Sequence and Structure*, ed. by M. O. Dayhoff, Washington DC, National Biomedical Research Foundation, 1978, 345 – 352, 353 – 358.

表 5.1 PAM250 打分矩阵

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	−2	6																		
N	0	0	2																	
D	0	−1	2	4																
C	−2	−4	−4	−5	12															
Q	0	1	1	2	−5	4														
E	0	−1	1	3	−5	2	4													
G	1	−3	0	1	−3	−1	0	5												
H	−1	2	2	1	−3	3	1	−2	6											
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	5										
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6									
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	5								
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6							
F	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3			
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17		
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4

M. Dayhoff 等用手工比较了当时数目有限的同源蛋白质序列，取实际观察所得的代换频度与随机背景序列的相应频度比值的对数，用统计方法得到对应 1PAM 的数据，再外插到 250PAM．表 5.1 给出常用的 PAM250 矩阵．实际计算中针对不同的演化距离，使用从 PAM100 到 PAM500 不等的打分矩阵．亲缘关系近者用 PAM100 到 PAM150，亲缘关系远者用更高号的矩阵，相当于容许更高的噪声背景．

表 5.1 中两个色氨酸 (W) 相匹配得最高分 17．这是因为在蛋白质中，平均含量只有 1.23% 的色氨酸在序列中具有较高的保守性，两个 W 相匹配是概率较小的非偶然事件．事实上，W 被其他多数氨基酸代换都得负分，正表明它的保守性．与此对照，在蛋白质中平均含量达 7.78% 的两个丙氨酸 (A) 相匹配，是概率较高的普通事件，只得 2 分 (作者感谢张春霆提供了从 SWISS-PROT 数据库第 37 版 8 万多条蛋白质序列计算出

的氨基酸平均含量).

R-620 **BLOSUM 矩阵.** 近来使用较多的 BLOSUM 矩阵是根据 BLOCK (见 [R-476]) 数据库中蛋白质序列的高度保守部分的联配自动产生的。见:

S. Henikoff, and J. G. Henikoff, *Proc. Natl. Acad. Sci.* USA **89** (1992) 10915 - 10919.

表 5.2 BLOSUM62 打分矩阵

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	−1	5																		
N	−2	0	6																	
D	−2	−2	1	6																
C	0	−3	−3	−3	9															
Q	−1	1	0	0	−3	5														
E	−1	0	0	2	−4	2	5													
G	0	−2	0	−1	−3	−2	−2	6												
H	−2	0	1	−1	−3	0	0	−2	8											
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

许多序列联配程序自动以表 5.2 所示的 BLOSUM62 作为首选打分矩阵. 从 BLOSUM30 到 BLOSUM90 都可能用到. 请注意, 与 PAM 矩阵相反, BLOSUM 矩阵大号对应近亲, 小号对应远亲.

试比较表 5.1 和表 5.2 中的数值, 可见两者的大趋势一致. 一般认为, BLOSUM 比 PAM 略好一些. 表 5.3 简要地比较这两类打分矩阵. 许多程序设有选项, 允许用户选用其中一种.

表 5.3 PAM 矩阵和 BLOSUM 矩阵的比较

	PAM	BLOSUM
得自	近缘序列整体联配	远缘序列的局域联配
矩阵号	大者对应远演化距离	大者对应近演化距离
矩阵元数值	由近缘者外插得到	直接计算得到
常见补缺	PAM250	BLOSUM62

这里要着重讲一下局域联配和整体联配. 生物序列中有重要功能的片段, 往往比较保守, 即变异的速率很低. 序列的其他部分可能具有较高的变异速率, 在演化过程中变得面目全非. 例如, 真核生物的 DNA 序列中, 往往是较短较少的保守片段, 被甚为丰富的高变异区淹没. 如果片面强调整体联配, 可能会漏掉真正的同源序列. 良好的局域联配往往会更有效地揭示同源关系. 更有甚者, 真核生物基因组中存在种种长短不一的重复片段, 两个序列的重复片段对齐, 可能给出得分很高的联配结果, 造成同源假象. 因此, 许多联配程序允许用户决定是否要"过滤"掉简单的重复序列.

请注意, 序列联配时讲究的"复杂"和"简单", 与人们研究"复杂性"和"复杂系统"时不大相同. 例如, 大量 aaaaaaaaaaaaaaa 、 ttttttttttttttt 或 atatatatatatatatat 这样的片段, 即所谓 Poly(a) 、 Poly(t) , 或蛋白质序列中富含脯氨酸 P 的片段, 它们明显包含较少信息, 被认为是简单的. 相反, 杂乱包含 a 、 c 、 g 、 t 四种字母的一串, 一般视为复杂片段. 由计算机去作决定时, 要有一套算法. J. Wootton 等专门研究过这类问题, 可以参考他在 [R-10] 和 [R-31] 中的综述. 读者如果想看一看"过滤"简单片段的结果, 可以调用 RepeatMasker [R-748] 程序, "加工"一条酵母染色体, 或者在向 BLAST 提交序列时, 对"过滤"做不同的选择.

绝大多数序列联配是针对蛋白质的. 提交一条蛋白质序列, 直接同蛋白质库里所有的序列对比, 不需对序列再作什么变换. 如果要把这条蛋白质序列, 同数据库里的 DNA 序列比较, 那就要把后者翻译成"蛋白质". 对于双链 DNA 的每个单链, 因为翻译起始点的不同, 要按照 3 个读框做翻译, 一共得到 6 条供比较用的"蛋白质"序列. 提交一条 DNA 序列, 去同核酸数据库中的序列做比较, 当然也无需变换. 如果要同蛋白

质序列库做比较，所提交的 DNA 序列也得按照 6 个读框先翻译出来。像 BLAST 和 FASTA 这类通用程序，序列的变换都包含在其功能之内，用户只需提出要求。一般地说，翻译成蛋白质序列再进行联配，结果比较灵敏。

两个蛋白质序列的“相似性”超过 25%，是同源性的一种佐证，但不是唯一的证明，必须靠生物学知识来论证。相似性大于 30%，一般比较有把握。相似性低于 25% 的情形，被 M. Dayhoff 称为序列对比的“晦暗区” (twilight zone)，就更需要生物学释疑，诸如两者是否同为细胞外蛋白或膜蛋白、同为多结构域的蛋白质，或具有类似的内含子类型，等等。

5.2.2 半经验的直观算法

假定已经选择好打分矩阵、空位罚分等参数，要求把一条给定的核酸或蛋白质序列，同数据库中所有现存序列进行联配，找出最相似的哪些序列，这是远非平庸的计算课题。如果进一步允许在联配时插入空位，计算难度就会空前增大。这首先是因为插入空位的位置和数目有大量可能的组合，一切靠“穷举”法挑出最佳方案的企图，都会超出现在和可以设想的未来的计算机能力。

从 20 世纪 80 年代以来，人们发展了一些半经验的直观算法。它们可以相当快地给出较好的结果，但不能保证所得结果是最优的。BLAST 和 FASTA 就是很成功的实例。我们极其简单地说明一下 BLAST 算法的基本思想。首先，BLAST 事先为数据库里的全部序列作了“索引”。它首先规定了一个字母串长度 (在 FASTA 中相应参数为 WORD 或 ktup)，对 DNA 序列是 11，蛋白质序列是 6。把每个序列所含的此种串的类型作为索引。提交一个新序列时，也先对它做索引。只有索引类型兼容的库中序列才用来做比较。这样就大为减少了搜索工作量。其次，从局域联配得分最高的片段开始，向左右两端延伸，直到一端到头或总计分下降超过事先设置的值。然后再把这样得到的结果作比较，选出统计上最显著的哪些，排队输出。

5.2.3 动态规划算法

动态规划是把大问题按时间步骤或空间分布分割处理、逐步寻求最

优结果的一套算法. 通常表示成一套逐点向前的递归关系, 再回溯找到最佳路径. 最早把动态规划算法用于寻求序列的整体最优联配的文章是:

R-621 S. B. Needle, and G. E. Wunsch, "A general method applicable to the search for similarities in the amino acid sequences of two proteins", *J. Mol. Biol.* **48** (1970) 443 - 453.

R-622 P. H. Sellers, "On the theory and computation of evolutionary distances", *SIAM J. Appl. Math.* **26** (1974) 787 - 793.

后来 Smith 和 Waterman 把动态规划算法用于寻求序列的局域最优联配.

R-623 **Smith-Waterman** 算法:

T. F. Smith, and M. S. Waterman, "Identification of common molecular subsequences", *J. Mol. Biol.* **147** (1981) 195 - 197; *Adv. Appl. Math.* **2** (1981) 482 - 489.

动态规划方法虽然有效地压制了计算量的"二项式"爆炸, 但计算量仍比例于 N^2, N 是问题的尺寸, 例如序列长度. 因此, 在很长时期里人们不能使用此法, 而不得不满足于半经验的直观算法, 诸如 BLAST 和 FASTA 所使用的办法. 随着计算机速度的增长, 这种限制已不严重. 因此, 越来越多的生物信息中心开始提供使用 Smith-Waterman 算法的服务. 这方面比较好的一个程序是 SSEARCH:

R-624 **SSEARCH3** 程序实现 Smith-Waterman 算法, 是 FASTA3 程序包的一部分, 也可以单独调用. 请参看 [R-642].

关于动态规划算法的详情, 可参看 M. S. Waterman 的专著 [R-29].

5.2.4 神经网络和隐马可夫链

神经网络算法是对生物神经系统信息处理过程的极其肤浅的模拟. 它有一个输入层面, 一个输出层面. 这两个层面之间还可以有若干隐含的"学习"层面. 每个层面中有许多结点. 结点之间的连接有种种方案. 每个结点如何把输入信号转变为输出 (传递函数), 也有不少选择. 要用大量已知前因后果的数据对神经网络进行训练, 也就是对其包含的大量参数作拟合. 经过训练的神经网络, 可以从同类的未经处理的数据中提取信息. 这虽然被人们视作计算机学习和提取知识的实例, 训练数据选择恰当

时，也能够解决某些实际问题，但是隐藏在大量参数中的“知识”，很难提炼成简单明白的指导原则，真正丰富人类的知识宝库。这是神经网络模型的基本弱点。

作为离散随机过程，各种马可夫链 (Markov chain) 模型有较坚实的数学基础。由 a 、 c 、 g 、 t 四种字母组成的一条长序列，如果是完全随机的，任何一个字母后随任意其他字母的概率都相同。例如， aa 、 ac 、 ag 、 at 出现的概率相等，都是 $1/4 = 0.25$ 。或者说， $a \to a$ 、 $a \to c$ 、 $a \to g$ 和 $a \to t$ 的“转移概率”都是 0.25 。用单个字母的“状态概率”和字母之间的“转移概率”，构造出一个离散随机过程，是为一阶马可夫链。从某物种的一条实际 DNA 序列可以计算出一套“状态概率”和“转移概率”，构造相应的马可夫链。用这样的模型可以检验给定的另一个 DNA 序列是否属于该物种。然而，这类简单马可夫链能处理的问题十分有限。

哺乳动物基因组的每条单链 DNA 中，从 5′ 端往 3′ 端计数， cg 的数目显著少于 gc 数目，而且分布不均匀。哪些 cg 比较集中的片段，称为 CpG 岛 (参看数据库 [R-323] 的简要说明)。它们往往是基因启动子区域的标记。可以分别构造两个马可夫链，对应 CpG 岛和非 CpG 岛的区域。然后引入在两个马可夫链之间的比较小的转移概率，描述两类区域的交替。这就是一个简单的隐马可夫链模型，可用于识别 CpG 岛的边界。为了反映序列中的空位或字母的“插删”、内含子与外显子的剪切等，都可以构造相应的隐马可夫链模型。

神经网络、马可夫链和隐马可夫链都是基于概率论的算法，都是数据库知识发现 (Knowledge Discovery in Databases，简称 KDD) 或数据采矿 (Data Mining，简称 DM) 中常用的方法。关于 KDD 和 DM，可从以下网址开始网上浏览：

R-625 **KDD** 数据库知识发现网页：

http://www.kdnuggets.com/

R-626 **DM** 数据采矿文献目录：

http://www.cs.bham.ac.uk/~anp/

贝叶斯统计 (Bayes statistics) 是从大量数据中用统计方法提取知识的基本工具。由于需要从“先验”分布出发，历史上曾经引发过争论。对于先验分布，现在有了较好的理论基础，而计算机使得反复迭代变得轻而易

举，最终结果可以与初始分布无关。因此，贝叶斯统计重新受到重视。有关进展，请参看 [R-18] 等书。

5.2.5 语言学方法

自从 20 世纪 70 年代测得第一个 DNA 序列以来，统计方法就是处理生物学符号序列的重要手段，而随机序列则是统计评估的基本参考。实际的生物序列，无论 DNA 还是蛋白质，当然都不是随机的。然而，如果刻画的角度不妥，所提取的许多特征量又离开随机不远。这表明，统计方法不足以充分放大 DNA 序列与随机序列之间，以及 DNA 序列之间的差别，必须寻求越出单纯统计方法的新途径。语言学方法可能值得注意。

数据库中的核酸和蛋白质记录，都是有方向的、可以按确定方式从左向右读的一维序列。核酸由 4 种字母组成，蛋白质由 20 种字母组成。它们都自然地满足语言的形式定义，可以借助语言学方法加以研究。事实上，从提出“中心法则”以来，转录、翻译、编辑、修饰等等具有语言背景的术语就在分子生物学中广泛应用。

生物遗传语言和人类自然语言有许多相似之处，例如多义性、冗余性、容错或纠错性、有长程关联、有某种语法框架但不能完全“生成”，存在多种方言和个体差异、都有演化和灭绝问题，都保留着少数“古语”或“化石”成分，等等。同时，它们又有深刻差别，例如标点和间隔的不同、两种或多种语言的相互作用、重复序列的数目和功能不同等。经过一定程度的抽象后，语言学 (language 而不是 philology) 的方法应能在生物信息学中发挥更大作用。

第一，在语言学中已经对生成语法和语言的复杂性有较好的分类。按照 N. Chomsky 的串行生成语法，语言区分为正规语言、上下文无关语言、上下文有关语言和递归可数语言四个层次。DNA 序列中的个别“字法”可以和某些层次对应，例如回文 (palindrome) 对应上下文无关语法，而假扭结 (参看 [R-272]) 对应上下文有关语法。使用语法规则寻找基因的尝试见 [R-703]。

第二，并行生成的 Lindenmayer 系统本身来自发育生物学的观察，预期在生物问题中会有更多应用，但目前尚未引起充分注意。形式语法很容易推广成模糊语法。然而，只有能进一步对模糊程度作定量刻画，才会

更贴近生物学。

第三，还应研究随机语法。隐马可夫链模型相当于随机正规语法，更复杂的层次仍有待钻研。广而言之，无论自然语言或遗传语言，都是基于离散的排列组合系统。组合学方法应能在建立生命现象的理论方面发挥更多作用。

本书作者们最近由细菌完全基因组出发定义了一种“可因式化”语言，严格解决了研究基因组中缺失和稀少字母串时遇到的一个计数问题，同时也用组合学方法得到一致的结果。希望这些初步尝试能收抛砖引玉之效：

R-627 B. L. Hao (郝柏林), H. C. Lee and S. Y. Zhang (张淑誉), "Fractals related to long DNA sequences and complete genomes", *Chaos, Solitons and Fractals* **11** (2000) 825 - 836.

R-628 Bai-lin Hao (郝柏林), "Fractals from genomes: exact solutions of a biology-inspired problem", *Physica* **A282** (2000) 225 - 246.

关于形式语言学的基本概念和引文，请参看谢惠民的专著：

R-629 谢惠民，《复杂性与动力系统》，上海科技教育出版社，1994。

R-630 Hui-min Xie, *Grammatical Complexity and One-Dimensional Dynamical Systems*, in *Directions in Chaos*, vol. 6, ed. by Bai-lin Hao, World Scientific Publishing Co., 1996.

§5.3 BLAST 、 FASTA 和类似服务

BLAST 和 FASTA 是使用得最为频繁的两套数据库搜索程序。它们的功能相近，都是把用户提交的一个核酸序列或蛋白质序列，拿去同指定的数据库中的全部序列作比较。它们的使用方法也大同小异：可用电子邮件提交序列并指定各种参数，可以在浏览器里填表提交作业，还可以把它们下载到自己的计算机上运行，不过这时要备有所需的数据库。比 BLAST 和 FASTA 晚出现 10 年的 BLITZ [R-651] 服务器，使用 Smith-Waterman [R-623] 算法，拿用户提交的蛋白质序列同数据库中的序列作比较。另一个集成了 BLAST 、FASTA 和 Smith-Waterman 算法、既可以比较蛋白质也可以比较 DNA 序列的服务器，是美国橡树岭国家实验室的 GenQuest

[R-652]。它们的用法与 BLAST 和 FASTA 相似，放在这一节里一起叙述。还有一些专门显示 BLAST 或 FASTA 输出结果的程序，也在本节最后提及。

一般认为，BLAST 运行速度快，对蛋白质序列的搜寻更为有效；FASTA 运行较慢，对核酸序列更为敏感。蛋白质序列的比较，往往可以揭示 20 亿～30 亿年前分道发展的同源关系，而 DNA 序列的比较只能回溯 2 亿～5 亿年[27]。因此，通常应先做蛋白质序列的比较，再对比核酸序列。只要条件允许，就应当 BLAST 和 FASTA 双管齐下，兼收并用。

用电子邮件提交序列时，用户应当知道如何选取参数和阅读服务器送回来的结果。不同的参数选择，可能返回差别极大的结果。我们在下面分别介绍 BLAST、FASTA、BLITZ 和 GenQuest 四种电子邮件服务的使用方法。

5.3.1 BLAST 服务

BLAST 是“基本局域联配搜寻工具”(Basic Local Alignment Search Tool") 的字头缩写。BLAST 算法的最初描述见：

R-631 S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.* **215** (1990) 403 – 410.

