第2版

生物信息学

手册

郝柏林
张淑誉 编著

上海科学技术出版社

责任编辑 ● 叶　剑
封面设计 ● 戚永昌

ISBN 7-5323-6774-6
9 787532 367740 >
定价：42.00 元

www.sstp.com.cn

China

生物信息学手册

第 2 版

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. —2版.
上海: 上海科学技术出版社, 2002.12
ISBN 7-5323-6774-6

Ⅰ.生… Ⅱ.①郝…②张… Ⅲ.生物信息学-手册 Ⅳ.Q811.4-62

中国版本图书馆 CIP 数据核字（2002）第090223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)
常熟市兴达印刷有限公司印刷 新华书店上海发行所经销
2000 年第 1 版
2002 年 12 月第 2 版 2002 年 12 月第 2 次印刷
开本 787×1092 1/16 印张 19.75 插页 6 字数 293 000
印数 3 001—8 200 定价: 42.00 元

内容提要

目前，从细菌到人类，从核酸、蛋白质到基因表达和信号传导，各种生物数据库的信息量正在迅猛增长．生物学不再是“数学等于零”的学科，也不再仅仅是基于观察和实验的科学．理论和计算将对生物学的进步发挥日益巨大的作用，计算机和网络技术正在为生物学研究插上腾飞的翅膀．生物信息学作为一门崭新的前沿交叉学科也应运而生．

狭义的生物信息学基于从海量信息中提取新的知识．这涉及到数据库、界面、算法、软件和网上服务．国际互联网上每日每时都在更新和重组的资源，很容易使人在浩如烟海的信息中迷失方向．本书扼要叙述了生物数据库、算法、程序和服务的基本情况，列举了上千条网址和引文，力图成为生物信息汪洋大海中的导航图．

本书可供广大生命科学和生物技术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．

再版前言

正如我们在本书初版前言中所说：“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。”再版前对原书所开列的网址进行过核查，删去了一些过时的地址，增加了一批新的网上资源. 由于人类基因组工作框架图的完成，相应介绍比原来简短，但网上信息更为充实. 主要的更动是增加了关于常用算法的第 5 章，把原来介绍软件和网上服务的第 5 章推后成第 6 章. 由于根本不可能在此书篇幅内讲解算法，第 5 章只是列举了若干备查的公式和引文，因而更具有手册性. 这一章基于 2001 年秋季在华大基因中心 [R-176] 同浙江大学联合举办的生物信息学研究生班开设的《生物信息学引论》课程的部分内容.

许多同行和读者对本书初版提出了宝贵的改进意见. 两年多来我们又从很多朋友那里学习了不少新知识. 这里特别要感谢苏州大学谢惠民，复旦大学钟扬，中国科学院上海生命科学研究院生物信息中心李亦学，国家人类基因组南方中心任双喜，美国加州大学圣巴巴拉校区李明、河滨校区江涛、旧金山校区李浩、圣地亚哥校区华泰立，美国橡树岭国家实验室徐鹰，美国冷泉港实验室张奇伟，美国普林斯顿 NEC 研究所汤超，美国北岸 LIJ 研究所李问天等诸位学者，特别是华大基因中心的王俊以及一大批实践在生物信息工作前沿的年轻人.

作者在生物信息学领域的研究工作，受到国家重点基础研究专项经费、中国科学院创新工程和北京市 248 创新项目资助. 我们还要特别感谢复旦大学提供的工作条件以及两年多来同华大基因中心在研究工作和人才培养方面的愉快合作.

作者　　2002 年 9 月 10 日

于上海复旦大学理论生命科学研究中心

初版前言（摘录）

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． 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学研究所 (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，重要面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。

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 36] 就是第 9 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–214] 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，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–474]、德国转录因子数据库 TRANSFAC [R–227]、美国的 RepBase [R–232] 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 141]、欧洲生物信息学研究所 EBI [R 138] 和北京大学生物信息中心 CBI [R-174] 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-86] 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 706] 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-219] 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. 当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 [1] . 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.

[1] 见 *Science* **284** (1999) 1742 .