最初的 BLAST 进行序列联配时不容许插入空位，后来取消了这一限制。参见：

R-632 S. F. Altschul 等 7 位作者，"Gapped BLAST and PSI-BLAST: a new generation of protein database search program", *Nucleic Acids Res.* **25** (1997) 3389 – 3402.

上文标题中的 PSI 是 Position Specific Iterated 的缩写。据 1999 年 4 月 12 日出版的 *The Scientist* 杂志报道，此文发表两年引用超过 500 次。因此，牛津大学出版社特别把它放到网页上，供学者免费下载。网址：

R-633 http://nar.oupjournals.org/

所谓 PSI，是先从给定的打分矩阵出发，由库中选出相似性高的一批序列，把这些序列同给定序列作联配，计算出新的打分矩阵，再到库中

[27] 参看 W. R. Pearson 在 [R-23] 一书第 186 页的论述。

搜索．这样叠代到结果收敛或达到一定次数．关于使用 PSI-BLAST 作递归搜索的较近描述见：

R-634 J. Gouzy, F. Corpet, and D. Kahn, *Comput. Chem.* **23** (1999) 333 - 340.

BLAST 是运行速度甚快的数据库搜索程序，许多生物信息中心都有专门运行 BLAST 的服务器．从 1998 年 10 月起， NCBI 服务器运行 Gapped BLAST [R-632]，又称 BLAST2.0．请注意，BLAST2.0 不同于华盛顿大学的 WU-BLAST2．下面 EBI 服务器运行的就是 WU-BLAST2．

R-635 主要的 BLAST 服务器网址如下：

`http://www.ncbi.nlm.nih.gov/BLAST/` (运行 BLASTR2.0)

`http://www.ebi.ac.uk/` (运行 WU-BLAST2)

`http://blast.wustl.edu/`

`http://www.blast.genome.ad.jp/`
(日本 GenomeNet，运行 BLAST2.0)

`http://rtcmain.rtc.riken.go.jp/BLAST/`
(日本理化研究所，运行 BLAST2.0)

`mailto:blast@ncbi.nlm.nih.gov`

`mailto:blast@ebi.ac.uk`

`mailto:blast@nig.ac.jp`

上面这些服务器的使用方法大同小异，主要差别在于所搜寻的本地 DNA 数据库．至于蛋白质库，大家都少不了 SWISS-PROT [R-401]、PIR [R-404] 和 PDB [R-441]．然而，数据库的具体组织和名称仍有差别．参数设置也有一些差别．使用任何一个服务器，都应事先弄明白这些细节。

用户可用电子邮件或通过网页向 BLAST 服务器提交序列，经过一段时间后获得搜寻结果．也可以把 BLAST 下载到本地计算机上运行，但要有相应的配套数据库．下面介绍用电子邮件提交序列到 NCBI 的 BLAST 服务器时，如何选择程序和设定参数．

第一，根据所提交的序列类型和要求搜索的数据库类型，以及是否要把核酸序列翻译成蛋白质，选取 BLAST 程序的一种工作方式或变种．BLAST 的五种可能的选择列在表 5.4 中．

表 5.4 BLAST 程序的几种工作方式

程序	查询序列	搜寻的数据库
BLASTN	核酸	核酸
BLASTP	蛋白质	蛋白质
BLASTX	核酸的 6 个读框	蛋白质
TBLASTN	蛋白质	核酸的 6 个读框
TBLASTX	核酸的 6 个读框	核酸的 6 个读框

从表 5.4 可见，在 BLAST 前面加 T 表示要求进行翻译，后面加 N、P 分别表示核酸和蛋白质库，X 则表示某种“交叉”比较．注意，运行 TBLASTN 和 TBLASTX 时，要对库中的大量 DNA 序列作 6 个读框的翻译，因而要求更多计算时间．TBLASTX 虽然是在比较核酸序列，但中间要按 6 个读框翻译成氨基酸序列，这样可以提高灵敏度．

第二，要规定请 BLAST 搜索哪些数据库．可能的选择列在表 5.5 中．

第三，要选择恰当的过滤程序．除 BLASTN 外，可以选用的过滤程序为 SEG、XNU 或其组合．SEG 过滤掉序列中的“低复杂度”区域，否则像 Poly(A)、Poly(T) 这样的片段会导致高分联配，漏掉真正的编码区．XNU 过滤简单的重复片段．BLASTN 只能选用或不用 DUST 过滤程序．另外还有需要与 RepBase [R-223] 数据库联合使用的 CENSOR 过滤程序．有些过滤程序可以单独访问．例如：

R-636 SEG 过滤程序，描述见：

J. C. Wootton, and S. Federhen, *Comput. Chem.* **17** (1993) 149 - 163.

网址：

`ftp://ncbi.nlm.nih.gov (/pub/seg/)`

请注意，SEG 的补缺参数原是为蛋白质序列设置的．对 DNA 序列应改为 window size 20 和 complexity threshold 0.8 .

R-637 XNU 过滤程序．网址：

`ftp://ncbi.nlm.nih.gov (/pub/xnu/)`

R-638 DUST 过滤程序．网址：

`ftp://ncbi.nlm.nih.gov (/pub/tatusov/dust/)`

R-639 CENSOR 过滤程序，描述见：

J. Jurka, P. Klonowski, V. Dagman, and P. Pelton, *Comput. Chem.*

表 5.5 BLAST 程序可以搜索的几种数据库

	数据库	说明
蛋白质数据库	NR	PDB+SWISS-PROT+PIR 的非冗余蛋白质序列和 GenBank 的翻译
	MONTH	最近 30 天内上述库中更新过的序列
	SWISSPROT [SP]	最新版 SWISS-PROT 蛋白质序列库
	YEAST	酿酒酵母蛋白质序列
	ECOLI	大肠杆菌基因组翻译的蛋白质序列
	PDB	PDB 库中的蛋白质序列
	KABATPRO [KABAT]	与免疫有关序列的 Kabat 库
	ACR	SWISS-PROT 库中的古老保守子集
	ALU	由部分 ALU 重复序列翻译出的数据库
核酸数据库	NR	GenBank+EMBL+DDBJ+PDB 的非冗余序列但不包括 EST, STS, GSS 和 HTGS
	MONTH	最近 30 天内上述库中更新过的序列
	YEAST	酿酒酵母基因组序列
	ECOLI	大肠杆菌基因组序列
	EST	GenBank+EMBL+DDBJ 中非冗余 EST
	STS	GenBank+EMBL+DDBJ 中非冗余 STS
	HTGS	高产出的基因组序列
	PDB	由 PDB 三维结构倒推的序列
	ALU	RepBase 中挑选的 ALU 序列

20 (1996) 119 - 121.

网址：

`ftp://ncbi.nlm.nih.gov (/pub/repository/censor/)`

还有 RepeatMasker [R-748] 程序，也值得参考。

第四，几个最重要的参数，需稍加说明。

期待 (EXPECT) 值 E 是假定所提交的序列和库中全部序列都是随机序列，所预期的符合数目。只有搜索时找到的期待值比 E 小的符合序列，才作为结果返回。E 值由用户给定，范围从 0~1000，补缺值是 10。

BLAST 程序返回的结果，由三部分组成。第一部分是对匹配序列的简单描述，由参数 DESCRIPTION 规定返回的行数，补缺值为 100。第二部分给出数据库中得分最高的序列的联配图，其组数由参数 ALIGNMENT 规定，补缺值为 50。第三部分是一个表明匹配统计的直方图，

由参数 HISTOGRAM 规定，补缺值是 YES．通过参数 ALIGNMENT、DESCRIPTION 和 HISTOGRAM 可以控制输出量或取消相应输出。

BLAST 的其余参数见表 5.6。

表 5.6 BLAST 程序的参数

参数	补缺值	可选值和说明
NEW	'TRUE'	'FALSE' 用老的 BLAST 1.4 版
PROGRAM	必需	见表 5.4
DATALIB	必需	见表 5.5
BEGIN	必需，且独占一行	后面随以 FASTA 格式的序列数据
EXPECT	10	可带小数点
CUTOFF	由 EXPECT 值算得	
MATRIX	BLOSUM62	PAM40，PAM120
		PAM250，IDENTITY
		BLASTN 不用 MATRIX
ALIGNMENTS	50	
DESCRIPTION	100	返回符合序列简短说明的行数
HISTOGRAM	YES	NO
ACKNOWLEDGE	120	
FILTER	DUST	NONE(BLASTN)
	SEG	NONE，XNU+SEG
		XNU, SEG+XNU
GCODE	1	按读框翻译时使用的密码表
		目前有 14 种选择
GAP_EXISTENCE	5	空位起始罚分 a
GAP_EXTEND	2	空位延长罚分 b
HTML	NO	YES：结果用 HTML 格式
NCBI_GI	NO	显示 NCBI 的 GI 序列编号
SPLIT	1000	电子邮件行数
STRAND	DOUBLE(BOTH)	SINGLE(TOP,PLUS,+)
		COMPLEMENTRAY(MINUS,-)
PATH	发信人 E-mail 地址	另一个 E-mail 地址

现在借助一个实例，演示如何用电子邮件提交查询序列。我们注意到在热球菌 *Pyrococcus abyssi* 的完全基因组中，有一个重复出现多次的长度为 18 的字母串 gttccaataagactaaaa．这本身已是远非平庸的事件，因为 *P. abyssi* 的全长 1 765 118 个字母的环形 DNA 如为随机序列，其中

特定的字长 18 的串出现一次的概率只有 $1\ 765\ 118\times4^{-18}=0.0\ 066$，重复出现的概率就更小了。我们发如下的电子邮件，请 BLAST 把这个短串与数据库中所有的 DNA 序列进行比较，看看它还在哪些序列中出现过（省略了一些次要语句）：

```
From: hao@itp.ac.cn
To: blast@ncbi.nlm.nih.gov
Subject:
PROGRAM  BLASTN
DATALIB      NR
BEGIN
> tmpseq_1 Pabyssi2
gttccaataagactaaaa
```

这封电子邮件中除了必须指定的程序 BLASTN 和数据库 NR 外，其余参数都使用系统设置的补缺值。返回的结果很长，我们只印出一小部分：

```
From blastsvc@miles.nlm.nih.gov Fri Jan 21 17:28 CST 2000
Date: Fri, 21 Jan 2000 04:11:39 -0500 (EST)
To: hao@itp.ac.cn
Subject: [E-mail Blast] tmpseq_1 Pyro2
From: BLAST E-Mail Server <blast@ncbi.nlm.nih.gov>
BLASTN 2.0.10 [Aug-26-1999]
Query= Pyro2
         (18 letters)
Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
           515,812 sequences; 1,484,651,443 total letters
                                                              Score     E
Sequences producing significant alignments:                  (bits)  Value

emb|AJ248288.1|CNSPAX06 Pyrococcus abyssi genome; segme...     36  0.019
emb|AJ248286.1|CNSPAX04 Pyrococcus abyssi genome; segme...     36  0.019
emb|AJ248283.1|CNSPAX01 Pyrococcus abyssi genome; segme...     36  0.019
dbj|AP000005|AP000005 Pyrococcus horikoshii OT3 DNA,994...     36  0.019
emb|X58253|LEUBI3 Tomato ubi3 gene for ubiquitin               32  0.30
......
```

根据 DESCRIPTION 值返回的 50 行简短说明，我们只印出前 5 行。$E=0.019$ 表明，它们绝不是偶然巧合。具体数值 0.019 的计算，需要知道一些“内部”算法，但我们很容易估计它的数量级。从上面返回的数据，

知道序列的平均长度是 1 484 651 443/515 812=2 878。假定没有环形序列，它们一共导致 (2 878-18+1)× 515 812 个长度为 18 的串。乘以某一个特定串出现的概率 4^{-18}，得到 0.086。这个从随机模型估计的值，与 0.019 相去不多。对于较长的询问序列，往往返回小于 10^{-100} 的 E 值，解读成“毫不偶然”就成了。

上面前 3 行都来自 EMBL 库中 *P. abyssi* 自己的完全基因组序列的不同片段，进一步验证了我们的观察，即 gttccaataagactaaaa 是重复出现多次的“字”。上面第 1 行简短说明对应如下的的详细说明：

```
>emb|AJ248288.1|CNSPAX06 Pyrococcus abyssi complete genome; segment 6/6
             Length = 265118
Score = 36.2 bits (18), Expect = 0.019
Identities = 18/18 (100%)
Strand = Plus / Plus
Query: 1      gttccaataagactaaaa 18
              ||||||||||||||||||
Sbjct: 260129 gttccaataagactaaaa 260146
......
```

它一共给出 27 组准确联配的图示，上面只印出第 1 组。第 4 行简短说明对应的情况中，也有 25 组是准确联配。但这个序列来自 DDBJ 库中另一个热球菌 *Pyrococcus horikoshii*，它与 *P. abyssi* 均为同一个属的细菌。所有其他结果中，都没有 100% 的准确联配。例如，第 5 行简短说明来自番茄的泛激素基因，但 18 个字母中只有 16 个完全匹配：

```
>emb|X58253|LEUBI3 Tomato ubi3 gene for ubiquitin
        Length = 2374
Score = 32.2 bits (16), Expect = 0.30
Identities = 16/16 (100%)
Strand = Plus / Minus
Query: 3   tccaataagactaaaa 18
           ||||||||||||||||
Sbjct: 576 tccaataagactaaaa 561
......
```

返回结果的最后部分，是这一次数据库搜寻所使用的参数（包括补缺值），以及一些计算出的参数值。我们只列出其中一部分。

```
Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
Posted date:  Jan 15, 2000  6:09 PM
Number of letters in database: 1,484,651,443
```

```
Number of sequences in database:  515,812
 Lambda     K      H
  1.37    0.711     1.31
 Gapped
 Lambda     K      H
  1.37    0.711     1.31
Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 4654
Number of Sequences: 515812
Number of extensions: 4654
Number of successful extensions: 4654
Number of sequences better than 20.0: 373
length of query: 18
length of database: 1,484,651,443
effective HSP length: 17
effective length of query: 1
effective length of database: 1,475,882,639
effective search space: 1475882639
effective search space used: 1475882639
T: 0
A: 0
X1: 6 (11.9 bits)
X2: 10 (19.8 bits)
S1: 12 (24.3 bits)
S2: 13 (26.3 bits)
```

上面所示的简单查询，根本未涉及基因及其产物的生物学解释。然而，它告诉我们，在各个核酸数据库迄今所收入的 515 812 个序列中，18 个字母的短串 gttccaataagactaaaa 只出现在 *Pyrococcus* 这一个属的两个细菌中，而且是高度重复出现。它能不能成为这个属的标记序列呢？

顺便提一下，美国橡树岭国家实验室的 GenQuest [R-652] 服务器也允许选择 BLAST 程序。

5.3.2 FASTA 服务

FASTA 是另外一套根据用户提交的单个序列进行数据库搜索的程序。一般认为，FASTA 对于核酸序列的比较更敏感一些。它的发展经历了三个阶段：

R-640 **FASTP** 是最早的版本，其描述见：

D. J. Lipman, and W. R. Pearson, *Science* **227** (1985) 1435 - 1441.

R-641 **FASTA** 是 FASTP 的改进版本，现在称为 FASTA2．描述见：

W. R. Pearson, and D. J. Lipman, "Improved tools for biological sequence analysis" *Proc. Natl. Proc. Acad.* USA **85** (1988) 2444 - 2448.

R-642 现行 3.0 版 FASTA，又称 FASTA3 的描述可以参看 [R23] 一书第 10 章．另外还可以参考 W. R. Pearson 本人的讲义 [R-836]．FASTA2 的某些程序尚未在 FASTA3 中实现，两者的源程序都是公开的．下载网址是：

`ftp://ftp.verginia.edu (/pub/fasta/)`

使用 FASTA 服务的作者，均请引用 [R-641] 所列文章．

许多生物信息中心提供 FASTA 数据库搜索服务，用户只需提交序列，即可获得结果．对于比较长的序列，即使通过网页上载，也只能用电子邮件收取结果．如果想要把 FASTA 下载到本地计算机上运行，可访问网址 [R-642]．

运行 FASTA 的 WWW 网页服务器和电子邮件服务器很多，例如：

R-643 部分运行 FASTA 的 WWW 服务器和电子邮件服务器的 URL：

`http://www.ebi.ac.uk/`

`mailto: fasta@ebi.ac.uk`

`http://www.fasta.genome.ad.jp/` (日本 GenomeNet)

`mailto: fasta@nig.ac.jp` (日本国立遗传所，运行 fasta3)

`http://rtcmain.rtc.riken.go.jp/fasta/` (日本理化所)

`http://www.rtc.riken.go.jp/pdb/index.html`

(日本理化研究所，运行 fasta3，只搜索 PDB 库)

`http://www2.ebi.ac.uk/fasta3/`

`http://iubio.bio.indiana.edu/search/fasta`

我们仍以电子邮件服务为例，介绍一下如何用 FASTA3 搜索数据库．同 BLAST 一样，FASTA 也有几种工作方式，表 5.7 列举其新版 3.0 所包含的程序即工作方式．

表 5.7 FASTA 程序的几种工作方式

程序	查询序列	搜寻的数据库
fasta3	DNA	DNA
	蛋白质	蛋白质
ssearch3	DNA	DNA
	蛋白质	蛋白质
fastx3	DNA 6 个读框，允许密码子间读框错位	蛋白质
fasty3	DNA 6 个读框，允许密码子内读框错位	蛋白质
tfastx3	蛋白质	DNA 的 6 个读框
tfasty3	蛋白质	DNA 的 6 个读框
fasts3	质谱仪多肽链	蛋白质
fastf3	混合多肽序列	蛋白质
tfastf3	混合多肽序列	DNA 翻译蛋白序列

各个生物信息中心的 FASTA 服务的差别，主要在于所搜索的数据库品种．例如，欧洲生物信息研究所 EBI 的 FASTA 服务，可以搜索表 5.8 中列举的数据库．

FASTA3.0 版设置了大量补缺参数，见表 5.9．用户甚至无须指定程序名字．FASTA 服务器根据所提交序列中的字母类型判断是 DNA 或蛋白质，大小写字母可以混用．

仍然使用前面讲 BLAST 时的例子．这封电子邮件完全使用服务器设置的补缺参数，因此颇简短：

```
From: hao@itp.ac.cn
To: fasta@ebi.ac.uk
Subject:
SEQ
gttccaataagactaaaa
END
```

FASTA 服务器自动选择 fasta3_t 程序搜索 EMALL 库，实际上比较了 23 个核酸数据库中的 5 655 840 个序列，总计 5 903 562 019 个碱基．由于提交的序列很短，参数 WORD(即 ktup)=1．返回的结果中，18 个字母 100% 严格匹配的序列，除了来自 *P. abyssi* 和 *P. horikoshii*，还增加了一个 *P. furiosus*．三者都是热球菌属的细菌．这更为加强了我们的猜测，即 18 个字母的寡核苷酸序列 gttccaataagactaaaa 可以作为该属

表 5.8 EBI 的 FASTA 程序可以搜索的数据库

数据库		说明
蛋白质数据库	SWALL	来自 PDB 和 SWISS-PROT 的非冗余蛋白质序列，PIR 和 GenBank 的翻译
	SWISSPROT	最新版 SWISS-PROT 蛋白质序列库
	SWNEW	上述库中最近更新的序列
	TREMBL	从 EMBL 翻译出的序列
	TREMBLNEW	TREMBL 中的最新序列
核酸数据库	EMBL	EMBL 核酸数据库
	EFUN	EMBL 真菌
	EINV	EMBL 无脊椎动物
	EHUM	EMBL 人类
	EMAM	EMBL 哺乳动物
	EORG	EMBL 细胞器
	EPHG	EMBL 噬菌体
	EPLN	EMBL 植物
	EPRO	EMBL 原核生物
	EROD	EMBL 啮齿动物
	ESTS	EMBL 序列标记
	ESYN	EMBL 合成序列
	EUNA	EMBL 未分类序列
	EVRL	EMBL 病毒
	EVRT	EMBL 脊椎动物
	EEST	EMBL 已表达的序列标记
	EGSS	EMBL 基因组总结序列
	EHTG	EMBL 高产出基因组序列
	EMNEW	EMBL 最近更新的序列
	EMALL	EMBL + EMBLNEW
	IMGT	免疫遗传学数据库
	EVEC	由 EMBL 导出的载体序列

表 5.9 FASTA 程序的参数

参数	补缺值	可选值和说明
HELP		取使用说明
PATH		返回结果的另外地址
TITLE		作业标题
PROGRAM	fasta3	表 5.7 中列出的程序
LIB	EMALL 或 SWALL	表 5.8 中列出的数据库
MATRIX	BLOSUM62	PAM250
WORD(ktup)	2	对蛋白质序列
	6	对 DNA 序列
ALIGN	25	
LIST	50	
STRAND		both,top,bottom(只对 DNA)
HISTOGRAM	NO	YES
SEQ	不能缺省	
END		结束标志

的标记.

我们注意到, EBI 的 FASTA 服务器搜索的核酸序列总数, 比 NCBI 的 BLAST 多 10 倍, 碱基总数多 4 倍, 而且确实返回来 BLAST 没有找到的结果. 这也印证了前面所说, 对于 DNA 序列 FASTA 往往更敏感, 而且搜索数据库时最好几种服务器多管齐下, 比较结果.

这个例子还可以说明序列比较问题的特点, 即演示一个已知结果是不难的. 但要从长度为 100 万字母的细菌基因组中把可能作为标记的寡核苷酸序列找出来, 就不是发两封电子邮件能做到的.

FASTA3 还包含几个不能在电子邮件中调用的程序, 他们都涉及对联配结果的统计评估. 这里只作扼要介绍:

R-644 **PRSS3** 程序对两个蛋白质或 DNA 序列的联配结果进行统计评估, 办法是不断把第二个序列随机地打乱, 用 Smith-Waterman 算法求相似分数, 然后估计极值 (extremee value) 分布 (也叫 Gumbel 分布) 的参数. 此法只适用于两个序列的局域联配. 整体联配或多序列联配的统计分布性质尚不清楚. 请参看:

S. F. Altschul, M. S. Boguski, W. Gush, and J. C. Wootton, *Nature*

Genetics **6** (1994) 119 – 129.

R-645 **sc_to_e** 程序根据序列长度、数据库长度、相似性的分数等，计算搜索结果的统计显著性水平。

R 646 **randseq** 程序产生一个长度和组分都与所提交的序列相同的随机序列。随机序列主要用于对数据库搜索结果的统计评估。

顺便提一下，美国橡树岭国家实验室的 GenQuest [R-652] 服务器也允许选择 FASTA 程序。

5.3.3 与 BLAST 和 FASTA 有关的后处理程序

BLAST 和 FASTA 的输出，都是普通的纯文本文件。有一些程序可对这些输出做后处理，主要是视象化表示。

R 647 **Blixem** 是 BLAST 输出的视象化程序，它要求先调用 MSPcrunch [R 648] 程序对数据进行过滤。描述见：

E. L. L. Sonnheimmer, and R. Durbin, *CABIOS* (*Bioinformatics*) **10** (1994) 301 – 307.

Blixem 的安装见：

`http://www.cgr.ki.se/cgr/groups/sonnhammer/Blixem.html`

R 648 **MSPcrunch** 是 BLAST 输出送往 Blixem [R-647] 显示之前的过滤程序，可从 NCBI 的 ftp 子目录下载。使用说明见：

`http://www.cgr.ki.se/groups/sonnhammer/MSPcrunch.html`

R 649 **Visual BLAST** 和 **Visual FASTA**，这是 P. Durand 等为分析 BLAST 和 FASTA 输出的蛋白质序列联配结果而编写的视象化程序，其描述见：

P. Durand, L. Conard, and J. P. Mormon, *CABIOS* (*Bioinformatics*) **13** (1997) 407 – 413.

此程序只在 PC 视窗系统 (95/98/NT) 下运行。程序可免费从以下网址下载：

`http ://www.lmcp.jussieu.fr/~ durand/`

网址中的 LMCP 是 Laboratoire de Mineralogie-Cristallographie de Paris，即巴黎矿物晶体研究室的缩写。这套程序的新版请参看 [R-650]。

R-650 **Octopus** 是 Visual BLAST 和 Visual FASTA 程序新版合并后的别名，适用于 PC 视窗系统 (95/98/NT)，必须联网使用。它包括显示、疏水性分析、多序列编辑等部分。需要者须向 P. Durand 提出请求：
mailto: durand@lmcp.jussieu.fr

5.3.4 BLITZ 服务

R-651 **BLITZ**，欧洲生物信息研究所 EBI [R-131] 提供的服务器。它使用 Smith-Waterman [R-623] 算法，把用户提交的序列送到 SWISS-PROT [R-401] 和 TrEMBL [R-402] 蛋白质序列数据库去搜索比较。这是一套灵敏而且速度很快的程序。电子邮件地址：
mailto: blitz@ebi.ac.uk

用电子邮件提交序列给 BLITZ 服务器的方法，同 FASTA 很相像。电子邮件主体中，每个参数占一行。除了 SEQ 及其后的序列之外，其他参数都可以省略，由服务器自动使用补缺值。表 5.10 中列出了主要参数以及它们的补缺值和可选值。

表 5.10 BLITZ 程序的参数

参数	补缺值	可选值或说明
HELP		取得 HELP 文件
DATABASE	SWALL	SWISSPROT，SWNEW，TREMBL TREMBLNEW，SPTR
PAM	100	150，200，250
BLOSUM	62	无其他选择，但可换 PAM
GAPCOST	15	空位罚分 a：5–20
GAPXCOST	1	空位延长罚分 b：0.05–2.0
ALIGN	10	送回的最佳联配数，最多 100
NAMES	50	回报的打分个数
TITLE	可有可无	本次作业标题，不含引号
SEQ	必需	独占一行，后随序列数据，可含空行空格，大小写均可
END		结束标志

5.3.5 GenQuest 服务

GenQuest 是美国橡树岭国家实验室提供的集成的数据库搜索服务。调用 GenQuest 的方法有四种：使用电子信；通过与 GRAIL 的共同网页，见 [R-719]；通过基于 X [R-52] 的一个在本地计算机上运行的服务器界面；或从 XGRAIL [R 719] 程序中调用。这里只介绍电子邮件服务。

R-652 **GenQuest** 电子邮件服务：

```
mailto:  Q@ornl.gov
```

GenQuest 把用户提交的一个 DNA 序列或蛋白质序列同指定数据库中的全部序列做比较。它允许用户选择 BLAST 、 FASTA 、 Smith-Waterman(SW) 等多种算法，以 SW 为补缺值。橡树岭实验室准备今后把其他行之有效的搜索方法继续集成到这个服务器中。GenQuest 电子邮件的参数见表 5.11 。

表 5.11 GenQuest 电子邮件的参数

参数	补缺值	可选值和说明
TYPE	必需	PROTEIN 、 DNA 、 DNA6
SEQ	必需	独占一行，后随序列数据
END	可有可无	序列结束标志
FILTER	不写	对提交的序列先行过滤
MATRIX	BLOSUM62	BLOSUM80 、 PAM100±10 × n 只用于蛋白质序列
METHOD	SW	FASTA 、 BLAST 、 FLASH
ALIGN	10	仅用于 METHOD 选 SW 时
SCORE	10	仅用于 METHOD 选 SW 时
TARGET	GSDB(对 DNA)	DBEST 、 REPITITIVE
	SWISSPROT(对蛋白质)	PDB,PROSITE,PIR,BLOCKS
COMMENT	可有可无	作业的注解、标题
HELP		取得 HELP 文件

§5.4 多序列联配程序

动态规划方法原则上可以推广到多序列联配，但对计算机的处理能力要求甚高。如果只要联配少数几个不很长的序列，可以把这些序列通过网页提交到以下服务器：

R-653 **BCM** 服务器，即 Baylor College of Medicine 所提供的 BCM Search Launcher 服务。网址：

`http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html`

R-654 瑞士 **ETH** 服务器：

`http://cbrg.inf.ethz.ch/MultiAlign.html`

更常见的多序列联配程序，使用“逐步联配法” (progressive alignment)，由简到繁，先把彼此最接近的序列联配成对，根据它们之间的距离用邻接法 [R-671] 形成导引树 (guide tree)，再按“演化”顺序实现全部序列的联配。还可以用自举法 [R-676] 对联配结果作统计评估。目前使用得最广泛的免费逐步联配程序，是早年由 D. G. Higgins 开始编写的 Clustal 系列程序：

R-655 D. G. Higgins, and P. M. Sharp, *Gene* **73** (1988) 237 – 244.

Clustal 程序接受多种输入格式，包括 FASTA、EMBL、SWISS-PROT、PIR 和 GCG/MSF 等，但所有输入序列必须在同一个文件中。如果输入序列的非空格符号中 85% 以上是 A、C、G、T、U、N(大小写均可)，就判定为核酸序列，否则作为蛋白质序列。但蛋白质序列和核酸序列不可混在一个文件里。Clustal 的输出文件也有多种格式供选择。

R-656 **ClustalW** 多序列联配程序使用纯文本对话控制输入、输出和参数选择。ClustalW 的描述见：

J. D. Thompson, T. J. Gibson, and D. Higgins, *Nucleic Acids Res.* **22** (1994) 4673 – 4680.

ClustalW 有适用于多种平台的源程序和可执行文件，可从以下网址下载：

`http://www.ebi.ac.uk/dos/clustalw/`

`http://iubio.bio.indiana.edu/align/clustal/`

ftp://ftp.ebi.ac.uk (/pub/software/)

ftp://ftp-igbmc.u-strassbg.fr (/pub/Clustal/)

R-657 **ClustalX** 是 ClustalW[R-656] 多序列联配程序的 UNIX 版本，它使用 X 窗口图形界面，图形显示在一个单独窗口中，备有选择菜单。程序描述见：

J. D. Thompson, T. J. Gibson, F. Plewiak, F. Jeanmongin, and D. G. Higgins, *Nucleic Acids Res.* **25** (1997) 4876 – 4882

下载网址：

ftp://ftp.ebi.ac.uk (/pub/software/)

ftp://ftp-igbmc.u-strassbg.fr (/pub/ClustalX)

Clustal 程序没有对多序列联配文件进行编辑加工的功能。当然可以使用普通的编辑程序作这件事，但不如调用专门的多序列联配编辑程序方便，例如：

R-658 **SeaView** 多序列联配编辑程序，描述见：

N. Galtier, M. Gouy, and C. Gautier, *CABIOS* (*Bioinformatics*) **12** (1996) 543 – 548.

网址：

ftp://biom3.univ-lyon1.fr (/pub/mol_phylogeny/)

R-659 **BOXSHADE** 程序，把多个蛋白质序列联配结果用阴影或颜色加工成适于发表的形式，获取有关信息的网址是：

http://www.isrec.isb-sib.ch/software/BOX_form.html

也可向以下网页提交序列，进行加工：

http://www.ch.embnet.org/software/BOX_form.html

R-660 **CINEMA**，交互式的彩色多序列编辑程序，网址：

http://www.bioinf.man.ac.uk/dbbrowser/

CINEMA2.1/cinema2hdr.html

R-661 **AMAS** 多序列分析程序包。它不能直接读取 Clustal 格式，但接受 PIR 格式的输入文件，描述见：

C. D. Livingstone, and G. J. Barton, *CABIOS* (*Bioinformatics*) **9** (1993) 745 – 756.

网址：

http://barton.ebi.ac.uk/servers/amas-server.html

R-662 **belvu**，是由 KISAC [R-145] 的 Eric Sonnhammer 编写的一个多序列联配的显示程序，适用于 UNIX 平台，它允许用户选择显示氨基酸的颜色，但不是一个功能齐全的序列编辑程序。详见：

`http://www.cgr.ki.se/cgr/groups/sonnhammer/Belvu.html`

可由以下网址 获取：

`ftp://ftp.cgr.ki.se (/pub/prog/SFS/)`

R-663 **LalnView** 程序显示序列联配的结果。描述见：

L. Duret, E. Gasteiger, and G. Perriere, *CABIOS* (*Bioinformatics*) **12** (1996) 261 – 282.

网址：

`ftp://ftp.expasy.ch (/pub/lalnview/)`

顺便指出，SeqPup[R-714] 程序也有对多序列联配结果进行编辑加工的功能。Clustal 程序虽有构建亲缘树的功能，但为此最好使用 §5.5 节中介绍的专用程序。

与逐步联配不同的另一种策略，是先联配序列中的保守片段，然后再把它们组装起来：

R-664 **Dialign** 程序，描述见：

B. Morgenstern, A. Dress, and T. Werner, *Proc. Natl. Acad. Sci.* USA **93** (1996) 12098 – 12103.

网址：

`http://bibiserv.techfak.uni-bielefeld.de/dialign/`

§5.5 亲缘树的计算和图示

演化是生物学的基本概念。许多生物学的事实表明，所有现存物种来自同一祖先，不同的核酸或蛋白质序列可能源于同一原始序列。亲缘关系远近的判别，曾经主要基于形态学的观察，因而与分类学密切相关。亲缘关系研究的一组特定对象，称为“操作性分类单元” OTU (Operational Taxonomic Units)。OTU 可以是基因、核酸序列、蛋白质序列、个体、物种或种群。分子生物学的发展，把亲缘关系的研究推进到分子水平。分子演化导致的亲缘分析，是生物信息学的重要篇章。然而，要使分子演化

理论有比较坚实的数学基础，还必须考察其基本假设。

第一，是对突变的认识。历史上占优势的是达尔文以来的选择演化观点，认为在自然选择的压力下，除了有害突变，就是有益突变，中性的、不好不坏的突变很少。20 世纪 60 年代后期，木村 (Motoo Kimura) 等人提出了中性演化观点，认为突变是与选择无关、随机产生的，除了有害突变，大部分都是中性的，有益突变很少。中性演化并不否认选择的作用，而是认为在长期自然选择的基础上，各个历史时期的物种都已接近当时条件下最优的水平，发生继续优化的有益突变的可能性甚微。选择演化观点很难严格表述，而中性演化理论却可以有较好的数学基础。两种观点引起激烈、持久的争论，中性演化逐步赢得较多支持。中性演化观点的早期文章见：

R-665 M. Kimura, "Evolutionary rate at the molecular level", *Nature* **217** (1968) 624 – 626.

R-666 J. L. King, and T. H. Jukes, "Non-Darwinian evolution: random fixation of selectively neutral mutations", *Science* **164** (1969) 788 798.

第二，分子钟假设，即核酸序列中每个碱基或蛋白质序列中每个残基都以恒定的速率发生突变。这个假设从整体上说是不成立的。核酸序列中非编码部分比编码部分突变速率高，内含子比外显子突变速率高。然而，如果选择特定的某种基因序列，突变速率则近似保持恒定。一般说来，蛋白质序列中的突变速率更为接近恒定。人们对分子钟假设作过各种检验，但争论仍在继续。

第三，是具体代表一个 OTU 的“矢量”的长度不能太短，否则容纳不了演化的历史过程；或者说，变异的积累达到“饱和”。在下一小节介绍过距离概念之后，我们再继续讨论这一点。这里先开列三本关于分子演化的较新参考书：

R-667 Wen-Hsiung Li, *Molecular Evolution*, Sinauer Associate, Inc., 1997, xv + 487.

R-668 Roderic D. M. Page and Edward C. Holmes, *Molecular Evolution: A Phylogenetic Approach*, Blackwell, 1998, 1999, v + 346.

R-669 Wen-Hsiung Li, and Dan Graur, *Fundamentals of Molecular Evolu-*

tion, Sinauer Associate, Inc., 1991, 1999.

后两本书比较简明易读，而第一本对分子演化理论的基础讨论较深。

5.5.1 距离和相异性

把一批同类对象 (OTU) 按照相似性分成聚类 (clustering)，是分类学 (taxonomy) 的基本要求。分类的关键首先在于确定一种或多种比较每个 OTU 用的性状。例如，对人群做身体检查，每人测 10 个指标。如果，对每项指标只问是“阴性” (0) 或“阳性” (1)，那每人的检查结果由 10 个 0 或 1 组成的一个矢量代表，人与人之间的差别由相应矢量之间的距离刻画。距离近者可聚成一类，距离远者差异也大。

定义“距离”的办法很多。我们先看一下距离应具备的基本性质。设有 A、B 和 C 三个 OTU，它们之间的距离分别记作 $d(A,B)$、$d(B,C)$ 和 $d(A,C)$。正确的定义，应当满足以下三条要求 (距离公理)：

第一，自己到自己的距离为 0： $d(A,A)=0$。

第二，从 A 到 B 的距离等于从 B 到 A 的距离： $d(A,B)=d(B,A)$ (对称性)。

第三，任意两个距离之和应当等于或大于第三个距离。这很容易想象成一个三角形三个边的关系，因此又称为三角形不等式。

距离代表相异性，但习惯上用来表示相异性的量，不一定满足上述公理。使用满足距离公理的定义，数学上有一些好处。事实上，从亲缘树出发可以定义物种之间的距离。这些距离不仅满足上述公理，而且满足更强的“超测度” (ultrametricity) 条件 (“测度”是和距离差不多的数学概念，读者可以不问细节)。超测度与中性演化和分子钟假定有密切关系。我们稍作介绍。

设有一株层次清楚的亲缘树。所有现存物种处于同一最底层，往上追溯到各代祖先。把两个现存物种到达共同祖先的“代”数，定义为它们之间的距离。不难看出，任意三个物种之间的距离，必有两个距离相等，且等于或大于第三个距离。或者说，上面三角形不等式中的三角形，只能是等腰或等边三角形。这就是超测度。超测度满足距离公理，只是三角形不等式的表现形式更具体。

早在 20 世纪 50 年代，分类学者就提出过一种系统的聚类方法，把两个小聚类成员间的最小距离取为聚类之间的距离。这样可以由下而上地构造出唯一的亲缘树。相应的约化后的距离比原始距离小，而且满足超测度条件。还可以取两个小聚类成员间的最大距离为聚类之间的距离，这样得到的亲缘树虽不唯一，约化距离也比原始距离大，但仍满足超测度关系。

在中性演化背景下，以及分子钟假定即突变速率恒定条件下，且代表每个 OTU 的矢量的分量数目趋向无穷多，则各个 OTU 之间的距离趋向超测度。来自实际生物序列的距离矩阵，不会满足超测度条件，但可以把它与超测度的差别，作为对分子钟假设的一种检验。

超测度的概念虽然有些抽象，但它为亲缘树的构建提供了一个理论上严格的框架。下面这篇关于超测度的综述文章，虽是写给物理学者的，但它有很长一节专门讨论生物分类和亲缘树。特别对于历史发展，有简要描述，值得一读。

R–670 R. Rammal, G. Toulouse, and M. Virasoro, "Ultrametricity for physicists", *Reviews f. Modern Phys.* **58** (1986) 765 – 788.

5.5.2 亲缘树算法简介

常用的亲缘树算法有两大类。一类是简单的聚类方法，另一类是体现某种优化要求的方法。后者往往属于 NP 完备问题，因而不得不寻求近似解法。我们只简要地提及一些方法的名字，而不叙述计算公式和理论背景。

聚类方法中常见的是 NJ 和 UPGMA 法。它们常用来迅速构建亲缘关系，然后再靠其他方法改进。

R–671 **NJ** 是邻接 (Neighbour Joining) 方法的缩写，它可以很快地导致单一的亲缘树。最初描述见：

N. Saito, and M. Nei, "The neighbor joining method: a new method for reconstructing phylogenetic trees", *Mol. Biol. Evol.* **4** (1987) 406 – 425.

R–672 **UPGMA** 是使用算术平均的不加权的成对分组方法 (Unweighted Pair Group Method with Arithmetic Means)。用此法构建的亲缘树

满足超测度关系.

对亲缘树与原始数据的拟合提出一定优化要求的方法，目前使用得比较普遍. 常见的名目有:

R-673 最大简约法 (Maximal Parsimony，简称 MP).

R-674 最短演化长度法 (Minimal Evolution，简称 ME).

R-675 最大似然法 (Maximal Likelihood，简称 ML).

无论用何种方法构建的亲缘树，都有一个对其统计置信度进行评估的问题. 这里最常用的是:

R-676 自举法 (bootstrap)，即对代表 OUT 的矢量用某种方式随机取样，多次重复构建亲缘树并观察所得结果稳定性的一套办法.

5.5.3 亲缘树计算软件

R-677 **PHYLIP** 程序包. 华盛顿大学 Joe Felsenstein 编写的这套 PHYLogeny Inference Package，是最常用的免费亲缘树计算软件，由大约 30 个程序组成. 它已有 15 年以上历史，最初描述见:

J. Felsenstein, *Evolution* **39** (1985) 783 - 791.

PHYLIP 程序包 1999 年 10 月发行第 3.5 版. 程序和使用说明书可一起下载:

`http://evolution.genetics.washington.edu/`

`ftp://evolution.genetics.washington.edu/`

文件 `phylip.tar.Z` 包含 C 源程序和说明书，而 `phylip.exe` 是适用于 DOS 的自动解压安装的程序.

R-678 **PAUP** 是简约法亲缘分析 (Phylogeny Analysis Using Parsimony) 的缩写. 它目前已是 GCG [R-792] 商业软件包的一部分.

R-679 **Phylo_Win** 是在 UNIX 平台上运行的一套免费亲缘树计算程序. 它的描述和网址均请参看 SeaView[R-658] 的说明.

R-680 **NJBafd** 程序从 DNA 中微卫星序列位点频度或其他遗传标记出发，用邻接法或 UPGMA 法构建亲缘树. 网址:

`ftp://iubio.bio.indiana,edu (/soft/molbio/evolve/njbafd)`

R-681 **NJPlot** 是根据文件绘制亲缘树的程序. 文件中对树的描述采用括号嵌套的形式. PHYLIP 程序就可以输出这样的文件. NJPlot 也可

以输出 PostScript 图形文件. 网址:

ftp://biom3.univ-lyon1.fr (/pub/mol_phylogeny/njplot)

R-682 **TreeView** 是 Rod Page 编写的亲缘树显示程序. 网址:

http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

R-683 **Phylodendron**, 是 D. Gilbert 编写的专门绘制亲缘树的程序, 它可以避免树枝交叉等不大雅观的输出. 网址:

http://iubio.bio.indiana.edu/java/apps/trees/

R-684 **PAML**, 由杨子恒编写的最大似然法亲缘分析 (Phylogenetic Analysis by Maximal Liklihood) 程序, 可以由核酸或蛋白质序列出发, 进行模型拟合或亲缘树的构建. 网址:

http://iubio.bio.indiana.edu/evolve/paml

ftp://abacus.gene.ucl.ac.uk (/pub/paml)

R-685 **Phyltest**, S. Kumar 编写的亲缘假设检验程序, 它可以比较三种亲缘树, 估计每对物种的平均距离等. 下载网址:

http://iubio.bio.indiana.edu/ibmpc/phyltest

R-686 **malign**, 多序列联配和亲缘树计算的服务器. 设在日本的 DDBJ [R-213], 可用电子邮件提交序列, 参数见表 5.12. 电子邮件中每个参数占一行. 其算法见:

J. J. Hein, *Methods Enzymol.* **183** (1990) 626 – 645. (见 [R-26])

网址:

mailto: malign@nig.ac.jp

表 5.12 malign 程序的参数

参数	补缺值	可选值和说明
ances		每个节点印一条祖先序列
gapa	8	空位罚分
gapb	3	空位延长罚分
moltype	DNA	protein
tree		印出亲缘树
begin	必需	独占一行, 后随两个以上 FASTA 格式的序列

互联网上还有一些分类学和亲缘关系的软件目录, 可以参考:

R-687 英国 Glasgow 大学的分类学软件目录:

http://taxanomy.zoology.gla.ac.uk/software.html

R-688 牛津大学动物学系 Paul Harvey 研究组的亲缘关系和种群遗传学软件目录:

http://evolve.zps.ox.ac.uk/software.html

R-689 **Tree of Life**, "生命之树"网点, 是一个多位作者合作的计划:

http://phylogeny.arizona.edu/tree/programs/programs.html

§5.6 与 DNA 测序和基因工程有关的软件

用于 DNA 序列大规模测序的软件包, 在几个主要的国际测序中心发展. 它们都是相当复杂的系统, 并且带有详细的使用说明书. 这里只能点名简介.

R-690 **Staden** 程序包是以 Rodger Staden 为首的小组经多年锤炼改进而成的. 它主要用于大规模 DNA 序列测序, 但其中一些程序可单独使用. 其最新的 2000.0 版可从以下网址下载:

http://www.mrc-lmb.cam.ac.uk/pubseq/downloads.html

ftp://ftp.mrc-lmb.cam.ac.uk

/pub/staden/downloads/staden_OS_RELNUM.tar.gz

文件名字中的 OS 是所需 UNIX 操作系统名, 如 solaris; RELNUM 是版本号如 2000.0. 所取得的程序包均在演示模式下运行. 无论学术性或商业性用户, 均须由网页

http://www.mrc-lmb.cam.ac.uk/pubseq/licence.html

得到使用许可, 才能全功能运行. Staden 程序包有长达 500 页的使用说明书, 可以 PostScript 文件打印, 或作为 HTML 文件在浏览器中阅读. 使用说明书的网址是:

http://www.mrc-lmb.cam.ac.uk/pubseq/

下面是 Staden 程序包中的几个主要程序:

1. PREGAP 对自动测序机的读数结果进行预处理.

2. GAP4 对测序得到的序列片段进行拼接组装.

3. SIP4 取代了原来的 SIP 程序. 这个序列比较和联配程序, 最初以 DIAGON 名称发表:

R. Staden, "An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences", *Nucleic Acids Res.* **10** (1982) 2951 - 2961.

4. NIP4 取代了原来的 NIP 程序。这个寻找读框和基因的程序，最初发表时的名字是 ANALYSEQ 。见：

R. Staden, "Graphic methods to determine the function of nucleic acid sequences", *Nucleic Acids Res.* **12** (1984) 521 - 538.

R-691 华盛顿大学为大规模 DNA 测序发展了一整套程序系统，它的主要模块有：

1. Phred 测序程序，它实现碱基识别和错误率估算。

2. Phrap 组装程序，拼接组装 Phred 提供的短片段，并进行质量评估。

3. Consed 校对程序，结合人工校对。

学术性用户原则上可以免费使用这套程序，但事先必须与发展 Phrap 系统的 P. Green 实验室取得联系，做出学术性使用的承诺。详细情形请参看网址：

http://www.phrap.org/

R-692 **CAP、CAP2 和 CAP3**，是黄晓秋编写的相邻片段组装程序 (Contig Assembly Program) 。它能够灵敏地查找片段的重复部分。算法和程序描述见：

X. Huang, "An improved sequence assembly program", *Genomics* **33** (1996) 21 - 31.

这个用 C 语言编写的命令行程序，可以从 GAP4 [R-690] 、 SeqPup [R-714] 等程序中调用。网址：

http://genome.cs.mtu.edu/cap/cap3.html

ftp://cs.mtu.edu (/pub/huang/)

其最新版本 CAP3 的文件在：

http://genome.cs.mtu.edu/sas.html

需要此程序者应直接与黄晓秋联系：

mailto: huang@mtu.edu

R-693 **Primer3** 是一个 PCR 和测序所需引物的设计程序。详细描述见 [R 23] 一书第 20 章。可以直接从 WWW 网页享用这一服务。网址：

http://www.genome.wwi.mit.edu/cgi-bin/primer/info.cgi/

也可以下载源程序自己实现：

ftp://genome.wi.mit.edu (/pub/software/)

R-694 **PrimerDesign** 引物设计程序．网址：

ftp://ftp.chemie.uni-marburg.de (/pub/PrimerDesign/)

R-695 **Primer-Master** 引物设计程序．网址：

http://www.ebi.ac.uk/software/software.html

ftp://ftp.ebi.ac.uk (/pub/software/)

商业性的 GCG 程序包 [R-792] 中也有几个与测序拼装有关的程序。

§5.7 DNA 序列分析程序

在大规模测序所得到的 DNA 序列中判认基因和对基因表达起调控作用的各种蛋白质结合位点，是序列分析的核心任务．DNA 序列的高产出，使得计算机判认成为不可替代的手段．

原核生物 DNA 中基因密度较高，而且基本上没有内含子，因此排除掉较易确定的 RNA 基因和简单重复序列之后，较长的开放读框 (Open Reading Frame，简称 ORF) 通常就对应基因。只有很短的基因容易被漏掉．为此可使用各种马可夫模型加以补救．

关于从 DNA 序列中识别基因的方法，可以参考近几年的一些综述文章：

R-696 J. W. Fickett, "Finding genes by computer: the state of the art", *Trends Genet.* **12** (1996) 316 – 320.

R-697 J. M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences", *Hum. Mol. Genet.* **6** (1997) 1735 – 1744.

R-698 C. B. Burge, and S. Karlin, "Finding the genes in genomic DNA", *Curr. Op. Struct. Biol.* **8** (1998) 346 – 354.

下面列举一批 DNA 序列分析的软件和服务．

R-699 **ReadSeq** 序列格式转换程序．Don Gibert 所写的这个 C 语言程序可从印第安那大学取得：

http://iubio.bio.indiana.edu/readseq/

它的缺点是不能处理包含间隙符号“-”的序列，也不能变换太长的序列。顺便提一下，不能处理太长的序列是多数老程序的弱点。这一方面曾受 FORTRAN 语言数据结构的限制，另一方面也因为人们过去主要关心同单个基因相对应的序列。随着模式生物完全基因组或整个染色体的测序，以及大量基因同时表达的研究，这种情形正在发生变化。 ReadSeq 的说明文件描述了它能够转换的 18 种格式.

R-700 **Artemis** 是 Sanger 中心 [R-299] 最近推出的 DNA 序列显示和注释工具。它用 Java 语言编写，在任何支持 Java 的平台上运行，可以处理任意长度的序列. 这是遵守 GNU [R-62] 协议的免费软件。网址：

http://www.sanger.ac.uk/Software/Artemis/

R-701 **GenScan** 程序基于隐马可夫链模型，目前被认为是最好的基因判认工具之一。果蝇全基因组 [R-369] 的注释也使用了这个程序，效果似比 Genie [R-702] 稍差。算法描述见：

C. Burg, and S. Karlin, *J. Mol. Biol.* **268** (1997) 78 - 94.

网址：

http://gnomic.stanford.edu/GENSCANW.html

R-702 **Genie** 程序基于隐马可夫链模型。最近对果蝇全基因组 [R-369] 的注释，其效果似比 GenScan 程序略佳。部分原因可能在于 Genie 恰好是用果蝇数据训练的。算法描述见：

D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, “A generalized hidden Markov model for the recognition of human genes in DNA”, in *Proceedings of ISMB96*, ed. by D. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. F. Smith, AAAI Press, 1996. 见 [R-826]

网址：

http://www-hgc.lbl.gov/inf/genie.html

R-703 **GeneLang** 是一个基于基因语法规则的模式识别程序。描述见：

S. Dong, and D. B. Searls, “Gene structure prediction by linguistic methods”, *Genomics* **23** (1994) 540 - 551.

网址：

http://cbil.humgen.upenn.edu/~sdong/genlang.html

R-704 **ECOPARSE** 程序使用隐马可夫链模型寻找大肠杆菌 DNA 中的基因。算法描述见：

A. Krogh, I. S. Mian, and D. Haussler, "A hidden Markov model that finds genes in *E. coli* DNA", *Nucleic Acids Res.* **22** (1994) 4768 – 4778.

网址：

`http://genome.cbs.dtu.dk/krogh/EcoParse.info`

R-705 **VEIL** 程序使用一批隐马可夫模型来描述内含子、外显子、基因间区域等各种不同的序列片段，然后借助动态规划的 Viterbi 算法来分析询问序列，以确定编码区。因此，英文名字是 Viterbi Exon-Intron Locator，简称 VEIL。描述见：

J. Henderson, A. Delcher, S. Kasif, and K. Fasman, "Finding genes in human DNA with a hidden Markov model", *J. Comp. Biol.* **4** (1997) 127 – 141.

网址：

`http://www.cs.jhu.edu/labs/compbio/veil.html`

R-706 **GeneParser** 是一个基于动态规划方法的基因识别程序。其算法描述见：

E. E. Snyder, and G. D. Stormo, "Identification of coding regions in genomic DNA", *J. Mol. Biol.* **248** (1995) 1 – 18.

网址：

`http://cbil.humgen.upenn.edu/~sdong/`

R-707 **AAT** 是 Analysis and Annotation Tools 的缩写。这个程序主要靠与数据库中已知的蛋白质和 cDNA 序列对比，来识别编码区和内含子、外显子剪切点。程序描述见：

X. Huang, M. D. Adams, H. Zhou, and A. R. Kerlavage, "A tool for analyzing and annotating genomic sequences", *Genomics* **46** (1997) 37 – 45.

网址：

`http://genome.cs.mtu.edu/aat.html`

R-708 **MORGAN** 是 Multiframe Optimal Rule-based Gene ANalyzer 的缩写。这个程序使用统计学中的决定树方法，以 19 种特性的集合来区分 DNA 的不同片段，并以大量实际序列来训练程序，即确定参数。

算法描述见：

S. Salzberg, "Locating protein coding regions in human DNA using a decision tree algorithm", *J. Comp. Biol.* **2** (1995) 473 - 485.

网址：

`http://www.cs.jhu.edu/labs/compbio/morgan.html`

R-709 **GenView** 程序使用双密码子统计性质来识别内含子和外显子的剪切位点．算法描述见：

L. Milanesi, "GenView: a computing tool for protein-coding regions prediction in nucletide sequences", in [R-3] , 1993 , 573 - 588.

网址：

`http://www.itba.mi.cnr.it/webpage/`

R-710 **ORF Finder** , 是 NCBI 提供的帮助用户寻找开放读框 的网上服务．用户可把核苷酸序列直接提交到网页上，也可按索取号从数据库中指定序列，并规定最小读框的长度．网址：

`http://ncbi.nlm.nih.gov/`

R-711 **GeneFinder** 程序，这是 BCM Search Launcher [R-653] 提供的综合服务的一部分，可以针对不同物种、使用多种方法寻找基因．网址：

`http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html`

R-712 **GeneID** 是一个基于规则的程序，用于识别脊椎动物基因中的编码区．算法描述见：

R. Guigo, S. Knudsen, N. Drake, and T. Smith, "Prediction of gene structure", *J. Mol. Biol.* **226** (1992) 141 - 157.

网址：

`http://www.imim.es/GeneIdentification/`
`Geneid/geneid_input.html`

R-713 **PROCRUSTES** 程序基于剪切位点联配算法，从外显子的各种可能组合中挑选最适应有关蛋白质者．算法描述见：

M. Gelfand, A. Mironov, and P. Pevzner, "Gene recognition via spliced sequence alignment", *Proc. Natl. Acad. Sci.* USA **93** (1996) 9061 - 9066.

网址：

http://www-hto.usc.edu/software/procrustes/index.html

R-714 **SeqPup 程序**。D. Gilbert 用 Java 语言编写的这套生物序列编辑和分析程序，是过去只适用于 Macintosh 平台的 SeqApp 程序 (本书未介绍) 的发展，适用于一切支持 Java 的平台。它包含对许多网络资源和网上分析程序的链接。可以进行多序列联配和编辑、支持多种序列格式、做 DNA 到蛋白质序列的翻译、求 DNA 序列的共轭序列，联配序列的带方框或阴影区的打印等等。所能调用的外部分析程序包括 ClustalW [R-656] 、 CAP [R-692] 和 TACG [R-715] 等。网址：

http://iubio.bio.indiana.edu/seqpup/

R-715 **TACG 程序**，H. Mangalam 编写的对 DNA 序列做内切酶分析的程序，可从 SeqPup [R-714] 等集成程序中调用。网址：

http://hornet.bio.uci.edu/~hjm/projects/tacg/

ftp://iubio.bio.indiana.edu (/restrict-enz/tacg)

R-716 **Glimmer** 是基因定位和内插马可夫模型 (Gene Locator and Interpolated Markov Modeler) 的缩写。这是用内插隐式马可夫模型方法识别编码和非编码序列的程序。其使用可参看：

S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models", *Nucleic Acids Res.* **26** (1998) 544 – 548.

源程序可自由下载：

http://www.cs.jhu.edu/labs/compbio/glimmer.html

R-717 **GeneMark 程序**使用隐马可夫链来识别内含子和外显子，寻找编码区。它对于原核生物比较有效。程序描述见：

M. Borodovsky, and J. McIninch, "GENMARK: parallel gene recognition for both DNA strands", *Comput. Chem.* **17** (1993) 123 - 133.

网址：

http://www2.ebi.ac.uk/genemark/

R 718 **SEView 程序**形象化地表示蛋白质或 DNA 序列已知和预测的各种元件。描述见：

T. Junier, and P. Bucher, "SEView: a Java applet for browsing molecular sequence data", *In Silico Biol.* **1** (1998) 13 – 20. (可从该刊网址

[R-7] 免费下载)

通过 WWW 浏览器使用这一程序的网址是:

ftp://ftp.isrec.isb-sib.ch/sib-isrec/SEView/

源程序在:

ftp://cmpteam4.unil.ch/

R-719 **GRAIL** 是"基因识别分析互联网链接" (Gene Recognition Analysis Internet Link) 的缩写. 这是美国橡树岭国家实验室在能源部人类基因组计划支持下编写的一套程序, 它使用神经网络来发现核酸序列中的编码外显子. 其几个版本除方法改进外, 所用训练序列不同: GRAIL 1 为人、家鼠和大肠杆菌, GRAIL 1a 为人和家鼠, GRAIL 2 为人、家鼠、拟南芥和果蝇. 各版本并存供选择. 用户可以用多种方法享用这套服务. 最简单的办法是用电子邮件提交序列:

mailto: grail@ornl.gov (用 HELP 获取使用说明)

GRAIL 和 GenQuest [R-652] 有一个共同的 WWW 网页界面, 其 URL 是:

http://compbio.ornl.gov/Grail-1.3/

此目录下有 help.html, 通常由 GRAIL 返回的结果, 再提交 GenQuest 与数据库中的全部序列比较. GRAIL 是目前使用得最为普遍的从序列中寻找基因的程序之一. 早在 1996 年, 它每个月平均就要处理 4000 万碱基对.

R-720 **GeneExpress**, 是以俄国学者为主与欧洲合作研制的一个程序系统, 它在 EBI [R-131] 的 SRS [R-203] 界面基础上集成了对真核生物基因组内调控序列的识别、分析和描述. 此系统的介绍见 ISMB98 [R-826] 会议文集中 N. A. Kolchanov 等 25 位作者的文章. 该文可从以下网址下载:

http://wwwmgs.bionet.nsc.ru/mgs/papers/kol/ismb98/

程序系统本身在:

http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/

R-721 **tRNAscan-SE** 程序专门在基因组序列中寻找 tRNA. 描述见:

T. M. Lowe, and S. R. Eddy, *Nucleic Acids Res.* **25** (1997) 955 – 964.

网址:

http://www.genetics.wustl.edu/eddy/tRNAscan-SE/

R-722 **RNABOB** 程序预测 RNA 二级结构，速度较快，但不及 Palingol [R-723] 灵敏。网址：

http://www.genetics.wustl.edu/eddy/software/#rnabob

R-723 **Palingol** 程序预测 RNA 二级结构。网址：

http://www.abi.snv.jussieu.fr/cgi-bin/wrap/viari/Palingol/

R-724 **TRADAT** 是一个集成的预测 DNA 序列中与基因有关的各种特性的程序。网址：

http://www.itba.mi.cnr.it/tradat/

R-725 **NIX** 是另一个集成的从 DNA 序列预测基因的程序。网址：

http://menu.hgmp.mrc.ac.uk/Nix/

R-726 **Pol3Scan** 服务器也可用于寻找 tRNA，因为许多 RNA 基因包含内部 RNA 聚合酶 Pol III 启动子。描述见：

A. Pavesi, *Nucleic Acids Res.* **22** (1994) 1247 - 1256.

网址：

http://irisbioc.bio.unipr.it/pol3scan.html

R-727 **PROMOTER SCAN**，启动子扫描程序，借助查找转录因子结合位点，预测第 II 类 RNA 聚合酶启动位点。描述见：

D. S. Prestridge, *J. Mol. Biol.* **249** (1995) 923 - 932.

网址：

http://biosci.umn.edu/software/proscan/promoterscan.htm

http://bimas.dcrt.nih.gov/molbio/

ftp://biosci.umn.edu (/pub/proscan/)

R-728 **SIGNAL SCAN** 程序，查找所提交的序列中是否包含已发表的信号序列，主要是转录因子。所找到的信号中有许多错误，必须结合序列来源的生物体、细胞和环境加以甄别。描述见：

D. S. Prestridge, *CABIOS* (*Bioinformatics*) **12** (1996) 157 - 160.

网址：

http://bimas.dcrt.nih.gov/molbio/signal/

R-729 **TFSEARCH** 程序查找转录因子结合位点，相当灵敏。网址：

http://www.genome.ad.jp/SIT/TFSSEARCH.html

R-730 **PatSearch** 程序寻找 DNA 序列中的调控元件。描述见 [R-735] 的

引文，网址：

http://transfac.gbf.de/cgi-bin/patSearch/patsearch.pl

R-731 **TESS** 软件是搜寻转录因子结合位点的网页服务。描述见：

J. Schug, and G. C. Overton, "TESS: Transcription element Search Software on the WWW", *Tech. Rep.* CBIL-TR-1997-1001-v0.0 Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania.

网址：

http://agave.humgen.upenn.edu/utess/tess31/

R-732 **MatInspector** 程序，利用 TRANSFAC [R-219] 数据库所提供的调控元件的代表序列和权重矩阵，在 DNA 中寻找这些结合位点。描述见：

K. Quandt, *Nucleic Acids Res.* **23** (1995) 4878 – 4884.

网址：

http://www.gsf.de/cgi-bin/mastersearch.pl

R-733 **FunSiteP** 程序预测 DNA 序列中的启动子位置。描述见：

Y. V. Kondrakhin 等 5 位作者， *CABIOS* (*Bioinformatics*) **11** (1995) 477 – 488.

网址：

http://transfac.gbf.de/dbsearch/funsitep/fsp.html

R-734 **PromDF** 启动子搜寻程序：

ftp://beagle.colorado.edu (/pub/PromFD.tar)

R-735 **FastM** 服务器，在一定距离内寻找同时存在的两个调控元件，因此可以显著降低转录因子结合位点的误报率。描述见：

T. Heinemeyer 等 12 位作者， *Nucleic Acids Res.* **26** (1998) 362 – 367.

网址：

http://transfac.gsf.de/cgi-bin/fastm.pl

R-736 **SaGa** 服务器使用遗传算法分析 DNA 结构 (Structural Analysis with Genetic Algorithm)，它从一组已经联配好的序列中发现共同的结构上的特征。用户须在网页上提交序列。网址：

http://transfac.gbf.de/TRANSFAC/cgi-bin/saga/saga.pl

R 737 **MFOLD** 是预测 RNA 二级结构的程序。它是 GCG [R-792] 程序

包的一部分，因而不能免费获得。但网上有一个 PC 机的版本，可以一试。网址：

ftp://ftp.fly.bio.indiana.edu

取文件 /molbio/ibmpc/pcfldsrc.uue (UUENCODED)。

R-738 维也纳大学有一个 UNIX 系统的预测 RNA 二级结构的程序：

ftp://itc.univie.ac.at

R-739 **HMMER** 是用隐马可夫链反映蛋白质结构域轮廓 (profile) 的程序包，可用以在新的蛋白质序列中发现这些结构域。Pfam [R-478] 数据库就包含一批用此程序产生的 HMM 模型。网址：

http://hmmer.wustl.edu/

R-740 **HMMPRO** 是 [R-18] 一书作者之一 P. Balsi 编写的用隐马可夫链模型做序列分析的程序，其最新的 2.2 版由 NET-ID 公司销售，从事学术研究的个人，可在注册取得允许后免费下载。网址：

http://www.netid.com/

R-741 **MZEF** 是冷泉港实验室 [R-159] 张奇伟编写的预测 DNA 序列中编码外显子的程序。它的算法已经发表：

M. Zhang, *Proc. Natl. Acad. Sci.* USA **94** (1997) 565 - 568.

非营利性用户可从以下 ftp 服务器下载：

http://sciclio.cshl.org/genefinder/

ftp://ftp.cshl.org (/pub/science/mzef/)

R-742 **CorePromoter** 是冷泉港实验室 [R-159] 张奇伟编写的预测人类基因中核心启动子的程序，其算法已经发表：

M. Zheng, "Identification of human gene core-promoters in silico", *Genome Res.* **8** (1998) 319 - 326.

非营利性用户可从以下 ftp 服务器下载：

ftp://ftp.cshl.org (/pub/science/promoter/)

R-743 **ESTScan** 服务，用户可提交 DNA 序列，以查找其中的 EST 片段。此系统对人和哺乳类动物优化，不可用于其他物种。网址：

http://www.ch.embnet.org/software/ESTScan.html

R-744 美国国家生物信息中心 NCBI 提供“电子 PCR”服务。用户可以提交核酸序列，查找其中包含的已知的 EST 片段。请由 NCBI [R-134] 的网址进入：

http://ncbi.nlm.nih.gov/

R-745 **TRF** 是寻找串联重复序列 (Tandem Repeats Finder) 的程序. 描述见:

G. Benson, *J. Comp. Biol.* **4** (1997) 351 – 367.

TRF 服务器的网址:

http://c3.biomath.mssm.edu/trf.upload.form.html

R-746 **Satellites** 程序, 寻找基因组中的卫星重复序列. 描述见:

M. F. Sagot, and E. W. Myers, *J. Comp. Biol.* **5** (1998) 539 – 554.

网址:

http://bioweb.pasteur.fr/seqanal/
interfaces/satellites.html

R-747 **HLA-Bind**, 白细胞抗原肽链结合位点的预测程序. 网址:

http://bimas.dcrt.nih.gov/molbio/hla_bind/

R-748 **RepeatMasker** 是华盛顿大学 A. F. A. Smit 和 P. Green 发展的一套程序, 它检查用户提交的序列中所包含的已知的重复序列和简单 (低复杂度) 序列, 并把相应字母 "掩去", 即换成 N(或 X). 这个程序可以下载到本地计算机上运行, 也可通过 WWW 网页或电子邮件提交序列. 它的运行要求调用 RepBase [R-223] 数据库. 网址:

http://ftp.genome.washington.edu/RM/RepeatMasker.html

mailto: repeatmasker@ftp.genome.washington.edu

可在电子邮件主体中写 HELP 以获取使用说明.

R-749 **Dotter** 程序, 是用点阵法进行两个序列对比的程序, 也可用于在同一个序列中查找重复或逆重复片段. 其优点是形象化, 缺点是只适用于不太长的序列. 它可以开窗口来显示指定段落的联配情况. 程序描述见:

E. L. L. Sonnhammer, and R. Durbin, *Gene* **167** (1995) GC1 – GC10.

可免费从 KISAC [R-145] 的网址下载, 亦见:

ftp://ncbi.nlm.nih.gov (/pub/esr/dotter/)

R-750 **RHMAPPER** 程序, 是由 WICGR [R-158] 发展的免费辐射杂交图谱软件. 它基于最大似然模型, 可从以下网址获取:

http://www.genome.wi.mit.edu/ftp/pub/software/rhmapper/

R-751 **BEND** 和 **BEND-TRI** 都是预测 DNA 链弯曲性和曲率的程序.

描述见：

D. S. Goodsell, and R. E. Dickerson, *Nucleic Acids Res.* **22** (1994) 5497 - 5503.

网址：

`http://www.scripps.edu/pub/goodsell/research/bend/`

R-752 **CURVATURE** 是预测 DNA 链曲率的程序。描述见：

E. S. Shipigelman, E. N. Trifonov, and A. Bolshoy, *CABIOS* (*Bioinformatics*) **9** (1993) 435 - 440.

网址：

`ftp://sgjsl.weizmann.ac.il (/pub/Curvature/)`

§5.8 蛋白质结构和功能预测

DNA 序列的测序速度，远远超过测定蛋白质三维结构的进展。因此，如何从蛋白质序列甚至由 DNA 读框翻译出的氨基酸序列预测可能的蛋白质结构，就成为迫切的任务。目前较为有效的预测方法，都要依靠已知三维结构的蛋白质的序列来预测折叠单元。基于轮廓 (profile) 的预测方法，通常假定两段被比较的蛋白质序列的氨基酸残基接近水环境的情形是保守的。另一种所谓“线串法” (threading) 则把残基环境中的疏水作用相加，因而效果稍好。但对于多结构域的蛋白质序列，线串法的效果不佳。当序列中残基数目超过 800 时，目前很难作出有意义的预测。

R-753 **Threader** 程序，使用线串法预测蛋白质序列的三级结构。描述见：

D. J. Jones, W. R. Taylor, and J. M. Thronton, *Nature* **358** (1992) 86 - 89; *Proteins* **23** (1995) 337 - 355.

此软件的 2.5 版学术性单位仍可在注册后自由下载。网址：

`http://globin.warwick.ac.uk/~jones/threader.html`

R-754 **LIBRA I** 是 LIght Balance for Remote Analogous proteins 的缩写。这是一套分析蛋白质结构和序列的程序，其主要手段是“线串法” (threading)。网址：

`http://www.ddbj.nig.ac.jp/E-mail/libra/LIBRA_I.html`

现在这套程序已经扩充到可以分析蛋白质反折叠问题，参见：

M. Ota, and K. Nishikawa, “Feasibility in the inverse protein folding

protocol", *Protein Sci.* **8** (1999) 1001 - 1009.

R-755 **PEDANT** 程序系统，允许用户综合使用多种方法从蛋白质序列预测其结构和功能，网址：

`http://pedant.mips.biochem.mpg.de/`

R-756 **SAPS**，这是一个蛋白质序列统计分析 (Statistical Analysis of Protein Sequences) 程序，其算法描述见：

V. Brendel 等 5 位作者，"Methods and algorithms for statistical analysis of protein sequences", *Proc. Natl. Acad. Sci.* USA **89** (1992) 2002 - 2006.

用户可从以下网页提交序列，进行分析：

`http://www.isrec.isb-sib.ch/software/SAPS_form.html`

R-757 **GeneFIND** 不是一个寻找基因的程序，而是帮助确定蛋白质家族的工具，即 Gene Family Identification Network Design 的缩写。它是基于 ProClass 数据库 [R-411] 的一组集成的搜寻和联配程序。它首先从快速神经网络程序 MotiFind 出发，然后使用 BLAST [R-631] 搜寻和 Smith-Waterman [R-623] 算法做联配，即调用 SSearch [R-624] 程序，进行模体搜寻。最终结果用 HTML 格式给出，包括整体和模体的打分数、与 PROSITE 数据库 [R-406] 和 PIR 数据库 [R-404] 中超家族的匹配清单、模体匹配，以及到 ProClass 数据库条目的链接等。网址：

`http://diana.uthct.edu/genefind.html`

R-758 **COILS** 服务器，根据蛋白质序列预测由两条 α 螺旋形成卷曲螺旋 (coiled coil) 的区域，算法描述见：

A. Lupas, M. Van Duke, and J. Stock, "Predicting coiled coils from protein sequences", *Science* **252** 1162 - 1164.

用户可从以下网页提交序列，进行预测：

`http://www.ch.embnet.org/software/COILS_form.html`

R-759 **CASP**(Critical Assessment of methods of protein Structure Prediction) 是在国际互联网上组织的蛋白质结构预测方法评估的竞赛，已经举行过三届。请参看网址：

`http://moult.carb.nist.gov/`

关于 2000 年正在进行的 CASP4，可访问网址：

http://predictioncenter.llnl.gov/

R-760 **PHD** 是一套集成的根据蛋白质序列搜索数据库和用多种方法预测蛋白质性质的程序。目前它包含一组程序：预测二级结构的 PHDsec、预测溶剂可及性的 PHDacc、预测跨膜螺旋的 PHThtm、预测拓扑的 PHDtopology 和线串法预测折叠的 PHDthreader 等。它返回的结果包含数据库中多相似序列的联配、PROSITE [R-406] 模体、ProDom [R-480] 结构域、卷曲螺旋区、球形区、跨膜区等。用户可以从网页或用电子邮件提交序列。从 1999 年初，原在海德堡欧洲分子生物学实验室的网址：

http://www.embl-heidelberg.de/predictprotein/

mailto: predictprotein@embl-heidelberg.de

已经转到美国纽约哥伦比亚大学的网址：

http://cubic.bioc.columbia.edu/predictprotein/

mailto: phd@dodo.cpmc.columbia.edu

PHD 系统有详细的使用说明，可用电子邮件 HELP 获取，或在网页上阅读：

http://cubic.bioc.columbia.edu/
predictprotein/help_entry.html

也可下载有关文件：

ftp://cubic.bioc.columbia.edu

在子目录 /pub/phd/ 中取 wwwPP.tar.gz 和 ReadMe 两个文件。

PHD 蛋白质预测程序的中国镜象点在北京大学生物信息中心，可通过 [R-166] 网页的 predictprotein 选项进入，或直接访问网址：

http://www.cbi.pku.edu.cn/predictprotein/

R-761 英国癌症研究基金会 (Imperial Cancer Research Fund，简称 ICRF) 的分子生物模型实验室有一个识别蛋白质结构域的服务器，通常称为 ICRF 服务器。网址：

http://www.bmm.icnet.uk/~domains/

R-762 **PREDATOR** 程序，从单个或多个蛋白质序列预测二级结构，由于计入了长程相互作用，精度较高。描述见：

D. Frishman, and P. Argos, *Prot. Eng.* **9** (1996) 133 – 142; *Proteins* **27** (1997) 329 – 335.

可从 PEDANT [R-755] 中调用，也可访问网址：

http://www.embl-heidelberg.de/argos/predator/

R-763 **Prof** 程序，预测蛋白质二级结构。这是原来 DSC 程序的第二代。网址：

http://www.bmm.ncnet.uk/~prof/

R-764 **NNSSP** 程序，由最近邻关系预测蛋白质二级结构。网址：

http://dot.imgen.bcm.tmc.edu:9331/pssprediction.pssp.html

R-765 **Jpred2**，这是目前公认为较好的由蛋白质序列预测二级结构的服务。它综合使用若干种方法，包括先搜寻 PDB [R-441] 库里有没有相似的序列。PDB 中的相似序列往往更有意义，不必重新预测二级结构。Jpred2 的算法还未发表，因此不能下载到本地计算机上运行，只能把序列提交到以下网页：

http://jura.ebi.ac.uk:8888/

R-766 **nnpredict** 程序，用双层神经网络预测蛋白质的二级结构。算法描述见：

D. G. Kneller, F. E. Cohen, and R. Langridge, "Improvements in protein secondary structure prediction by an enhanced neural network", *J. Mol. Biol.* **214** (1990) 171 – 182.

用户可从网页或用电子邮件提交序列，每次只能提交一个序列。可以使用氨基酸的单字母名字 (但不许用 B 和 Z)，也可以用三字母名字，但须以空格分开。网址：

http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

mailto: nnpredict@celeste.ucsf.edu

R-767 **SignalP** 程序，用神经网络预测氨基酸序列中的信号肽。算法描述见：

H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, *Prot. Eng.* **10** (1997) 1 – 6.

可从网页或用电子邮件提交序列：

http://www.cbs.dtu.dk/services/SignalP/

mailto: signalp@genome.cbs.dtu.dk

R-768 **Structer**，一个由蛋白质三维结构产生接触图的程序，它实际上给出一个矩阵，反映蛋白质序列中哪些氨基酸残基在三维结构中互相

接触.

ftp://ncbi.nlm.nih.gov (/pub/esr/structer/)

下面是一批预测跨膜蛋白的程序或服务。我们仅把网址记录在案，以供参考。同时，请注意 [R-760] 中的 PHThtm 程序.

R-769 **TMpred** 是预测蛋白质中跨膜区及其取向的程序，它调用 TMbase [R-471]。网址:

http://www.ch.embnet.org/software/TMPRED_form.html

http://ulrec3.unil.ch/software/TMPRED_form.html

R-770 http://www.biokemi.su.se/~server/ TopPred2DAS

R-771 http://www.cbs.dtu.dk/services/TMHMM-1.0/

R-772 http://globin.bio.warwick.ac.uk/psipred/

R-773 使用表面紧致排列 (Dense Alignment Surface，简称 DAS) 方法，预测跨膜蛋白质的程序。网址:

http://www.biokemi.su.se/~server/DAS/

R-774 **TMAP**，欧洲分子生物学实验室预测蛋白质序列中跨膜螺旋的网上服务:

http://www.embl-heidelberg.de/tmap/tmap_info.html

原来的电子邮件服务 tmap@embl-heidelberg.de 已经停止。

使用二维凝胶电泳 (见 3.6.4 小节和 2DPAGE 数据库 [R-419])，可在一次实验中分辨上千种蛋白质。互联网上有一些服务，协助用户比较电泳图斑。

R-775 **Flicker** 服务网页，可以比较来自任何两个 URL 的 2D-PAGE 图象。描述见:

P. F. Lemkin, "Comparing 2D electrophoretic gel images across the Internet", *Electrophoresis* **18** (1997) 461 – 470.

Flicker 服务可在 Netscape 4.0 和 Internet Explorer 4.0 以上的浏览器中运行。网址:

www-lecb.ncifcrf.gov/flicker/

R-776 **Melanie 3** 服务，是与 SWISS-2DPAGE [R-419] 数据库配套的观察比较蛋白质二维凝胶图象的服务，可从 ExPASy [R-142] 的网页进入。例如，可以访问北京大学生物信息中心的镜象:

http://expasy.pku.edu.cn/melanie/

§5.9 显示蛋白质和核酸结构的程序

显示大、小分子结构的程序早就是结构生物学软件包的重要模块。现在有大量美妙的、可以动态显示甚至做动画模拟的程序。这里只提几种可从网上获取的、使用较为广泛的免费软件。

R-777 **RasMol** 程序，由 R. A. Sayle 编写。这是使用得最为广泛的一个显示 DNA 和蛋白质等分子三维结构的免费程序，可以用骨架图、条带图、空间填充图等各种方式显示，并可在显示时随意转动分子。其描述见：

R. A. Sayle and E. J. Milner-White. "RasMol: biomolecular graphics for all", *Trends in Biochem. Sci.* **20** (1995) 374 – 376.

它有适用于各种平台的版本，可从 E. Martz 维护的网页下载：

http://www.umass.edu/microbio/rasmol/

那里还可以获取 RasMol 使用说明书，示例分子和一些其他文件。RasMol 也可从 NCBI [R-134] 的 Cn3D [R-779] 显示程序的网页下载。

R-778 **Chime**，功能与 RasMol [R-777] 类似，也是显示分子三维结构的免费程序。但与 RasMol 不同，Chime 不能独立运行，而要在网页浏览器内显示。学术性用户可以从 MDL Information Systems 公司的网址下载软件和获得信息：

http://www.mdli.com/

R-779 **Cn3D** 是与 MMDB [R-463] 配套的一个三维分子结构和 NMR 模型的显示程序，可在 NCBI [R-134] 的网址直接使用或下载到用户的计算机上执行。它有适用于 PC 视窗系统和多种 UNIX 工作站的版本。描述见：

C. W. V. Hogue, *Trends in Biochem. Sci.* **22** (1997) 314 – 316.

R-780 **Protein Explorer**，蛋白质结构显示程序，简称 PE。可在网页上使用，也可在 PC 视窗系统中运行，目前还没有 UNIX 版本。网址：

http://www.umass.edu/microbio/chime/explorer/

R-781 **MolScript** 是 Per Kraulis 编写的从三维坐标显示大致或详尽的分子图象的程序。它要求特殊格式的输入文件。它的新版中有一个 MolAuto 程序协助产生输入文件。MolScript 程序由作者拥有的 Avatar Software AB 公司销售，不能自由获取和转让，但学术性机构为了从事纯学术研究，可以免费获得使用许可证。1999 年 1 月发行 2.1.2 版，请参看 MolScript 的正式网址：

`http://www.avatar.se/molscript/`

关于 MolScript 的使用说明和其他一些图形程序的信息，还可访问以下网址：

`http://graph.sci.osaka-cu.ac.jp/~teddy/`

§5.10 大规模基因表达的算法

由于 DNA 芯片和微阵列的迅速发展，对某一物种或组织中全部基因的表达关系进行整体性研究已经提上日程。例如，对酵母全部 6 300 个基因在一个生活周期中进行多次采样，然后研究它们之间的表达关系。首先是按照同步或反同步表达，以及表达强度的变化，把几千个基因分成聚类 (clusters)。然后由之构建基因调控网络、代谢网络，提取调控过程和生化反应的各种参数。这正在成为极其活跃的研究领域。目前在这一领域还没有成熟的服务网页。关于算法问题的讨论，可以从每年的太平洋生物计算研讨会文集 [R-825] 查到线索。

关于 DNA 芯片和微阵列，以及使用此类新技术对大规模基因表达的研究，可以访问一些实验室的网页，例如：

R-782 **MGuide**，斯坦福大学 Pat Brown 实验室的“微阵列导引”，有详细的技术介绍，其目的就是协助读者建立设备、进行研究。网址：

`http://cmgm.stanford.edu/pbrown/`

R-783 V. Cheung 所领导的基因组学研究室把 DNA 微阵列技术和 GMS (Genome Mismatch Scanning) 结合起来，直接用 IBD(Identical-by-Descent) 作图谱。用这种技术可以从共享 IBD 的个体分离 DNA 片段和作图谱。它们维护着人类 BAC 图谱数据库 GenMapDB [R-316]。网址：

http://w95vcl.neuro.chop.edu/vcheung/

R-784 英国曼彻斯特大学生物信息组的微阵列小组，专门研究有关算法和软件。请访问他们的网址：

http://www.bioinf.man.ac.uk/microarry/

R-785 关于细菌功能基因组学，特别是大肠杆菌基因的大规模表达研究，可参看网页：

http://bomi.ou.edu/faculty/tconway/global.html

他们的论文亦可参考：

T. Tao 等， *J. Bacteriol.* **181** (1999) 6425 – 6440.

这一新领域的文献尚不多。洛克菲勒大学基因统计学实验室李问天的网页上有一个随时更新的文献目录：

R-786 李问天微阵列文献目录，网址：

http://linkage.rockefeller.edu/wli/microarray/

§5.11 细胞过程模拟

细胞是生命活动的基本单元。随着对亚细胞结构的了解和对细胞内各种代谢途径与信号转导过程的知识和数据的积累，现在已经可以尝试模拟活细胞的生活过程。这是比核酸和蛋白质序列更高层次的生物信息学研究课题，1999 年已经有良好开端。1999 年 4 月 2 日的美国《科学》周刊介绍了两套程序：

R-787 **E-Cell** 。日本庆应义塾 (Keio) 大学生物信息学教授富田 (Masaru Tomita) 所编写的 E-Cell 程序，在 Red Hat Linux [R 48] 环境下运行，目前已进入 β 试运行阶段。它允许用户规定细胞中有哪些基因和其他分子，以及它们的位置和浓度，由程序去模拟生化过程的发展。运行中可剔除某个基因，或改变营养状况，以观后效。目前这个程序还只是演示基本上已知的生化反应途径。从长远看，它可能用 *in silico* 实验代替某些烦琐的 *in vitro* 研究，导致新的发现。1999 年 12 月 2 日的英国《自然》周刊在讨论 2010 年之前大规模科学计算的前景时，也提到 E-Cell 。此程序目前源代码公开，并有较详细的使用说明书，可以下载到本地计算机上运行。网址：

http://www.e-cell.org/

英文说明书见：

http://www.e-cell.org/manual/indexE.html

R-788 **Virtual Cell**，美国康涅狄格大学生物医学成象技术中心 (Center for Biomedical Imaging Technology，简称 CBIT) 所属国家细胞分析和模拟资源处 (National Resource for Cell Analysis and Modeling，简称 NRCAM) 的生理学家 L. Loew 和计算科学家 J. Schaff 联手，用 Java 语言编写的这个“虚拟细胞”程序，提供一个检验各种模型的环境，以便把三维细胞中描述单个反应过程的生化反应和电生理数据同实际的显微镜成象结合，模拟亚细胞层次的细胞活动。较详细的描述见：

J. Schaff, and L. M . Loew, "The virtual cell", *Pacific Symposium on Biocomputing* **4** (1999) 228 – 239. (参看 [R-825])

目前用户只能进入该大学的网页运行：

http://www.nrcam.uchc.edu/

R-789 从细胞内生物化学过程的整体性研究，自然地走向整个组织、器官生理过程的研究。已经有人仿照 genome(基因组) 和 proteome(蛋白质组)，制造了 Physiome 一字，意在从整体上研究生理过程。请参看网址：

http://www.physiome.org/

§5.12 向数据库提交序列的软件和服务

由实验确定的新核酸序列，只要提交到 GenBank [R-212]、EMBL [R-211] 或 DDBJ [R-213] 三家之一即可。有一批软件和网页服务协助用户做这件事。

向 NCBI[R-134] 所管理的 GenBank 提交序列，可使用 BankIt 或 Sequin 两种办法。

R-790 **Sequin** 是在本地计算机上运行的程序，它可以协助用户对序列做注释和进行一些分析，最终形成的文件可用电子邮件送往：

mailto: gb-sb@ncbi.nlm.nih.gov

也可用软盘寄往 NCBI。

R-791 **BankIt** 是 NCBI 网页的一个选项：

`http://www.ncbi.nlm.nih.gov/BankIt/index.html`

进入之后根据文字说明操作，提交新序列或更新旧序列(只有原始序列提交者才有权更新)。

§5.13 商业性生物信息资源

5.13.1 商业性软件

国际市场上有大量通用或专用的生物计算软件。它们大多价格昂贵，但通常有较好的服务。事实上，价格的相当部分在于购买服务。下面简单介绍的几种软件，或是因为文献中经常提到，或是由于我们自己偶有接触。我们的简介完全没有对这些软件作评价或推荐之意。

R-792 **GCG** 程序包，最初是 Wisconsin 大学的 Genetic Computer Group 发展的，后者现在是位于美国加州的 Oxford Molecular Group 的子公司。GCG 包含 130 多个与分析有关的程序，有 6 种重要生物数据库随程序包一起提供，并由 GCG 公司负责每两个月更新一次。我国有些单位或自行引进、或在外国公司支持下获得了这套程序，但严格限于单位内部使用。GCG 有详细的使用说明书，这里只作简要介绍。

第一、GCG 程序包的内容，我们只分类点名：

1. 序列比较程序：双序列联配有 Gap、BestFit、Compare、DotPlot 等，多序列联配有 PileUp。

2. 数据库搜索和分析程序：LookUp、StringSearch、BLAST、FASTA、NetBLAST 等，其中 BLAST 和 FASTA 都搜索本地数据库，而 NetBLAST 经过互联网去搜索 NCBI [R-134] 的数据库。

3. 演化和亲缘关系程序：Distances、Growtree、Diverge 等。

4. 序列片段组装程序：Gelstart、Gelmerge、Gelassemble、Geldisassemble 等。

5. 寻找基因和识别模体的程序：Frames、Motifs、Repeat、Findpattern、Xnu、Seg 等。

6. 蛋白质分析程序: Profilescan 、 Peptidesort 、 Moment 、 Helicalwheel 、 Isoelectric 、 Pepplot 等。

7. RNA 二级结构预测程序: Mfold 、 Plotfold 、 Foldrna 、 Stemloop 、 Circles 等。

8. 还有用于引物设计、格式转换、打印输出、图谱处理、数据库和序列查询等方面的程序，此处从略。

第二, GCG 有三种基本的运行方式:

1. UNIX 工作站上的命令行方式。只要在每次运行开始时，自动设置好所有的 GCG 环境变量，所有 GCG 程序调用起来就像是普通 UNIX 命令，但要熟悉命令行参数的意义和写法。用户可以从 PC 机远程登录到 UNIX 工作站上去运行 GCG 。

2. 通过 SeqLab 图形用户界面，打开若干使用 GCG 程序的窗口。 PC 机用户需有 X 窗口的模拟程序。

3. 通过网络在 Netscape 或 Internet Explorer 浏览器中使用 SeqWeb 接口系统，调用 GCG 中的一批核心程序。

R-793 **Peptool**™ 程序和 **Genetool**™ 程序是 Biotools 公司提供的可在任何计算机平台上运行的集成序列分析软件。公司网址:

`http://www.biotools.com`

其中 Peptool 的基础是加拿大 Alberta 大学发展的适用于 UNIX 系统的免费软件，描述见:

D. S. Wishart, "SEQSEE: a comprehensive program suite for protein sequence analysis", *CABIOS* (*Bioinformatics*) **10** (1994) 121 – 132.

D. S. Wishart, "Contraint multiple alignment using XALIGN". *CABIOS* (*Bioinformatics*) **10** (1994) 687 – 688.

D. S. Wishart, "A platform independent graphical user interface for SEQSEE and XALIGN", *CABIOS* (*Bioinformatics*) **13** (1997) 561 – 562.

R-794 InforMax 公司的 **VectorNTI** 程序，可以说是与 GCG 程序包恰成对比的一套小巧玲珑的软件，可以从常用载体的内切酶位点设计做到基本的 DNA 序列分析。全部程序在一张光盘上。公司网址:

`http://www.informaxinc.com/`

R 795 **DNATools** 是一套处理 DNA 和蛋白质序列的程序，适用于 PC 视

窗系统。它虽由一个大学实验室研制，但要收取许可证费用。因此，我们把它列入商业软件。对于学术性用户，它价格较低，而且有可以延长的四个月试用期．网址：

http://www.crc.dk/dtmain/

R-796 **GenTerpret** 是 Rabbithutch 生物技术公司所发展的核酸序列自动解释程序。它用 SorFind 2.0 预测编码的外显子，用 RepFind 2.0 查找重复片段，用 PromFind 2.0 预测启动子和 CpG 岛，并有图形接口．详见：

http://www.rabbithutch.com/

SorFind 、 RepFind 和 PromFind 三个程序的早期版本，保存在印第安那大学的档案中：

ftp://iubio.bio.indiana.edu /molbio/ibmpc

5.13.2 一些公司网页

生物技术、基因工程和生物信息公司多如雨后春笋。少数公司维护着免费的公用数据库和其他信息资源，多数公司网页以广告营销为主，但从其一般介绍特别是与公司研究成果有关的出版物目录，有时可以获得一些有益的信息。下面列举的一些网址，多系我们偶然遇见、顺手记下。是否入选，与公司业绩无关，也绝无评价或推荐之意．已经在前一小节介绍商业软件时提到过的公司，也不再重复．

R-797 **NEB** 是新英格兰生物实验公司 (New England BioLab) 的缩写。它维护着限制性内切酶和甲基化酶数据库 REBASE [R-424] 及蛋白质剪接数据库 InBase [R-436] 。据说 30% 已知的限制性内切酶是在 NEB 发现的． NEB 也是限制性内切酶和许多其他生化产物的重要供应商之一．网址：

http://www.neb.com/

R-798 **Celera** 公司，由原来在 TIGR [R-156] 的 J. C. Venter 参与组织、用霰弹法进行 DNA 测序，已经基本测完果蝇的基因组，正在人类基因组方面与国际人类基因组计划竞争．网址：

http://www.celera.com/

R-799 **D'Trends** 是一家生物信息公司．网址：

http://www.d-trends.com/

R-800 **Net-ID, Inc.** 是一家生物信息软件公司，它目前的主要产品 HMMpro2.2 [R-740] 对于学术界仍是免费的。网址：

http://www.netid.com/

R-801 **Affymetrix** 是世界上最大的 DNA 芯片公司之一，其科学出版物目录值得参考。网址：

http://www.affymetrix.com/technology/papers.html

R-802 **NanoGen Inc.** 是一家微电子技术与分子生物学结合的芯片公司。网页上一些与公司研究成果有关的出版物目录和专利简介值得参考。网址：

http://www.nanogen.com/

R-803 **Hyseq Inc.** 是使用杂交法的一家 DNA 芯片公司。它的长约 20 个碱基的寡核苷酸片段是垂直立在芯片上的。据称它拥有 1 200 万份 DNA 样品的分析结果，90 万个部分或完整的基因序列，已经找到 35 000 个基因。网址：

http://www.sbh.com/

为了出售基因，它最近又建立了一个名为 GeneSolutions 的子公司，请参看：

http://Gene.Solutions.com/

R-804 **InCyte** 公司最近改名为 InCyteGenomics，这家药物公司拥有的 LifeSeq 数据库中有大量人类基因和 EST 序列。它提供某些免费信息服务 (LifeSeq Public) 以促进销路。网址：

http://www.incyte.com/

它经营的一种集成生物信息软件叫做 Life Tools，并有一个生物信息网页：

http://www.incyte.com/Globe/bioinfo.html

R-805 **MSI**(Molecular Simulation, Inc.) 公司发展了大量结构生物学、大分子模拟、蛋白质工程、药物设计等方面的商业软件，包括使用甚广的三维分子模拟 图形环境 Insight II 软件。请参看网页：

http://www.msi.com/life/index.html

http://www.msi.com/life/products/insight/modules/Insight2.html

R-806 **Ambion, Inc.** 是一家专门提供 RNA 实验分析工具的公司，它发行不定期的电子通讯 RNA FlashNotes，可以自由订阅．网址：
`http://www.ambion.com/`

§5.14 其他网上生物医学信息资源

除了公开数据库和免费软件，国际互联网上还有大量其他生物医学信息资源和讨论组，以及电子出版物、会议消息和文集、讲义和课程等等。下面分类列举一些网址。

5.14.1 网上论坛：BIOSCI 新闻组

BIOSCI/bionet 是组织得很好的网上论坛，它又分成一百多个专题新闻组，见表 5.13 。

BIOSCI 是专业生物工作者共享的一个生物学论坛，不是业余爱好者的谈天空间．可以用三种方式之一参加：

R-807 最方便的办法是进入 BIOSCI/bionet 论坛的网页：
`http://www.bio.net/`
然后选取 Access the BIOSCI/bionet News Groups .

R-808 在本地计算机上安装 USENET 软件，或利用浏览器已经配备的访问新闻组功能．它可以替用户组织管理同指定的新闻组来往的信息。

R-809 用电子邮件订阅和参与一个或多个新闻组．这个办法最简单，但有不少缺点：必须指名订阅特定的新闻组；如果同时订阅多个新闻组，来信同其他电子邮件一起随机地进入电子邮件箱，必须自行组织整理；如果本地计算机的电子邮件服务出现故障，来信被连续退回，新闻组就会自动停送，必须重新订阅，才能恢复．初次订阅前，宜先发电子邮件到：
`mailto: biosci-server@net.bio.net`
在邮件主体中写一个字 lists，以获得当前的 BIOSCI 专题名单。选定专题 listname 之后，在下一封邮件主体中写 subscribe listname；停止订阅时写 unsubscribe listname．注意，所有命令都要写在邮件主体中，而不要放在 Subject 后面。

表 5.13 BIOSCI/usenet 新闻组

简称	新闻组地址
ACEDB-SOFT	bionet.software.acedb
AFCR	bionet.prof-society.afcr
AGEING	bionet.molbio.ageing
AGROFORESTRY	bionet.agroforestry
AIBS	bionet.prof-society.aibs
AMYLOID	bionet.neuroscience.amyloid
ARABIDOPSIS	bionet.genome.arabidopsis
ASCB	bionet.prof-society.ascb
AUDIOLOGY	bionet.audiology
AUTOMATED-SEQUENCING	bionet.genome.autosequencing
BIOCAN	bionet.prof-society.cfbs
BIOFILMS	bionet.microbiology.biofilms
BIOFORUM	bionet.general
BIO-INFORMATION-THEORY	bionet.info-theory
BIONAUTS	bionet.users.addresses
BIONEWS	bionet.announce
BIO-JOURNALS	bionet.journals.contents
BIO-MATRIX	bionet.molbio.bio-matrix
BIOPHYSICAL-SOCIETY	bionet.prof-society.biophysics
BIOPHYSICS	bionet.biophysics
BIO-SOFTWARE	bionet.software
BIO-SRS	bionet.software.srs
BIOTECHNIQUES	bionet.journals.letters.biotechniques
BIOTHERMOKINETICS	bionet.metabolic-reg
BIO-WWW	bionet.software.www
CARDIOVASCULAR-RESEARCH	bionet.biology.cardiovascular
CELEGANS	bionet.celegans
CELL-BIOLOGY	bionet.cellbiol
CHLAMYDOMONAS	bionet.chlamydomonas
CHROMOSOMES	bionet.genome.chromosomes
COMPUTATIONAL-BIOLOGY	bionet.biology.computational
CSM	bionet.prof-society.csm
CYTONET	bionet.cellbiol.cytonet
DEEPSEA	bionet.biology.deepsea
DIAGNOSTICS	bionet.diagnostics
DROSOPHILA	bionet.drosophila
ECOPHYSIOLOGY	bionet.ecology.physiology

表 5.13 (续表)

简称	新闻组地址
EMBL-DATABANK	bionet.molbio.embldatabank
EMF-BIO	bionet.emf-bio
FASEB	bionet.prof-society.faseb
FLUORESCENT-PROTEINS	bionet.molbio.proteins.fluorescent
FREE-RADICALS	bionet.molecules.free-radicals
G-PROTEIN-COUPLED-RECEPTOR	bionet.molbio.proteins.7tms_r
GDB	bionet.molbio.gdb
GENBANK-BB	bionet.molbio.genbank
GENETIC-LINKAGE	bionet.molbio.gene-linkage
GENSTRUCTURE	bionet.genome.gene-structure
GLYCOSCI	bionet.glycosci
GRASSES-SCIENCE	bionet.biology.grasses
HIV-MOLECULAR-BIOLOGY	bionet.molbio.hiv
HUMAN-GENOME-PROJECT	bionet.molbio.genome-program
IMMUNOLOGY	bionet.immunology
INFO-GCG	bionet.software.gcg
INSULIN-ACTION	bionet.cellbiol.insulin
JOURNAL-NOTES	bionet.journals.note
METHODS-REAGENTS	bionet.molbio.methds-reagnts
MICROBIOLOGY	bionet.microbiology
MOLECULAR-EVOLUTION	bionet.molbio.evolution
MOLECULAR-MODELLING	bionet.molec-model
MOLECULAR-REPERTOIRES	bionet.molecules.repertoires
MOLLUSC-MOLECULAR-NEWS	bionet.molbio.molluscs
MYCOLOGY	bionet.mycology
NAVBO	bionet.prof-society.navbo
NEUROSCIENCE	bionet.neuroscience
N2-FIXATION	bionet.biology.n2-fixation
P450	bionet.molecules.p450
PARASITOLOGY	bionet.parasitology
PEPTIDES	bionet.molecules.peptides
PHOTOSYNTHESIS	bionet.photosynthesis
PLANT-BIOLOGY	bionet.plants
PLANT-EDUCATION	bionet.plants.education
PLANT-SIGNAL-TRANSDUCTION	bionet.plants.signaltransduc
POPULATION-BIOLOLGY	bionet.population-bio

表 5.13 (续表)

简称	新闻组地址
PRENATAL-DIAGNOSTICS	bionet.diagnostics.prenatal
PROTEIN-ANALALYSIS	bionet.molbio.proteins
PROTEIN-CRYSTALLOGRAPHY	bionet.xtallography
PROTISTA	bionet.protista
PSEUDOMONADS	bionet.organisms.pseudomonas
RAPD	bionet.molbio.rapd
RECOMBINATION	bionet.molbio.recombination
SCHISTOSOMA	bionet.organisms.schistosoma
SCIENCE-RESOURCES	bionet.sci-resources
STADEN	bionet.software.staden
STRUCTURAL-NMR	bionet.structural-nmr
SYMBIOSIS-RESEARCH	bionet.biology.symbiosis
TIBS	bionet.journals.letters.tibs
TOXICOLOGY	bionet.toxicology
TROPICAL-BIOLOGY	bionet.biology.tropical
URODELES	bionet.organisms.urodeles
VECTOR-BIOLOGY	bionet.biology.vectors
VIROLOGY	bionet.virology
X-PLOR	bionet.software.x-plor
YEAST	bionet.molbio.yeast
ZBRAFISH	bionet.organisms.zebrafish

5.14.2 网上医学信息资源

本手册以生物学特别是分子生物学信息资源为主，关心医学、药学信息的读者可参阅 [R-14] 一书所列信息。这里只介绍几个网站，从他们出发可以链接到大量有关网址。

R-810 **MedMatrix** 是集临床医学信息大成的一个网页，可以在注册后免费进入。通过 MedMatrix 也可以访问许多新闻组和论坛。网址：

`http://www.medmatrix.org/`

R-811 北京生物技术和新医药产业促进中心主办的“新生命 - 北京生物医药在线”，是一个值得注意的中文网站。这里除新闻外，还有医药资源导航信息和一些可下载的通用软件。网址：

`http://www.newlifebp.org.cn/`

5.14.3 网上期刊和出版社

网上电子期刊有两大类，一是没有印刷本的“纯”电子刊物，二是与印刷本同步的电子版，两者数目都在增加，尤以后者为最．预计不久的将来，每一种重要学术期刊都会有电子版．然而，多数刊物规定，只有印刷版的订户才能免费阅读电子版和下载文章．生命科学期刊“上网”也越来越多，查找网上“在线” (On Line) 刊物的方便办法，是访问斯坦福大学图书馆的名为 High Wire 的网上服务 [R-812] ．

R-812 斯坦福大学图书馆 **HighWire** 服务，网址：

http://intl.highwire.org/

这里有一张刊物名单，逐一说明是摘要还是全文上网，是全免费还是对一定时间之前的过期刊免费，同时也列出各刊物的网址．

下面列举一些重要刊物的电子版．

R-813 美国《科学》(*Science*) 周刊．在我国自然科学基金委员会、科技部、教育部和中国科学院资助下，凡具有 .cn 域名的中国用户可以自由阅读其网络版，即 *Science-on-Line* ．网址：

http://china.sciencemag.org/

事实上，网络版比纸面版的内容更丰富，例如可通过“超链接”访问引文，了解引用情况．特别是生命科学方面的某些文章加有“超注释” (hypernote) ，能帮助读者迅速从网上追溯引文和掌握有关背景知识．

R-814 根据中国科学院与美国科学院的协议，凡具有 .cn 域名的中国读者可自由阅读或下载美国科学院院报 (*Proceedings of the National Academy of Sciences* USA ，简称 *PNAS*) ，否则只能免费阅读 18 个月之前的过期刊．网址：

http://intl.pnas.org/

R-815 英国《自然》(*Nature*) 周刊，可以免费订阅由电子邮件送达的每期目录，请查询：

http://www.nature.com/ 或

http://www.natureasia.com/

R-816 *MMBR* ，《微生物分子生物学评论》(*Microbiology and Molecular Biology Review*) ，十一个月之前的过刊可免费阅读或下载．网址：

http://intl-mmbr.asm.org/

下面一些杂志的网络版是完全免费阅读的.

R-817 *Protein Science*, 《蛋白质科学》:

http://www.proteinscience.org/

R-818 *BMJ*, 《不列颠医学杂志》 (*British Medical J.*):

http://www.bmj.com/

R-819 *AJP*, 《美国生理学杂志》的《教育》分卷 (*Am. J. Phisiology. Adv. in Physiol. Ed.*):

http://intl-ajpadvan.physiology.org/

AJP 的其他分卷只能免费阅读一定时间之前的过期刊物.

R-820 *J. Clin. Invest.*, 《临床研究杂志》:

http://www.jci.org/

R-821 **ESP** 是一个网上电子学术出版社 (Electronic Scholarly Publishing) 组织, 它提供可以免费下载的遗传学经典文献. 它的目录从 1798 年马尔萨斯 的人口论著作、 1865 年孟德尔 的植物杂交论文 (德文原文和英译本), 到包括摩尔根 实验室在内的 20 世纪上半叶的重要贡献. 当然还有达尔文 (Charles Darwin) 的著作. 有些文献在中国原是难得一睹的. 我们奉劝学者访问下面的网址:

http://www.esp.org/

5.14.4 会议消息和会议文集

许多生物信息中心的网页上都有会议消息, 有些系列会议有专门网页, 某些会议文集的电子版本可以免费下载. 我们列举一些网址.

R-822 北京大学生物信息中心的网页上, 有较为丰富的会议消息. 请看:

http://www.cbi.pku.edu.cn/conferences.html

R-823 在 TIGR 研究所 [R-156] 的网页上可以查到将由该所组织的一些会议的消息. 网址:

http://www.tigr.org/conf/

R-824 在冷泉港实验室 CSHL [R-159] 的网页上也有会议消息:

http://nucleus.cshl.org/meetings/

R-825 **PSB** 年会, 从 1996 开始, 每年 1 月在夏威夷举行太平洋生物计算

研讨会 (Pacific Symposium on Biocomputing，简称 PSB)．它比较注重算法．从过去的 PSB1996 到将要召开的 PSB2001，均可在加州大学旧金山校区的网址查看：

http://www.cgl.ucsf.edu/psb/

历届会议文集都由新加坡世界科学出版社 (World Scientific Publishing Co.) 印行．大部分论文也收入相应电子文集，可以从下面的网址下载 (XX=96 到 00，表示 1996 到 2000)：

http://www-smi.stanford.edu/projects/helix/psbXX/

R-826 **ISMB**，即分子生物学中的智能系统 (Intelligent Systems for Molecular Biology) 国际会议，自 1993 年以来已每年举行，2000 年举行了第 8 届．2001 年的第 9 届会议将在哥本哈根举行．这是着重研讨算法的会议，有不少序列分析和数据库搜索方面的文章，特别是神经网络和隐马可夫链模型的讨论．会议文集由 AAAI Press 出版．历次会议概况可参阅网页：

http://ismb00.sdsc.edu/prev_mtgs.html

http://www.aaai.org/

R-827 **RECOMB**，计算分子生物学国际年会 (Annual International Conference on Computational Molecular Biology)，自 1997 年起举行．第 4 届即 RECOMB2000 已在东京开过．前几届的会议文集已收入美国计算机协会的数字化图书馆 (ACM Digital Library)．虽然下载文章需付费，但可免费读取目录．网址：

http://www.acm.org/pubs/contents/proceedings/

R-828 **BOSC**，2000 年首次举行的生物信息学公开源程序会议 (Bioinformatics Open Source Conference 2000) 是 1999 年 BioPerl99 会议的继续，有可能发展成一个系列会议．请参看网址：

http://ismb00.sdsc.edu/bosc2000/

R-829 **RNA** 国际会议，1997 和 1998 的文集见：

http://www-smi.stanford.edu/people/altman/rna97.html

http://www.wisc.edu/union/info/conf/rna/rna.html

R-830 **STRUBE**，欧洲结构生物学会议 (Structural Biology in Europe)，请参看网址：

http://www.biodigm.com/strube.htm

R-831 **DDPS**，药物发现与蛋白质科学会议 (Drug Discovery and Protein Science)，亦请参看 STRUBE [R-830] 的网页。

R-832 **TMMeC**，分子模拟电子会议 (The Molecular Modeling E-Conference)，在全球有多处镜象，可从以下网址开始查阅：

`http://fcindy5.ncifcrf.gov/tmmec/`

5.14.5 讲义和课程

R-833 **Biorithms** 是 ICGEB [R-152] 的 S. Pongor 为几次生物信息学讲习班所写讲义的电子版的总题目，详见：

S. Pongor, "Algorithms for molecular biology" (1998)

网址：

`http://www.icgeb.trieste.it/net/courseware/`

R-834 北京大学生物信息中心在 1999 年 4 月与 ICGEB 合作举办了分子生物学数据库和分析工具的国际研讨会和讲习班，有 3 个报告保存在网页上：

`http://www.cbi.pku.edu.cn/meeting/icgeb/talk.html`

这些报告是：

1. Ed. Wingender, "Database modelling of gene regulation".
2. Bruno Gaeta, "Database similarity search".
3. Bruno Gaeta, "Patterns, profiles, and motif search".

R-835 法国 Rouen 大学 C. Charras 和 T. Lecroy 编写的《序列比较讲义》，可从网址

`http://www.dir.univ-rouen.fr/~charras/seqcomp/`

下载文件 seqcomp.ps。

R-836 FASTA 的作者 W. R. Pearson 本人关于其 3.0 新版的讲义，可在多个网点查看，例如：

`http://www.techfak.uni-bielefeld.de/`
`bcd/Lectures/pearson3.html`
`http://www.biotech.ist.unige.it/`
`bcd/Lectures/pearson3.html`
`http://merlin.mber.bcm.tmc.edu:8001/bcd/`
`bcd/Lectures/pearson3.html`

这些 URL 都属于自然科学虚拟学校 (Virtual School of Natural Science，简称 VSNS) 的生物计算部 (BioComputing Division，简称 BCD)。从所列 URL 往上查，还可以找到其他电子课程的记录。

R-837 **VSMS**，即医学科学虚拟学校 (Virtual School of Molecular Sciences)，其网址也值得访问：

http://www.vsms.nottingham.ac.uk/vsms/

5.14.6 一些有益的个人网页

R-838 **Amos' WWW Links Page**，瑞士的 Amos Bairoch 汇编的 WWW 链接地址清单，包含一千多处网址。其中不少本书前面已经提及。这个清单的好处，是已经分门别类，便于查询。请参看：

http://www.expasy.ch/alinks.html/

北京大学生物信息中心有镜象：

http://expasy.pku.edu.cn/amos_www_link.html

R-839 **Pedro** 网页，是 Pedro M. Coutinho 当研究生时建立的，由于内容比较丰富，曾经被广为引用，可惜自 1996 年初以来更新不及时。网址：

http://www.public.iastate.edu/~pedro/

R-840 **Willy** 网页，搜集了一批与生物学有关的超链接。网址：

http://genome1.bio.bnl.gov/

R-841 **Ranst** 网页，由瑞典的 Marc van Ranst 维护。网址：

http://www.ng.bik.se/~nstrna/mvr.htm

R-842 美国洛克菲勒大学统计基因组学实验室李问天 (Wentian Li) 的网页上，除了前面已经提到的微阵列文献目录 [R-786]，还有很多有益的信息，特别是关于 DNA 序列中核苷酸关联的研究情况。网址：

http://linkage.rockefeller.edu/wli/

李问天网页的许多内容在北京大学生物信息中心 [R-166] 有镜象。

R-843 石家庄华北制药集团金坦生物技术开发有限公司的谈杰建立了一个生物、化学免费软件网页。许多国际上的自由软件都已下载保存在那里，并有一个总目录供查询。网址：

http://www.ncpcgt.col.com.cn/zhigong/tanjie/index.html

有时不能直接访问这个子目录，要从一级主页经“职工园地”进入。

5.14.7 法律、伦理和社会影响

人类基因组计划的实施，引起了许多伦理道德、法律、社会问题，即所谓 ELSI (Ethical, Legal, and Social Implications)；还有一个生物技术的安全性，都是公众关心的问题．通过互联网向社会进行宣传，协助学校教师提高基因和生物知识的教育水平，是生物信息中心不可忽视的责任。有关信息可参考以下网址和它们的链接：

R-844 **BINAS** 生物安全服务，即生物安全信息网与咨询服务 (Biosafety Information Network and Advisory Service)，是联合国工业发展组织 (UNIDO) 提供的服务，它反映全球关于生物技术的各种法规的状况，可以通过北京大学生物信息中心 [R-166] 网页上的 UNIDO 链接进入．

R-845 **GeneLetter** 网上基因通信：
`http://www.geneletter.org/`

R-846 美国橡树岭国家实验室的 **ELSI** 链接：
`http://www.ornl.gov/hgmis/resource/elsi.html`

R-847 美国国家基因资源中心 NCGR [R-135] 的遗传学与公众网页：
`http://www.ncgr.org/gpi/index_gpi.html`

R-848 美国堪萨斯大学医学中心的基因教育中心 (Genetic Education Center，简称 GEC) 有一个面向广大公众的网页：
`http://www.kumc.edu/gec/`
请参看 [R-616]．

R-849 美国华盛顿大学基因中心设有针对中学生物学教师的“高中人类基因组计划” (High School Human Genome Project，简称 **HSHGP**)。除了一般信息，还为教师准备了程序模块、虚拟测序等．网址：
`http://hshgp.genome.washington.edu/`

§5.15 生物信息资源的近期发展动向

我们在这本手册中已经列举了上千个网址．这当然不能覆盖所有重要的生物信息资源．在结束全书之前，再简单讨论一些生物信息资源的近期发展动向．

第一，各种软件和数据库的集成化．这在第 4 章和第 5 章中已经多次提到，不再复述。

第二，数据库和软件系统集成化的统一标准．我们只提一下 CORBA 和面向对象这两个互相关联的问题。

R-850 **CORBA** 即 Common Object Request Broker Architecture，乃是国际对象管理协作组 (Object Management Group，简称 OMG) 制定的、使 OOP 对象与网络接口统一起来的一套跨越计算机、操作系统、程序语言和网络的共同标准。OMG 的网址是：

`http://www.omg.org/`

CORBA 标准的汉译本见：

OMG 著，《CORBA：系统结构、原理和规范》，电子工业出版社，2000．

其实，CORBA 并不是专为生物学制定的．由于历史原因，生物数据库的组织方式多种多样，而 WWW 只处理超文本文件．为了各种应用程序能方便地经互联网联接各种数据库，欧洲国家已经决定采纳 CORBA 的协议及其界面定义语言 (Interface Definition Language，简称 IDL) 作为共同标准．详情请参看 EBI [R-131] 的网页．目前已经纳入这个框架的数据库有 EMBL [R-211] 核酸序列库、PIR [R-404] 蛋白质库、SWISS-PROT [R-401] 蛋白序列库、MSD [R-443]、GDB [R-283]、TRANSFAC [R-219]、RHdb [R-281]、p53 [R-324] 等。关于 CORBA 在生物信息学方面的应用，还请参看新近建立的网页 [R-61]：

`http://biocorba.org/`

R-851 **ACeDB**，即基于面向对象的程序设计 (OOP) 思想的线虫数据库本身，是一套可以用于其他生物数据库的自由软件．目前许多基因组测序计划均采用 ACeDB 作数据管理系统．例如 GrainGenes [R-572]、MsqDB [R-375]、TreeGens [R-583]，RiceGenes [R-571]、CSNDB [R-558]，以及中国科学院遗传研究所基因中心进行的泉生热袍菌 (*Caldotoga fontana*) 的基因组测序计划．ACeDB 的简单描述可参看 [R-17] 一书的第 13 章和 [R-21] 一书的第 22 章．数据库本身可以从许多网址下载：

`http://alpha.crbm.cnrs-mop.fr/`

ftp://lirmn.lirmn.fr (/pub/acedb/)

ftp://ftp.sanger.ac.uk (/pub/acedb/)

ftp://cele.mrc-lmb.cam.ac.uk (/pub/acedb/)

ftp://ncbi.nlm.nih.gov (/repository/acedb/)

还有一批从 ACeDB 派生出来的软件，主要是各种界面，例如：

1. Web 界面 Webace，详见网址：

http://webace.sanger.ac.uk/

2. Java 界面 Jade，描述见：

L. Stein, "Jade: an approach for interconnecting bioinformatics databases", *Gene* **209** (1998) 39 – 43.

3. Perl 界面 AcePerl，以及 AQL、WinAce 等。

安装和使用 ACeDB 的许多经验，可以向以下网址或新闻组查询：

http://probe.nalusda.gov:8000/acedocs/acedbfqa.html

http://www.bio.net:80/hypermail/ACEDB/

ftp://rtfm.mit.edu

访问 /pub/usenet/news.answers/acedb-faq。

mailto: mail-server@rtfm.mit.edu

news: bionet.software.acedb

第三，知识环境 (Knowledge Environment，简称 KE) 的建设。这里只举一个刚刚开始的实验系统。

R-852 **STKE** 信号转导知识环境是由美国《科学》周刊和斯坦福大学图书馆共同建立的第一个网上知识环境。不同于刊物上的综述文章，这里由专家撰写的总结文章处于经常更新之中。例如，对发育过程重要的 Wnt 转导途径，此网页上有华盛顿大学 Randall Moon 提供的知识。用户可就一般情况和具体物种、组织或细胞类型两个层次提出询问。具体到 Wnt，还可参看 Wnt 网页 [R-430]。STKE 的网址：

http://www.stke.org/index.html

第四，在统一标准下群策群力发展自由生物信息软件的努力。例如：

R-853 **EMBOSS** 是英国 Sanger 中心正在实现中的欧洲分子生物学开放软件系统。这是一套基于 UNIX 命令行的、具有统一风格、不断发展扩大的程序包。其长远目标是为学术界建立高质量的、种类齐全的免费

软件系统，可以按照 GNU [R-62] 自由软件协议享用．除了 EMBOSS 小组自己编写的程序，它也欢迎各国学者按同样的风格体例做出贡献．由于是公开的免费系统，它也可能把许多现存的自由软件集成进去．现在 EMBOSS 系统中已经有一批可以使用的程序，包括本书前面提到过的 BLAST [R-631]、Blixem[R-647]、dotter[R-749]、Gap4[R-690]、Phrap[R-691] 等程序．详见：

http://www.sanger.ac.uk/Software/EMBOSS/

目前试运行的 β 版本，可以下载：

ftp://ftp.sanger.ac.uk (/pub/EMBOSS/)

取文件 EMBOSS-0.0.4.tar.Z．

最后，但可能是最重要的动向之一，是如何发展对于用户透明的获取网上资源的系统，即不必关心 URL 而方便、直接地联接到信息源的手段．请参看：

R-854 **TAMBIS** (Transparent Access to Multiple Biological Information Sources) 计划，即多种生物信息资源的透明获取．这是英国曼彻斯特大学生物科学和计算科学两个学院正在合作进行的项目．它的最终目标，是使用户只通过一个 URL，访问全部 WWW 上的相关信息．请看网址：

http://www.cs.man.ac.uk/mig/tambis/

R-855 **ISYS** 是美国国家基因资源中心 NCGR [R-135] 正在发展的一个软件平台，它具有网上导航、浏览、视象化和进行分析的功能，帮助用户利用序列、基因图谱、代谢途径、基因表达等各方面的数据库和应用程序．这个系统基本上用 Java 语言编写，但是在用户代理服务器和数据库以及应用程序服务器之间，保证了与 CORBA [R-850] 接口协议的兼容性．2000 年 8 月，为软件发展人员提供的 ISYS 核心程序已经可以下载，但为最终用户服务的版本要再等一段时间．ISYS 系统对纯学术用户免费，详情请参看网址：

http://www.ncgr.org/research/isys/

TAMBIS 计划比较侧重数据库端的界面，而 ISYS 系统更着眼于用户端的功能．因此，从目前的描述判断，这两套系统在相当程度上将是互补的．

索　引